International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Analysis of Continuous and Discrete Time-to-Event Data Using Parametric Techniques

Joseph Okello Omwonylee[1], Thomas Okello[2]

[1]Department of Mathematics, Kyambogo University, Kyambogo, Uganda

[2]Department of Land, Faculty of Engineering, Kyambogo University, Uganda

Abstract: ***In this paper, the effect of discretization of time-to-event data on parameter estimates is investigated with the objective of finding out how discretization of nearly continuous or continuous survival data affects the outcome of the parameter estimates. Monte Carlo simulation was used to simulate data with different sample sizesfor the study. Discretisation of the simulated data was made. The parameters of the Weibull and the exponentially distributed models were estimated using maximum likelihood estimation techniques with the help of Davidon-fletcher-Powell optimization formula in MATLAB program for both the continuous and the discretized data. Using the two-sample Kolmogorov-Smirnov test, the hypothesis that the discrete and continuous samples come from the population with the same distribution could not be rejected for samples with sizes of less than 100 but rejected for sample of more than 100 sample sizes. It was also found out that discretization of survival data reduces their precision by increasing the parameter estimates. Researchers studying time-to-event data** are therefore advised to avoid over discretization in order to reduce biasedness **in the parameter estimates. Smaller counting units should be expressed as a proportion of the bigger counting unit used and in the event that there is no event in a given interval, they should** resort into interpolation to find the missing value.*

Keywords: Continuous survival data, discrete survival data, Monte Carlo simulation, Kolmogorov-Smirnov test, Weibull and the exponential models, Maximum likelihood estimation

1. Introduction

The time which is the backbone of survival analysis can be measured in days, weeks, months, years, in which case is often discretized. The three main objectives of time-to-event (Survival) analysis are; to compare time-to-event between two or more group; to assess the relationship of the co-variables to time-to-event; and to estimate time-to-event for a group of individuals (cohort). A lot of literature is available on survival/reliability (time-to-event) analysis and survival data but the treatment of the variation arising from continuous and discrete survival data analysis is lacking. For instance, Omwonylee, et al., (2014), in their study on modelling the return time of persons who had been displaced by Lord Resistance Army measured return time in years. This means a family that returned in January was considered to have returned at the same time with the one who returned in December of the same calendar year. Saleem, et al., (2012), in the study about the coronary artery bypass graft surgery (CABG) patient also measured the event in years and in which case the event that occurred in January might have been considered to have taken place at the same time as that of December if the months were considered in such manner. The question is therefore whether the finding would still have remained the same if the researcher considered January through December as the proportion below of a year?

Table 1.1: Months Expressed as Year

Jan.	Feb.	Mar	Apr.	May	Jun.
0.0833	0.1667	0.2500	0.3333	0.4167	0.5000
Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
0.5833	0.6667	0.7500	0.8333	0.9167	1.0000

With the growing application in sociology, An educationist wishing to study dropout from school might face level in accuracy if the counting process of dropout is made yearly because he/she will consider a child who dropout in first term the same way the treatment is given to the other child who dropout in third term with progress both is term one and two.

1.1. Survival Analysis Techniques

According to Collett, D. (2003), Survival analysis is a phrase used to describe the analysis of data in the form of time from a well-defined time origin until the occurrence of the particular event of interest or the end point of the study.It is therefore a class of statistical techniques used for studying the occurrence and timing of events. They were originally designed for the event of death occurrence and hence name survival analysis. The techniques is extremely useful for studying many different kinds of events in both the social and natural sciences, such as the onset of disease in Biostatistics, equipment failures in engineering, earthquakes, automobile accidents, stock market crashes, revolutions, job terminations, births, marriages, divorces, promotions in job places, retirements, Contracting Lung cancer due to smoking, arrests and many other time to event data. According to Omwonylee, et al., (2014), In Biostatistics, this techniques are often referred to as clinical trials, in Engineering they are referred to as Reliability analysis or failure time analysis, in econometric they are either duration analysis or transition analysis, and in Sociology it is often referred to as event history analysis. This is because Survival analysis techniques have been adopted by researchers in several different fields.

The three well known techniques for analyzing time to event data are; parametric, semiparametric and nonparametric each with its own limitation. With **Parametric models,** the outcome is assumed to follow a certain known distribution. It is also thought that parametric approach may yield better results provided the assumptions made in the analysis are correct as seen in Omwonylee, et al., (2014) and Buis, (2006).

In **non-parametric models**, which include life table and **Kaplan-Meier estimates**, there is no assumption about the shape of the hazard function or about how covariates may affect that shape. It is therefore mainly descriptive and fails to control for covariates, requires categorical predictors, and cannot accommodate time-dependent variables.

