Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, March 2020, pp.481-490

Employment of the genetic algorithm in some methods of estimating survival function with application

Sabah Manfi Redha[1], Anwar Taher Abdel Hadia[2]
[1,2] College of Business and Economics, University of Baghdad

ABSTRACT

Intended for getting good estimates with more accurate results, we must choose the appropriate method of estimation. Most of the equations in classical methods are linear equations and finding analytical solutions to such equations is very difficult. Some estimators are inefficient because of problems in solving these equations. In this paper, we will estimate the survival function of censored data by using one of the most important artificial intelligence algorithms that is called the genetic algorithm to get optimal estimates for parameters Weibull distribution with two parameters. This leads to optimal estimates of the survival function. The genetic algorithm is employed in the method of moment, the least squares method and the weighted least squares method and getting on more efficient estimators than classical methods. Then, a comparison will be made between the methods depending on the experimental side. The best method is evaluated based on mean square error of the survival function and the methods will be applied to real data for patients with lung and bronchia cancer.

Keywords: Survival function, Two parameter Weibull distribution, Genetic algorithm, Method of moment, Least squares method, Weighted least squares method, Censoring data

Corresponding Author:

Anwar Taher Abdel Hadia
College of Business and Economics
University of Baghdad
E-mail: anwertaher17@gmail.com

1. Introduction

For the purpose of studying the time of survival in the case of a dangerous disease, it is necessary to determine the appropriate model or distribution that follows the time to be studied. After we have identified the appropriate model, this model is estimated and some estimators are inefficient due to problems in solving non-linear equations using traditional methods. This may lead to an inaccurate estimate of the survival function and hence the objective of this paper to estimate the survival function by employing the method of genetic algorithm to getting on optimal estimates of the parameters of Weibull distribution with two parameters and thus to getting an optimal estimate of the survival function.

1.1. Two parameter Weibull distribution model

Weibull distribution is one of the important distributions in the study of human life times and it is also used to study the times of survival machine and stop it. It was first found by Walodd Weibull in 1951, and this distribution has the ability to describe all stages of failure that it is going through, Whether it is a human or a machine, such as the stage of increasing or decreasing failure, and the probability density function of the distribution of the two parameter Weibull take the formula [1].

$$f\,(\,t\,,\alpha\,,\beta\,) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right) \quad \alpha\,,\beta\, > 0\,, I_{(0,\alpha)} \qquad \dots (1)$$

Notation:
β: Shape parameter
α: Measurement Parameter
As for the distribution function, it takes the following formula:

The Weibull distribution parameters have an effect on the shape of the probability density function as shown in the forms below.

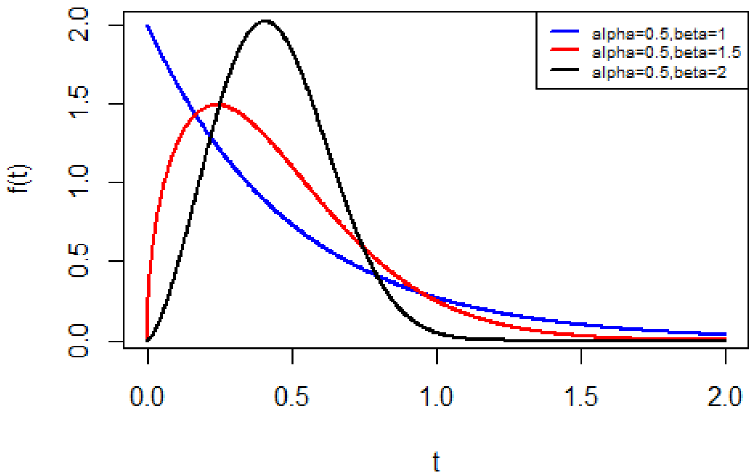

Figure 1. The effect of the shape parameter in the form of a probability density function for Weibull distribution

