Journal of Mathematics and Statistics Studies (JMSS)
www. al-kindipublisher.com/index.php/jmss

The Effect of Censoring Percentages on the Performance of Gamma Distribution in Analysing Survival Data

Yusuf Abbakar Mohammed
Department of Mathematical Sciences, University of Maiduguri, PMB1069, Maiduguri, Nigeria.
Corresponding Author: Yusuf Abbakar Mohammed, E-mail: yusufabbakarm@gmail.com

ARTICLE INFORMATION

Received: September 03, 2020
Accepted: October 05, 2020
Volume: 1
Issue: 1

KEYWORDS

Gamma; Random censoring; Parameter; Mean square error; Survival data

ABSTRACT

The Gamma distribution was employed to investigate the performance of the model in estimating the maximum likelihood parameter of the model. Simulated data were employed to investigate the performance of the model by considering five different censoring percentages (0%, 10%, 20%, 30% and 40%) and three sets samples of size (100, 300 and 500) observations. The parameters of the Gamma Distribution were estimated successfully. The simulation was repeated 300 times and the mean square error (MSE) and root mean square error (RMSE) were estimated to assess the consistency and stability of the model. The simulated data used to compare the effect of different censoring percentages revealed that the model performed much better with small percentage of censored observations. As the censoring percentage increases the model seems to under estimate the shape parameter and overestimate the scale parameter. The Gamma model showed that survival model is affected by the increase in the percentage of lost information in the data set. However, increasing the sample size helps the model to estimate the parameter of interest much more precise and consistent.

1. Introduction

Survival analysis usually used to analyse an event of interest that occurs within a specified period of time. The techniques of survival analysis are commonly employed in different fields such as engineering, biological sciences, sociology, economic and engineering to mention few. The nonparametric methods are frequently used to analyse survival data. Pure classical parametric survival models are very powerful methods in survival analysis; they perform better than the nonparametric methods when the chosen distribution fit the data properly. The Gamma distribution is frequently employed in analysing survival data (Ibrahim, Chen and Sinha, 2001; Kalbfleisch and Prentice, 2002; Lawless, 2003; Lee and Wang, 2003). The Gamma model has been in the literature long time ago. The gamma distribution was used to model the glass tumblers survival time in circulation in a cafeteria (Brown and Flood, 1947). The distribution was also employed in modelling the life length of materials (Birnbaum and Saunder, 1958). In the recent decades, many authors employed the Gamma distribution in modelling survival data by applying mixture model technique.

Two components survival mixture models of Gamma-Gamma were used to model survival data (Erisoglu, Erisoglu and Erol, 2012), they implemented model selection technique to select the model which better represents the real data. The effect of censoring percentage on the performance of both parametric and non-parametric method have been in the literature and it was observed that the model performs better with the low percentage of censoring. Mohammed, Yatim and Ismail (2019) investigated the performance of survival mixture model of three components of different distributions of exponential, Gamma and Weibul with different censoring percentages and different set of mixing probabilities. The study showed that the model performed better with low levels of censoring percentages. Another study considered the mixture model of three components of Gamma distribution with different censoring percentages where it was concluded that the low levels of censoring percentage allow the model to estimate the parameter much better (Mohammed and Ismail, 2019).

The arrangement of the paper is as follows. In section two the survival analysis and some properties of the Gamma distribution are highlighted. Section three devoted to data application to evaluate the parameters of the model and compare the different censoring percentages. Section four devoted for summary and conclusion.

2. Survival Analysis and Probability distributions

Survival analysis concern with the application of some statistical method to model and analyse survival data. The focus of interest is the occurrence of a particular event of interest within a given period of time. The response of variable *T* is a non-negative random variable which gives the survival time of an object or an individual which can be expressed as a probability density function (pdf) denoted by $f(t)$, which is written as

$$f(t)=\frac{dF(t)}{dt}$$

Where $F(t)$ is the distribution function of response variable *T.* The probability density function can also be presented graphically, the graph of $f(t)$, is known as the density curve. The density function $f(t)$ is a nonnegative function and the area between the curve and the *t* axis is equal to 1. The survival function denoted by $S(t)$ can be written as

$$S(t)=1-F(x)$$

Which gives the probability that an individual will survive beyond a particular time *t*. Note that the survival function $S(t)$ is a monotonic decreasing continuous function with $S(0)=1$ and $S(\infty)=0$. The hazard function can be represented by $h(t)$, and is given by

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

which gives the probability of an individual to fail within a small interval $(t,t+\Delta t)$, provided that the individual was a life until the beginning of that interval.