Semi-parametric models such as the Cox and the piecewise constant exponential model are particularly flexible since they make no assumption about the shape of the hazard but they make a strong assumption about how the covariates affect the shape of the hazard function between groups over time.

2. Methodology

In this paper the parametric models of Weibull and Exponential distribution is estimated using MLE with the data set of discrete and continuous properties. Detail discussion of the Weibull and Exponential time-to-event models can be found in Abernathy, (1998), Klein and Moeschberger, (1997, 2003), Kleinbaum and Klein, (2005), Lawless (2005) and Leemis (1995). Application of Weibull and exponential models can be seen in Khan, et al., (2011), Omwonylee, et al (2014)and Saleem, et al., (20012). The method of Monte Carlo simulation was used to generate data sets of different sample sizes which had properties of continuous or nearly continuous. To discretize the data, the data are grouped into countable interval where all the data are moved to the upper counting numbers as shown in the example of table 4.1 with a sample of size 40. Most of the data analysis was done using MATLAB software with the application of Davidon-Fletcher-Powel optimization technique.

These models are chosen, not only because of their popularity among researchers who analyze survival data, but also because they offer insight into the nature of the various parameters and functions, particularly, the hazard rate and survival function.
After discretization, two set of samples (discrete and continuous samples) were form which were then tested whether the two samples are drawn from the same distribution using two-sample Kolmogorov-Simonov goodness-of-fit hypothesis test. The parameters of the Weibull and exponential distribution were estimated using the maximum likelihood estimation techniques at 5% level of significance and their properties are investigated using total deviation, root mean square errors and biasedness.

2.1 Maximum Likelihood Estimation.

Lawless, (2003) proposed the form of likelihood function for the survival model in the presence of censored data. The maximum likelihood method works by developing a likelihood function based on the available data and finding the estimates of parameters of a probability distribution that maximizes the likelihood function. The likelihood function for all observed and censored Subjects were defined by:

$$L(t_i,\underline{\theta}) = \prod_{i\in u}[f(t_i,\underline{\theta})] \times \prod_{i\in c} S(t_i;\,\underline{\theta})$$

$$= \prod_{i=1}^{n}[f(t_i,\underline{\theta})]^{f_{t_i}} \prod_{i=1}^{n}[S(t_i;\,\underline{\theta})]^{c_{t_i}} \quad (2.1)$$

where f_{t_i} are the number of observed subjects until the event of interest has happened in the interval i and c_{t_i} are the number of censored individuals in the interval i each of length t, $f(t_i,\underline{\theta})$ is probability density function (pdf), a parametric model with survivor function, $S(t_i,\underline{\theta})$ and the hazard function, $h(t_i,\underline{\theta})$ with the vector parameter $\underline{\theta} = (\alpha,\beta)$ of the model for the case of the Weibull distribution model.

Since we are dealing with a complete sample that has no censored individual, then the equation (2.1) becomes;

$$L(t_i,\underline{\theta}) = \prod_{i=1}^{n}[f(t_i,\underline{\theta})]^{f_{t_i}} \quad (2.2)$$

To obtain maximum likelihood estimates of parameters of a Weibull model, logarithm is taken on both sides of the above equation (Likelihood function) and therefore by setting $l(t_i,\underline{\theta}) = lnL(t_i,\underline{\theta})$ (log-likelihood function) results into:

$$l(t_i,\underline{\theta}) = \sum_{i=1}^{n} f_{t_i} ln\left[f(t_i,\underline{\theta})\right] \quad (2.3)$$

It is worth noting that $S(t_i,\underline{\theta}) = 1 - F(t_i,\underline{\theta})$ and equation (2.3) is the same as the equation (2.2) but several events are considered to have happened in the interval I.

Also since $f(t_i,\underline{\theta}) = h(t_i,\underline{\theta}) \times S(t_i,\underline{\theta})$, then equation (2.3) becomes

$$l(t_i,\underline{\theta}) = \sum_{i=1}^{n} f_{t_i} ln\left[h(t_i,\underline{\theta})\right] + \sum_{i=1}^{n} f_{t_i} ln\left[S(t_i,\underline{\theta})\right] \quad (2.4)$$

Where, the first summation is for failure and the second summation is for all censored individuals.

For the estimation of the parameters, there is need to find out the hazard function and the survival function to be substituted in the log likelihood function and hence apply suitable iteration techniques to come out with the parameter estimates.