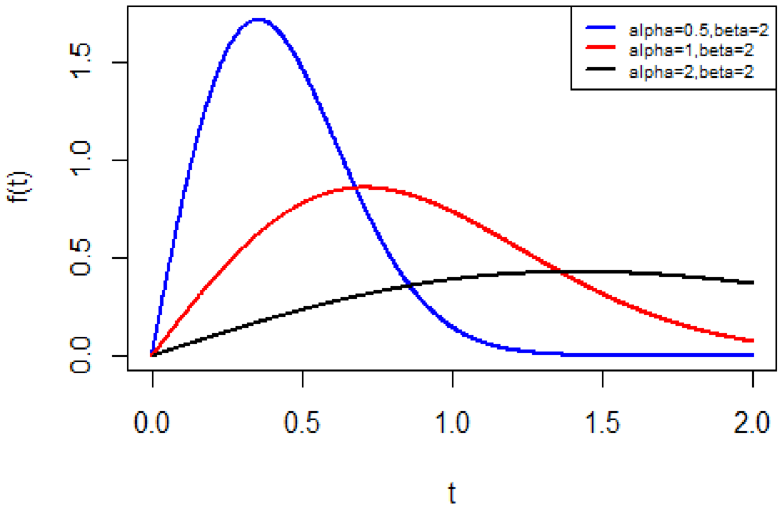

Figure 2. The effect of the measurement parameter in the form of a probability density function for Weibull distribution

1.2. The survival function of Weibull distribution

The equation for the survival function can be found as follows [2]**:**

$$S(t) = 1 - F(t, \alpha, \beta, \theta)$$

$$F(t,\alpha,\beta) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right)$$

$$S(t) = 1 - \left(1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right)\ \right)$$

1.3. Two parameter Weibull distribution properties

The characteristics are the finding of the arithmetic mean, variance and, the mode as follows [3]:

Arithmetic Mean

$$\mu = \alpha\Gamma\left(1 + \frac{1}{\beta}\right) \qquad \ldots(4)$$

Variance

$$\sigma^2 = \alpha^2 \Big[\Gamma \Big($$

The Mode

$$M_o = \alpha\left(1 - \frac{1}{\beta}\right)^{\frac{1}{\beta}}, \qquad \beta \geq 1 \qquad \ldots(6)$$

2. Parameters estimation method

In this section, we will have the most important methods of estimating the two parameter Weibull distribution in the case of the censored data from the right in order to estimate the survival function using the genetic algorithm and these methods include:

- Least Squares Method (LSM)
- Weighted Least Square Method (WLSM)
- Method of Moments (MOM)

2.1. Least squares method

The least squares method (LSM) method is extensively used in many practical problems in process of estimating the parameters of the models, if we have the following model [4]:

The least squares estimates of the parameters β_1, β_o are the values of the parameters which miniaturizing the function:

$$Z(\beta_{0,}\beta_1) = \sum_{i=1}^{n} wi(y_i - \beta_0 - \beta_1 x_i)^2$$

Therefore, the estimates of β_1, β_o take the following formula:

$$\hat{\beta}_o = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

In order to estimate the Weibull parameters β, α by using the least squares method in the beginning the distribution form must be converted into linear form as follows:

$$\ln\left(-ln\big(1-F(t)\big)\right) = \beta \ln(t) - \beta \ln(\alpha)$$

By making the equation above similar to the equation of regression number (7) and being as follows:

$$y_i = \ln\left(-\ln\big(1-F(t)\big)\right)$$

The values of x_i take the formula:

$$x_i = \ln(t_i)$$

and,

$$\beta_0 = -\beta \ln(\alpha)$$

$$\beta_1 = \beta$$

So if we have a random sample $t_1, t_2, \dots \dots, t_n$ drawn from the Weibull distribution with order statistics,

$$T_{(1)} < T_{(2)} < \cdots < T_{(n)}$$

and let,

$$t_{(1)} < t_{(2)} < \cdots < t_{(n)}$$

By arranging observations, the distribution function is estimated using the average rank as follows:

Here, i refers to i^{th} as the smallest value of $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ and ($\mathrm{i} = 1,2, \dots n$). Therefore the estimates of β_1, β_o becomes:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} ln(\ t_{(i)}) ln\left(-ln\left(1-\hat{F}\left(t_{(i)}\right)\right)\right) - \sum_{i=1}^{n} ln\left(t_{(i)}\right)\sum_{i=1}^{n} ln\left(-ln\left(1-\hat{F}\left(t_{(i)}\right)\right)\right)}{n\sum_{i=1}^{n} ln^2\left(t_{(i)}\right) - \left(\sum_{i=1}^{n}\left(t_{(i)}\right)\right)^2}$$

and,

$$\beta_1 = \beta$$

It means that $$\hat{\beta} = \hat{\beta}_1$$

and,

$$\beta_0 = -\beta\, ln(\alpha)$$

this implies that,

$$\alpha = exp\left(-\left(\frac{\beta_0}{\beta}\right)\right)$$

hence,

After finding estimators by using least square method, we estimate the survival function through the following formula:

$$\hat{S}(t) = \exp\left(-\left(\frac{t}{\hat{\alpha}_{LS}}\right)^{\hat{\beta}_{LS}}\right).$$

2.2 Weighted least square method

The Weighted Least squares method is extensively used to estimate the parameters of the multiple linear regression model. Bergman and others in 1993 used them to estimate the two parameter Weibull distribution parameters [5].

So if we have the following model:

The estimates of Weighted Least squares for the above model are minimized by:

$$Z(\beta_{0,}\beta_1) = \sum_{i=1}^{n} wi(y_i - \beta_0 - \beta_1 x_i)^2$$

Where , w_i , $i = 1,2, \dots, n$ represents weights

In order to estimate the two Weibull distribution parameters using weighted least squares, the shape of the distribution function must be converted into linear as follows:

$$\ln\left(-ln\big(1 - F(t)\big)\right) = \beta \ln(t) - \beta \ln(\alpha)$$

By making the equation above similar to the equation of regression number (12), it will be as follows:

$$y_i = \ln\left(-\ln\big(1 - F(t)\big)\right)$$

The values of x_i take the formula:

$$x_i = \ln(t_i)$$

and,

$$\beta_0 = -\beta \ln(\alpha)$$

$$\beta_1 = \beta$$

A random sample of $t_1, t_2, \dots \dots, t_n$ is drawn from the Weibull distribution with order statistics as below:

$$T_{(1)} < T_{(2)} < \cdots < T_{(n)}$$

Suppose that:

$$t_{(1)} < t_{(2)} < \cdots < t_{(n)}$$

Arranged views, the distribution function for it is estimated using the mean rank as follows:

The rank of median is:

Bergmam proposed a weight function that could be used to estimate parameters that take the following formula:

$$w_i = \left[\left(1 - \hat{F}\big(t_{(i)}\big)\right) ln\left(1 - \hat{F}\big(t_{(i)}\big)\right)\right]^2$$

Faucher and others in 1988 also proposed a weight function that takes the following formula:

$$w_i = 3.3\,\hat{F}\left(t_{(i)}\right) - 27.5\left[1 - \left(1 - \hat{F}\left(t_{(i)}\right)\right)^{0.025}\right]$$

For the purpose of obtaining the estimates of α, β using this method as follows:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} w_i\, y_i - \hat{\beta}_1 \sum_{i=1}^{n} w_i t_i}{\sum w_i}$$

$$\hat{B}_1 = \frac{\sum w_i \sum w_i t_i y_i - \sum w_i t_i \sum w_i y_i}{\sum w_i \sum w_i\, t_i^2 - (\sum w_i t_i)^2}$$

In compensation for the values of w_i , y_i , t_i , it is equal to the distribution function of the distribution function of the Weibull distribution in the estimates of $\hat{\beta}_{1,}\hat{\beta}_0$ as follows:

$$\hat{\beta}_1 = \frac{\sum w_{(}}{}$$

Thus, the estimates are:

$$\hat{\beta} = \beta_1$$

Note that the weight function proposed by bergmam will be used and takes the following formula:

$$w_i = \left[\left(1 - \hat{F}\left(t_{(i)}\right)\right) ln\left(1 - \hat{F}\,\left(t_{(i)}\right)\right)\right]^2$$

After finding the estimates using the weight least square method, we estimate the survival function through the following formula:

$$\hat{S}(t) = \exp\left(-\left(\frac{t}{\hat{\alpha}_{MWLS}}\right)^{\hat{\beta}_{MWLS}}\right)$$