Pure classical parametric survival models are powerful method in survival analysis; when the chosen probability distribution appropriately represents the data. The Exponential, Gamma and Weibull densities are commonly employed in the analysis of survival data. (Kalbfleisch and Prentice, 2002; Lawless, 2003; Lee and Wang, 2003; Erisoglu, Erisoglu and Erol, 2011; Erisoglu, Erisoglu and Erol, 2012). The probability density function $f(t)$ and survival functions $S(t)$ of these distributions are highlighted below.

Gamma distribution

$$f_G(t)=t^{\alpha-1}\frac{e^{-t/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \qquad t \text{ and } \alpha,\beta>0$$

$$S_G(t)=1-\frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$$

Where $\Gamma_x(\alpha)$ is known as the incomplete Gamma function.

3. Data Analysis

The performance of the Gamma distribution model was investigated by employing simulated data generated with ($\alpha_1 = 20$, $\beta_1 = 6$) as shape and scale parameters respectively. The sample generated was of size 100, 300 and 500 observations with (0%, 10%, 20%, 30% and 40%) as different censoring percentages. In this study the random right censoring set of data was used in with the censoring observations were considered for each of the sample generated in which, t_j=min(T_j,C_j) was taken as the minimum of the survival time and the censored time of the observed time *T* where

$$T=\begin{cases}\delta_i=1, & if \quad X\le C,\\ \delta_i=0, & if \quad X>C.\end{cases}$$

Table 1: The Estimated Parameters of the Simulated Data

Simulated Sample of Size 100					
parameters	0%	10%	20%	30%	40%
α = 20	20.373060	18.713064	17.150743	15.69430	13.25195
β = 6	5.934415	6.597421	7.378351	8.31274	10.21333

The parameters of the samples of size 100 observations with different censoring percentages were estimated successfully. Table 1 showed that the estimated parameters for smaller censoring percentages are most close to the postulated parameters used in the data generation. It can be clearly observed that as the censoring percentage increases the shape parameters were under estimated and the scale parameters were over estimated. The probability density function of the simulated data of the proposed model, with 0%, 10%, 20%, 30% and 40% censoring percentages respectively, and the histogram of the simulated data are displayed in Figures 1, 2, 3, 4 and 5.

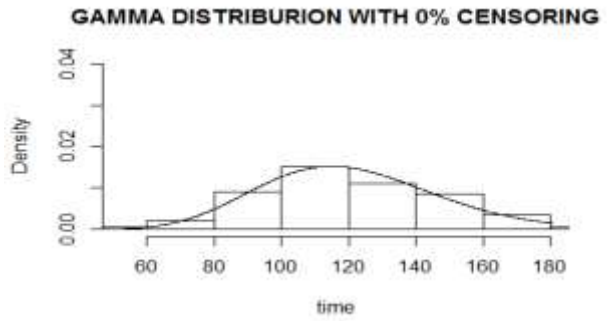

Figure 1: Density Function of the Simulated Data with 0% Censored Observations.

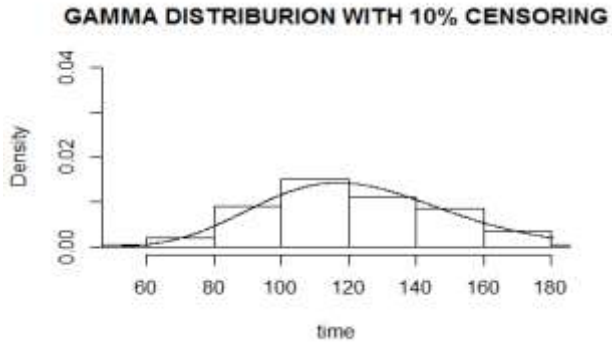

Figure 2: Density Function of the Simulated Data and 10% Censoring.