2.2 Survival function and Hazard function

For the parametric survival model, the survival function is defined by

$$S(t;\underline{\theta}) = \int_{t}^{\infty} f(x)dx \quad (2.5)$$

2.2.1 Weibull model

We define a Weibull distribution's probability density function (pdf), mathematically by:

$$f(t;\,\underline{\theta}) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} exp^{\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]} \quad (2.6)$$

$t \geq 0\ \alpha\ (scale) > 0\ ,\&\ \beta(slope) > 0$

Therefore

$$S(t;\underline{\theta}) = \int_t^\infty \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^\beta} dx$$

$$S(t;\underline{\theta}) = e^{-\left(\frac{x}{\alpha}\right)^\beta}\Big|_t^\infty = e^{-\left(\frac{t}{\alpha}\right)^\beta} \quad (2.7)$$

Where; $\underline{\theta} = [\alpha, \beta]$, $f(x)$ is the probability density function of the Weibull distribution function for this case.

The hazard function, also called the force of mortality in Biostatistics and epidemiology especially in clinical trials is the instantaneous failure rate. Mathematically the hazard function is defined by

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

$$h(t;\underline{\theta}) = \frac{f(t;\underline{\theta})}{S(t;\underline{\theta})} = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^\beta}}{e^{-\left(\frac{t-\gamma}{\alpha}\right)^\beta}}$$

$$h(t;\underline{\theta}) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} \quad (2.8)$$

$= P$(Experiencing the event of interest in the interval $(t, t + \delta_t)$|survived past time, t)
$= P(t < T < t + \delta_t | T > t)$

2.2.2 Exponential Model

Exponential distribution has a probability density function (pdf), mathematically defined by:

$$f(t;\ \lambda) = \left(\frac{1}{\lambda}\right) exp^{\left[-\left(\frac{t}{\lambda}\right)\right]} \quad (2.9)$$

$t \geq 0 \; \lambda > 0$

Therefore

$$S(t;\lambda) = \int_t^\infty \left(\frac{1}{\lambda}\right) exp^{\left[-\left(\frac{x}{\lambda}\right)\right]} dx$$

$$= exp^{\left[-\left(\frac{x}{\lambda}\right)\right]}\Big|_t^\infty = exp^{\left[-\left(\frac{t}{\lambda}\right)\right]} \quad (2.10)$$

The hazard function, for the exponential model is

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = \left(\frac{1}{\lambda}\right) \quad (2.11)$$

$= P$(Experiencing the event of interest in the interval $(t, t + \delta_t)$|survived past time, t)

$$= P(t < T < t + \delta_t | T > t)$$

This is constant for an exponential model

3. Investigation of the Properties of the Estimates

Since the estimators for a Weibull model do not exist in a close form solutions then the estimates cannot be computed analytically. This means that the properties of the parameter estimates can only be investigated through numerical techniques.

The most fundamental and desirable properties of an estimators are;

Unbiasedness which means on average the estimates equal the true parameter they estimates, ***minimum variance*** which means that the variance of the estimates are less than that of the original true parameter, *efficiency* meaning that the expected value of the estimator is equal to the parameter it estimates and *consistency* when the estimate converge in probability to the true parameter with the increased sample size.

In this study, the biasness, Root mean square error and total deviation were calculated so as to make inference about the data simulated

The mean square error for a parameter estimates is mathematically defined by

$$MSE = E\left[\left(\hat{\theta} - \theta\right)^2\right]$$

$$= \left[Bias(\hat{\theta}, \theta)\right]^2 + Var(\hat{\theta})$$

Where;

$\theta$is the true parameter and $\hat{\theta}$ is the parameter estimates

$Bias(\hat{\theta}) = E[\hat{\theta} - \theta] = E(\hat{\theta}) - \theta$for θ the actual parameters and $\hat{\theta}$ the estimates

$Var(\hat{\theta})$can be obtained from the estimated Fisher information matrix.Total deviation for the parameter estimates of a Weibull distribution function is calculated from the expression

$$TD(\hat{\theta}, \theta) = \left|\frac{\hat{\beta} - \beta}{\beta}\right| + \left|\frac{\hat{\alpha} - \alpha}{\alpha}\right|$$

The root mean square error of a parameter estimates are then calculated by

$$RMSE(\hat{\theta}, \theta) = \sqrt{E\left[\left(\hat{\theta} - \theta\right)^2\right]} = \sqrt{\left[Bias(\hat{\theta}, \theta)\right]^2 + Var(\hat{\theta})}$$

4. Results and Discussion

By using the DFP optimization method in the MATLAB program, the parameters estimates for which value of the likelihood function is maximum are obtained. MATLAB DFP program for the parameters estimation of the distribution model is developed. The optimal estimates of the scale and shape parameters (α, and βrespectively) of the Weibull distribution are obtained by maximizing the log-likelihood function. The optimal estimates of the parameter lambda of the exponential distribution is also obtained by maximizing the log-likelihood function.