2.3 Method of moments

This method is the essence of the equality between the moment of population and the corresponding sample and therefore we will get a number of equations for the parameters of population and by solving these equations we get the required estimates as this method is used to get the probability distribution of a random variable through the moment generation function, and the parameters of the two parameter Wiebull distribution are get by using the sample mean $\bar{t}$ and the sample variance s^2, where [6]:

$$\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}\,(\,t_i - \bar{t}\,)^2}{n-1}.$$

By equalizing the moment of population with the determination of the sample, the variance of population with the variance of sample is:

$$\mu = \bar{t}$$

$$\bar{t} = \alpha\Gamma\left(1+\frac{1}{\beta}\right)$$

$$s^2 = \alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]$$

The $\hat{\beta}$ parameter can be gotten by dividing the variance on the square mean:

$$= \frac{\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]}{\left[\alpha\Gamma\left(1+\frac{1}{\beta}\right)\right]^2}$$

By simplifying the equation above, the estimators are:

and,

After finding the estimators using the method of determination, we estimate the function of survival through the following formula:

$$\hat{S}(t) = \exp\left(-\left(\frac{t}{\hat{\alpha}_{mom}}\right)^{\hat{\beta}_{mom}}\right)$$

Genetic Algorithm (GA) is an intelligent algorithm based on the genetic evolution process of biomimetic natural population to solve the global optimal solution. Because of the advantages of global optimization, strong extensibility and good robustness, this paper combines traditional genetic algorithm with three parameter of Weibull distribution [7].

3. Results and discussion

The simulation includes a comparison of the traditional methods of estimation with two parameter of Weibull distribution parameters based on the mean square error criterion. These methods are small squares (LS), method of moment (MOM), and weighted least squares (WLS). The simulation includes comparison between the methods of Estimation when using the methods of intelligence by using GA. Different sample sizes (20 , 40 , 60 , 100) and different initial values of parameters for the experiment were repeated 1000 times and the following results were obtained.

Table 1. MSE when β = 1.2302 & α = 92.938

	n=20	n=40	n=60	n=100
Moment	0.0171381324203749	0.0157979985911530	0.0261172121826833	0.0228725782251598
LS	0.107182976109739	0.107872595644358	0.116953039014751	0.0968795236049614
WLS	0.00229587131687498	0.0173079743957992	0.0227340432818551	0.119414027136930
Best	WLS	Moment	WLS	Moment
Moment_GA	0.000694552069589193	0.000151679530006394	5.00471182607 e-05	1.1510545982 e-05
LS_GA	0.00025781657956210	6.27764098640677e-05	1.8383394688 e-05	2.8014760686 e-06
WLS_GA	0.000428361053084686	9.49686900790413e-05	2.9744732752 e-05	1.01253721586 e-05
Best	LS_GA	LS_GA	LS_GA	LS_GA

Table 2. MSE when β = 1.50 & α = 90.20

	n=20	n=40	n=60	n=100
Moment	0.071061578359983	0.069995396062682	0.088148406278593	0.080342637308021
LS	0.107174481619469	0.107870102107717	0.116897085499951	0.096843699204611
WLS	0.003969185562138	0.026868565323276	0.038565102789301	0.162465102987633
Best	WLS	WLS	WLS	Moment
Moment_GA	0.000777219892194	0.000176351397725	6.17436919871e-05	1.46545267553e-05
LS_GA	0.000264447372241	7.091284200334e-05	2.27591652183e-05	3.97802704681e-06
WLS_GA	0.000459611811341	0.000106023751410	3.59002053475e-05	1.29192610092e-05
Best	LS_GA	LS_GA	LS_GA	LS_GA

Table 3. MSE when β = 1.20 & α = 95.20

	n=20	n=40	n=60	n=100
Moment	0.0128356641733582	0.011613753155469	0.02066055444940	0.01782770466134
LS	0.107176803682558	0.107859153948333	0.11690832603015	0.09682649285122
WLS	0.00214383319311772	0.016357722823996	0.02117928308706	0.11447535268917
Best	WLS	Moment	Moment	Moment
Moment_GA	0.00069086089481638	0.000150969304654	4.98870825618 e-05	1.16137283595e-05
LS_GA	0.00025687629877521	6.20970576256e-05	1.81554889833e-05	2.79721208147e-06
WLS_GA	0.00042527864374103	9.403878615665e-05	2.94218145391e-05	9.94250372817e-06
Best	LS_GA	LS_GA	LS_GA	LS_GA