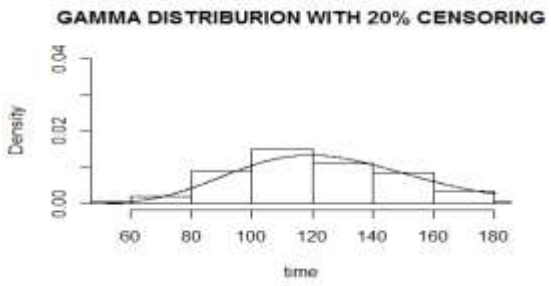

Figure 3: Density Function of the Simulated Data and 20% Censoring.

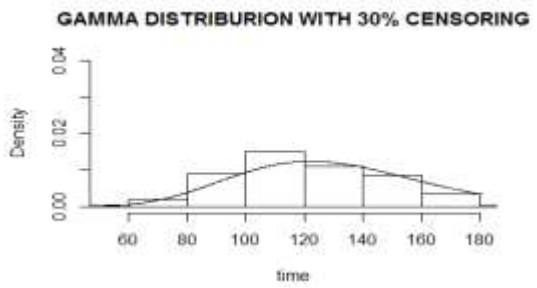

Figure 4: Density Function of the Simulated Data and 30% Censoring.

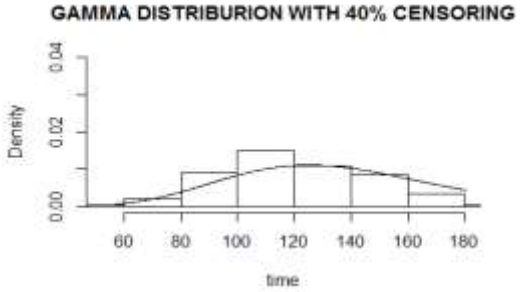

Figure 5: Density Function of the Simulated Data and 40% Censoring.

From the figures above it can be observed clearly that the graph with smaller censoring observation fit the histogram of the simulated data better.

Another sample of size (300 observations) was generated with (0%, 10%, 20%, 30% and 40%) as different censoring percentages. The sample was used to estimate the parameter of the Gamma model investigate the performance with large sample size of 300 observations.

Table 2 showed that the estimated parameters for smaller censoring percentages are most close to the postulated parameters used in the data generation. It can be clearly observed that as the censoring percentage increases the shape parameters were under estimated and the scale parameters were over estimated.

Table 2: The Estimated Parameters of the Simulated Data

Simulated Sample of Size 300 Observations					
para	0%	10%	20%	30%	40%
$\alpha = 20$	19.978438	18.683521	17.014600	15.86455	14.870795
$\beta = 6$	6.033255	6.589333	7.413317	8.19728	9.070262

The probability density function of the simulated data of the proposed model, with 0%, 10%, 20%, 30% and 40% censoring percentages respectively, and the histogram of the simulated data are displayed in Figures 6, 7, 8, 9 and 10.

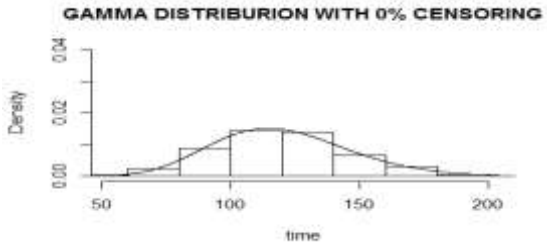

Figure 6: Density Function of the Simulated Data and 0% Censoring.

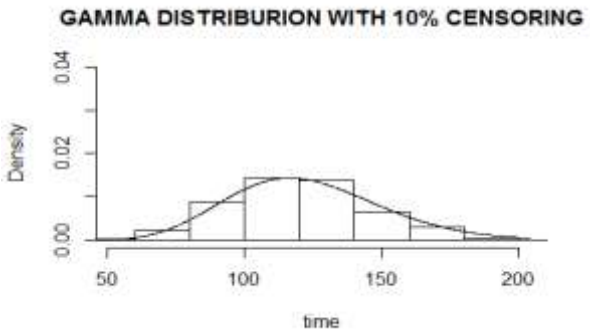

Figure 7: Density Function of the Simulated Data and 10% Censoring.