Table 4.1. Shows how discretization of the continuous data was done.

Table 4.1: Discretization of the Sample Size of 40

Continuous	2.2633	1.5728	7.1827	1.5051	3.3849	7.6281	5.6532	3.8843	1.0419	0.9453
Discrete	3	2	8	2	4	8	6	4	2	1
Continuous	6.7963	0.8638	1.0462	4.2510	2.3600	6.9870	4.6457	1.4835	2.4132	1.0167
Discrete	7	1	2	5	3	7	5	2	3	2
Continuous	3.2480	9.1273	2.0220	1.3066	3.1126	2.6335	2.7243	4.8371	3.2496	6.6427
Discrete	4	10	3	2	4	3	3	5	4	7
Continuous	2.9499	9.2834	5.6657	8.7684	7.6349	2.2037	3.0170	5.3585	1.1298	9.1765
Discrete	3	10	6	9	8	3	4	6	2	10

The null hypothesis that the two samples (continuous and discretized) come from a population with the same distribution when tested using the two-sample Kolmogorov-Smirnov test could not be rejected at 5% significance level for sample of sizes 40 and 80 but stronglyrejected for bigger sample sizes of 120 and 160.

In Fig.1, the curve for discrete and continuous data are draw for the samples of size 40 and 80. Much as the null hypothesis that both the discrete and the continuous data samples come from a population with the same distribution could not be rejected for the sample of sizes 40 and 80 at 5% when tested two-sample Kolmogorov-Smirnov test, the discrete survival curve for both samples lies far above their continuous survival counterparts. Discretization of time-to-event data therefore leads to overestimation of the survival proportion. The failure to reject the hypothesis that the two samples come from the population with the same distribution could have been because of the fewer data points in the samples.

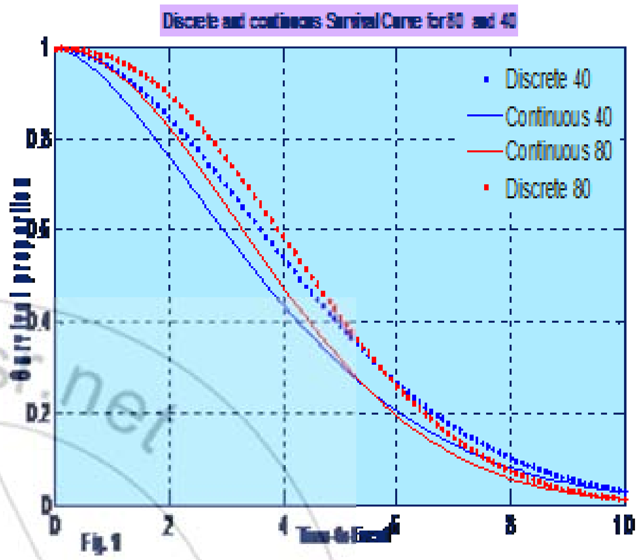

Fig.1

Table 4.2.below shows the Weibull parameter estimates at 5% level of significance for both continuous and discretized data with their corresponding percentage total deviation, confidence interval and the log likelihood values.It can be seen that the percentage deviation decreases with increase in sample size. It can also be seen that the parameter estimates is higher for discrete samples compared to that of the continuous sample. This therefore means that when prediction of the future event is made using the discrete data then the results will be misleading. For instance, Okello Omwonylee and Diongue, (2014) predicted the time when all those who were displaced by Lord Resistance Army would return to their ancestral homes but return time were counted in years which could have influenced their result.

Table 4.2: Parameter Estimates for the Weibull Model

Sample size		Continuous data			Discretized data			Deviation (%)
	Parameters	Estimates	Con. Interval		Estimates	Con. interval		
40	Scale, α	4.5109	3.6758	5.5358	5.1811	4.3535	6.1660	42.52%
	Shape, β	1.6002	1.2535	2.0429	1.8840	1.4818	2.3953	
	Log-likelihood	89.9276			90.8641			
80	Scale, α	4.6692	4.1468	5.2574	5.2801	4.7697	5.8450	30.04%
	Shape, β	1.9472	1.6358	2.3178	2.2773	1.9212	2.6994	
	Log-likelihood	172.5375			172.4529			
120	Scale, α	5.2377	4.7980	5.7177	5.7735	5.3389	6.2434	22.56%
	Shape, β	2.1489	1.8674	2.4729	2.4138	2.1027	2.7709	
	Log-likelihood	264.0610			264.0999			
160	Scale, α	4.9515	4.5834	5.3492	5.5250	5.1568	5.9190	24.74%
	Shape, β	2.1062	1.8612	2.3834	2.3657	2.0944	2.6721	
	Log-likelihood	347.0647			349.3734			