3.1. Analysis of simulation results

We note from Table 1 that the best method at the size of a sample 20 is the weighted least squares (WLS) because it has the least mean squares error. When the sample size is 40, the best method is the method of moment. At what time the sample size is increased to 60, the method of weighted least squares again is the best. When the sample size became 100, the method of moment is the best among the methods. We note that when the methods of estimation were employed in the genetic algorithm, the least square method is the best of all methods for different sizes of the sample.

From Table 2, we note that the best method at the sample size of (20, 40, and 60) is based on method of weighted least squares, because it has the least mean squares error. When the sample size is 100, the best method is the method of moment. We note that when the methods of estimation were employed in the genetic algorithm, the least square method is the best of all methods for different sizes of the sample.

We note from Table 3 that the finest method at the size of a sample 20 is based on method of WLS, because it has the least mean squares error. At what time the sample size is (40 or 60 or 100), the method of moment is the best. Besides, we note that when the methods of estimation were employed in the genetic algorithm, the least square method is the best in all methods for different sizes of the sample.

Based on the results of Tables 1-3, the best traditional estimation methods for estimating the two parameter Weibull distribution is the method of weighted least squares. On the other hand, when these methods are used in the genetic algorithm, the finest method is the least squares method.

4. The application

The data was collected from the patient's records in terms of patient's name, age, date of entry, type of disease and treatment schedules. The record contains details of the patient as well as the type of disease, the stage of the disease and the amount of dose given at each stage. Specialized doctors in the disease were interviewed and questioned about doses and patients.

These diseases need to be closely censored, and the development of the patient's condition in case of response or non-response to the dose must be recorded. These records and details are taken from Al-Amal hospital. The national oncology and data were recorded from the records of lung cancer patients including a time period for the difference between the last patient's review and the first review measured in days for 2018

4.1. Survival function estimate

The survival function was estimated based on summary study produced. The results showed that the best method is the method of least squares method with the genetic algorithm after taking the censored data from the right (right-censored). The MATLAB simulator has been used to obtain the results as in Table 4.

Table 4. Survival function estimate using the least square method with the genetic algorithm of lung cancer

Time	$\hat{s}(t)_{LS-GA}$	Time	$\hat{s}(t)_{LS-GA}$	Time	$\hat{s}(t)_{LS-GA}$	Time	$\hat{s}(t)_{LS-GA}$
1	0.3105	23	0.3104	45	0.3103	67	0.3101
2	0.3105	24	0.3104	46	0.3103	68	0.3101
3	0.3105	25	0.3104	47	0.3103	69	0.3101
4	0.3105	26	0.3104	48	0.3103	70	0.3101
5	0.3105	27	0.3104	49	0.3103	71	0.3101
6	0.3105	28	0.3104	50	0.3103	72	0.3100
7	0.3105	29	0.3104	51	0.3103	73	0.3100
8	0.3105	30	0.3104	52	0.3103	74	0.3100
9	0.3105	31	0.3104	53	0.3102	75	0.3100
10	0.3105	32	0.3104	54	0.3102	76	0.3100
11	0.3105	33	0.3104	55	0.3102	77	0.3099
12	0.3105	34	0.3104	56	0.3102	78	0.3099
13	0.3105	35	0.3104	57	0.3102	79	0.3098
14	0.3105	36	0.3104	58	0.3102	80	0.3097
15	0.3105	37	0.3104	59	0.3102	81	0.3097
16	0.3105	38	0.3104	60	0.3102	82	0.3096
17	0.3105	39	0.3104	61	0.3102	83	0.3096
18	0.3105	40	0.3104	62	0.3102	84	0.3096
19	0.3105	41	0.3103	63	0.3102	85	0.3095
20	0.3104	42	0.3103	64	0.3101	86	0.3094
21	0.3104	43	0.3103	65	0.3101	87	0.3091
22	0.3104	44	0.3103	66	0.3101		