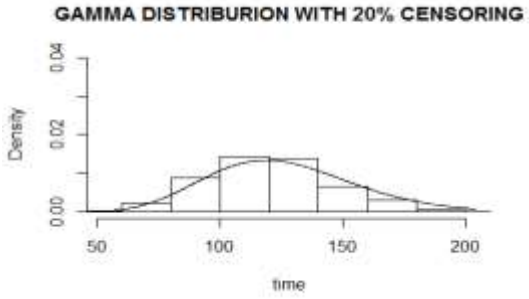

Figure 8: Density Function of the Simulated Data and 20% Censoring.

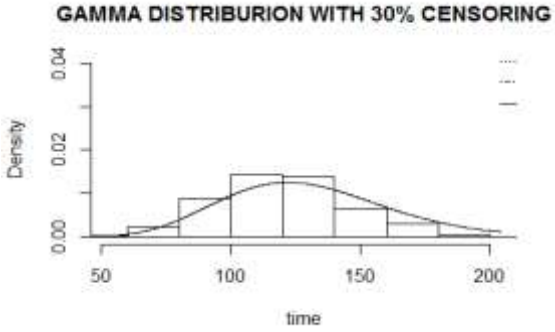

Figure 9: Density Function of the Simulated Data and 30% Censoring.

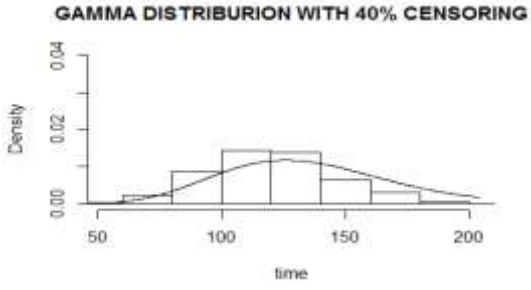

Figure 10: Density Function of the Simulated Data and 40% Censoring.

From the figures above it can be observed clearly that the graph with smaller censoring observation fit the histogram of the simulated data better.

A sample of size 500 observations was generated with (0%, 10%, 20%, 30% and 40%) as different censoring percentages. The sample was used to estimate the parameter of the Gamma model investigate the performance with larger sample of size 500 observations. Table 3 shows that the estimated parameters for smaller censoring percentages are most close to the postulated parameters used in the data generation. It can be seen that as the censoring percentage increases the shape parameters were under estimated and the scale parameters were over estimated.

Table 3: The Estimated Parameters of the Simulated Data

Simulated Sample of Size 500 Observations					
parameters	0%	10%	20%	30%	40%
$\alpha = 20$	19.567500	18.40257	17.022226	16.78245	14.943753
$\beta = 6$	6.161628	6.69645	7.424673	7.77155	9.040239

The probability density function of the simulated data of the proposed model, with 0%, 10%, 20%, 30% and 40% censoring percentages respectively, and the histogram of the simulated data are displayed in Figures 11, 12, 13, 14 and 15.

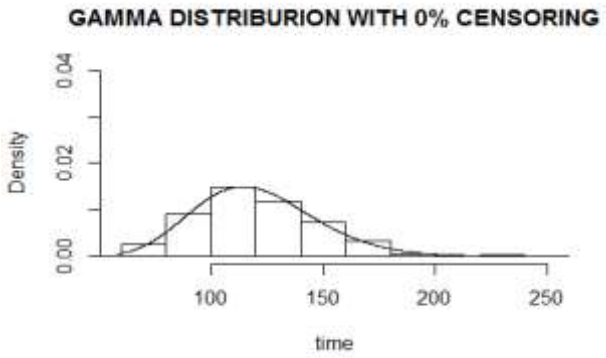

Figure 11: Density Function of the Simulated Data and 0% Censoring.