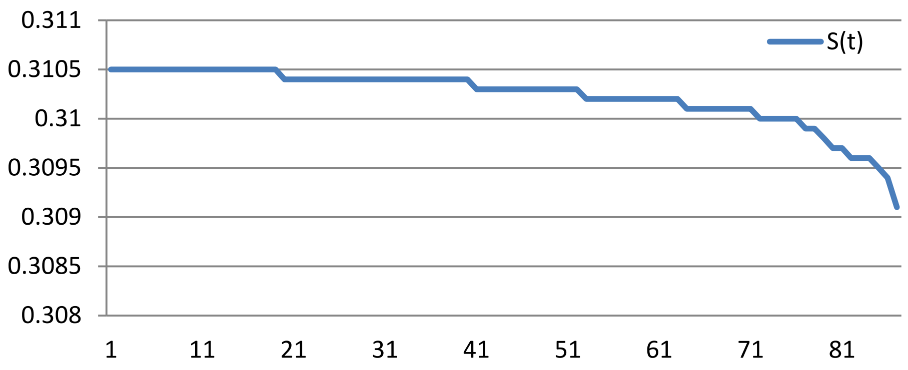

Figure 3. Survival function estimate using the least square method with the genetic algorithm of lung cancer

5. Conclusions

Through the presented simulation study and practical study, we have reached the following conclusion:

- The best method to estimate the two parameter of Weibull distribution is to use the weighed least squares among the traditional method.
- The best method to estimate the two parameter of Weibull distribution when the traditional methods were employed in the genetic algorithm is the least square method .
- In the case of small sample sizes, the best method is weighted least squares among the traditional methods and the least squares of the methods employed in the genetic algorithm .
- In the case of the size of the intermediate samples, there is a parity in the preference between the method of moment and weighted least squares method of size. But, when employing the genetic algorithm, the method of least squares is the best.
- In the case of large sample sizes, the method of attribution is the best among the traditional methods and the method of least squares of the genetic algorithm.

6. Recommendations

From what we have concluded, we recommend the following:

- Use of the least squares with the genetic algorithm when estimating the survival function of the two parameters of Weibull distribution, especially in the medical aspect and in dangerous diseases.
- Use the genetic algorithm to analyze survival and reliability functions.
- Conducting similar studies and taking the distribution of Weibull for more than two parameters.
- Conducting similar studies and taking other distributions that are also suitable for use in the analysis of survival functions.

References

[1] H. Rinne, *The Weibull distribution*. Justus-Liebig-University, Giessen, Germany, published by Chapman & Hall/CRC, ISBN: 13: 978-1-4200-8743-7, 2009.

[2] E. Lee and J. Wang, *Statistical methods for survival data analysis*.University of Oklahoma Health Sciences Center , Oklahoma city , india, John Wiley & Sons, 2003.

[3] M. A. Nielsen, “Parameter Estimation for the Two-Parameter Weibull Distribution,” M.Sc. Thesis, Department of Statistics, Brigham Young University, 2011.

[4] P. Osatohanmwen, F. O. Oyegue, and E. Joseph, “An Appraisal on Some Methods for Estimating the 2-Parameter Weibull Distribution with Application to Wind Speeds Sample” Sri Lankan Journal of Applied Statistics, Vol :18,No:3 ,pp. 146–166,2018.

[5] H. Lu, C. Chen, and J. Wu, “A Note on Weighted Least-squares Estimation of the Shape Parameter of the Weibull Distribution,” Journal of Quality and Reliability Engineering, vol .20 , pp. 579–586, 2004.

[6] A.M Razali, A.A. Salih and A.A. Mahdi, 'Estimation accuracy of Weibull distribution parameters', Journal of Applied Sciences Research, vol. 5, no. 7, pp. 790-795.2009

[7] C.-J. Wen, X. Liu, and X. Cheng. "Parameter Evaluation of 3-parameter Weibull Distribution based on Adaptive Genetic Algorithm.", 2nd International Conference on Machinery, Electronics and Control Simulation (MECS 2017). Atlantis Press, 2016.