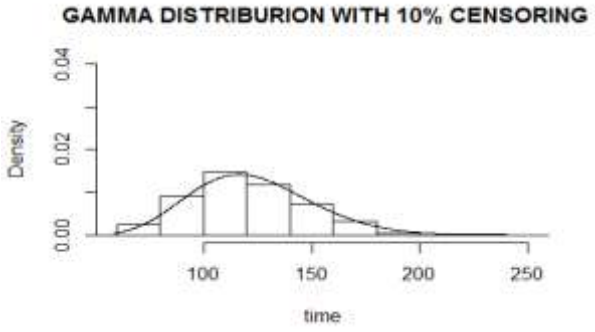

Figure 12: Density Function of the Simulated Data and 10% Censoring.

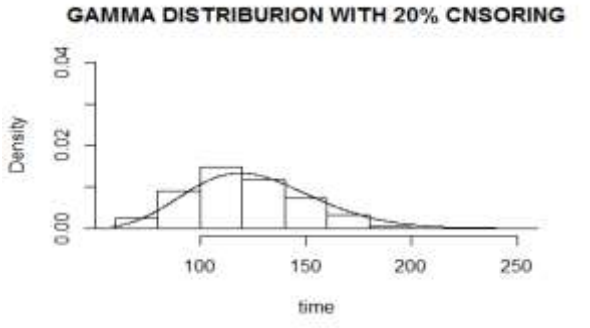

Figure 13: Density Function of the Simulated Data and 20% Censoring.

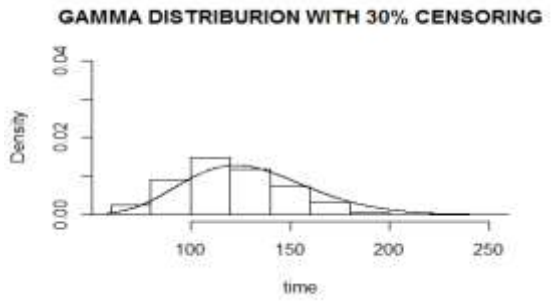

Figure 14: Density Function of the Simulated Data and 30% Censoring.

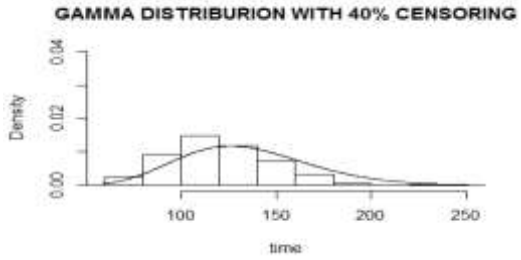

Figure 15: Density Function of the Simulated Data and 40% Censoring.

From the figures above it can be observed clearly that the graph with smaller censoring observation fit the histogram of the simulated data better.

The three sets of the generated data of samples of size 100, 300 and 500 observations with 0%, 10%, 20%, 30% and 40% censored observations were repeated 300 times to check the consistency and stability of the Gamma distribution in estimating the parameter of the model. The averages, the mean square errors (MSE) and root mean square error (RMSE) of estimated parameters were listed in Table 4, 5 and 6.

Table 4: The Repeated Simulation of Sample of Size 100 Observations

Sample of 100			
0% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	20.733651	0.030948259	0.17503055
β = 6	5.920621	0.002604427	0.04846923
10% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	19.835459	0.029692822	0.17231605
β = 6	6.276676	0.002842168	0.05331199
20% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	18.09863	0.027384804	0.1654835
β = 6	7.09276	0.004215905	0.0649300
30% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	16.952183	0.025898626	0.16093050
β = 6	7.757616	0.005597519	0.07481657
40% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	15.86206	0.024953440	0.15796658
β = 6	8.53225	0.007605433	0.08720913

Table 5: The Repeated Simulation of Sample of Size 300 Observations

Sample of 300			
0% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	20.215193	0.0075751772	0.08703549
β = 6	5.968878	0.0006477936	0.02545179
10% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	18.947366	0.0077709949	0.0009556476
β = 6	6.502743	0.08815325	0.03091355
20% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	17.394775	0.007417883	0.08612713
β = 6	7.292215	0.001423152	0.03772469
30% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	15.555158	0.006332478	0.07957687
β = 6	8.467811	0.002196520	0.04686705
40% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	14.295562	0.005649070	0.07516030
β = 6	9.525263	0.003135803	0.05599824

Table 6: The Repeated Simulation of Sample of Size 500 Observations

Sample of 500			
0% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	20.120729	0.0049350173	0.07024968
β = 6	5.985371	0.0004410226	0.02100054
10% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	18.62434	0.0046602692	0.06826616
β = 6	6.63218	0.0006281286	0.02506249
20% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	16.961157	0.0041637022	0.06452676
β = 6	7.501936	0.0008883411	0.02980505
30% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	15.317132	0.003821888	0.06182142
β = 6	8.615659	0.001349532	0.03673597
40% censoring observations			
PARAMETER	ESTIMATE	MSE	RMSE
α = 20	13.893223	0.003666863	0.06055463
β = 6	9.879999	0.002238125	0.04730883

The tables above show that the model was able to estimate the shape and scale parameters successfully with their corresponding Mean Square Errors (MSE) and Root Mean Square Errors (RMSE). The MSE and RMSE values are relatively small for the three sample sizes and different censoring percentages. It can be observed that the model was able to estimate the parameter more accurately for samples with small censoring observation. As the censoring percentage increases the model tends to under estimate the values of the shape parameter and overestimate the value of the scale parameters. Also it can be

seen that the values of the MSE and the RMSE are getting small as the sample size is increased which shows that as the sample size increase the variation decreases.

4. Conclusion

The article investigated the performance of Gamma distribution with different censoring percentages and sample sizes. Simulated data were used to evaluate and assess the performance of the Gamma model. The maximum likelihood estimator technique was employed to estimate the parameters of the model. The simulated data used to compare the effect of different censoring percentages revealed that the model performed much better with small percentage of censored observations. The investigation of the model showed that the percentage of the censored observation affects the performance of the model; the model seemed to perform much better with lower censored observation. Moreover, with the increase in the sample size of the data the model performed more consistently in estimating the maximum likelihood estimators.

References

[1] Birnbaum, Z. W., & Saunders, S. C. (1958). A statistical model for life-length of materials. *Journal of the American Statistical Association*, *53*(281), 151-160.

[2] Brown G. W. & Flood, M. M., (1947). Tumbler mortality. Journal of the American statistical Association. 42, 562-574.

[3] Erişoğlu, Ü., Erişoğlu, M., Erol, H. 2011, A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences, 5(2).* http://www.scopus.com/inward/record.url?eid=2-s2.0-78449258657&partnerID=40&md5=901faa5759d0767b0b2676000e17839c.

[4] Ibrahim, J. G., Chen, M. H., Sinha, D., 2001, *Bayesian survival analysis*. New York: Springer-verlag.ISBN 0-387-95277-2.

[5] Kalbfleisch J. D., Prentice R. L. .(2002). *The statistical analysis of failure time data* (2nd ed.). John Wiley & Sons, Inc. Hoboken, New Jersey. ISBN 0-471-36357-X.

[6] Lawless J. F., (2003) *Statistical models and methods of lifetime data*, (2dn ed.). John Wiley and Sons, Inc. Hoboken, New Jersey. ISBN 0-471-37215-3.

[7] Lee, E. T., Wang, J. W., 2003, *Statistical methods for survival time data analysis* (3rd ed.). John Wiley & son. New Jersey. ISBN 0-471-36997-7.

[8] Mohammed, Y. A. and Ismail, S., (2019). Mixture model of different distributions: A simulation study with different censoring and mixing probabilities. *International Journal of Science and Research, 8*(5)

[9] Mohammed, Y. A., & Ismail, S. (2019). A Simulation Study of Survival Mixture Model of Gamma Distributions with Different Set of Censoring Percentages and Mixing Proportions. IOSR Journal of Mathematics,15(6) Ser. VI (Nov – Dec 2019), PP 39-48.

[10] Shelby, C. (2013). The Use of Survival Analysis Techniques Among Highly Censored Data Sets.