Table 4.3 below shows the exponentially distributed model parameter estimates at 5% level of significance for both the discrete and the continuous samples. Just like in the Weibull model, the parameter estimates of discretized data is higher than those of their continuous counterpart. This therefore means that when prediction of the future event is made using the discrete data then the results would be influenced. The percentage deviation for the exponential model also decreases with the increase in sample size but the deviation are much lower than those of their Weibull counterparts.

Table 4.3: Parameter Estimates for the Exponential Model

Sample size	Continuous Data				Discretized Data			Deviation (%)
	Parameters	Estimates	Con. intervals		Estimates	Con. Intervals		
40	*Lambda, λ*	4.0254	3.0201	5.6346	4.5750	3.4325	6.4038	13.65%
	Log-likelihood	95.7051			100.8243			
80	*Lambda, λ*	4.1301	3.3558	5.2086	4.6625	3.4325	5.8800	12.89%
	Log-likelihood	193.4636			203.1641			
120	*Lambda, λ*	4.6356	3.9064	5.5911	5.1083	4.3047	6.1613	10.20%
	Log-likelihood	304.0521			315.7048			
160	*Lambda, λ*	4.3946	3.7859	5.1638	4.8937	4.2159	5.7502	11.36%
	Log-likelihood	396.8619			414.0734			

5. Conclusion and Recommendation

The table 5.1. below shows the results of the Two-Sample Kolmogorov-Smirnov test results. The test results shows that the null hypothesis that the two samples (continuous and discretized) come from a population with the same distribution when tested using the two-sample Kolmogorov-Smirnov test could not be rejected at 5% level of significance for sample of size 40 and 80 except for sample of sizes of 120 and 160. This is because of the few data points in a smaller sample.

Table 5.1: Two-Sample Kolmogorov-Smirnov Test

Sample size	*Hypothesis*	*p-values*	*KSSTAT*	*Decision*
40	H=0	0.3613	0.2000	Don't reject
80	H=0	0.0708	0.2000	Don't reject
120	H=1	0.0446	0.1750	**Reject**
160	H=1	0.0041	0.1938	**Reject**

In addition, the parameter estimates for the discretised survival time is higher than that of the continuous data. This implies that discretisation of the survival time increases the value of the parameter being estimated. The Root mean square errors decrease with increase in the sample sizes but the discrete parameter estimates consistently deviated from the continuous parameter estimates. Since the discrete parameter estimates are higher than the simulated continuous parameter estimates then the study of discrete survival data overestimates the parameter of the parametric models under consideration leading into inaccurate decision and future researchers should avoid this.

Researchers studying time-to-event analysis should reduce reliance on discrete data as it ignores the richer information that the continuous data possess. Smaller counting units can be made a proportion of the bigger one and in the cases of no occurrences, it should be interpolated.

References

[1] Abernathy, R. B. (1998). *The New Weibull Handbook*. 3rd ed. SAE Publications, Warren dale. PA.
[2] Buis, M.L., (2006). An introduction to Survival Analysis, Department of Social research Methodology, Vrije Universiteit Amsterdam.
[3] Collett, D. (2003).*Modelling survival data in medical research*, second edition, Chapman and Hall/CRC
[4] Joseph. O. Okello, D. Abdou Ka, (2014), Parametric Models and Future Event Prediction Base on Right Censored Data, *American Journal of Mathematics and Statistics*, **4**(5); 205-213. (http://article.sapub.org/10.5923.j.ajms.20140405.01.html)
[5] Khan K.H, Saleem M and Mahmud. Z. (2011). Survival Proportions of CABG Patients: A New Approch. 3(**3**).
[6] Klein.P.J and Moeschberger.L.M (1997, 2003). *Survival Analysis Techniques for Censored and Truncated Data*.
[7] Kleinbaum, D.G. and Klein, M. (2005). *Survival analysis: a self-learning text*.
[8] Lawless, J.F. (2003), *Statistical models and methods for lifetime data*, second edition, Wiley-Inter-Science, A John Wiley & Sons, Inc., Publication, Hoboken, New Jersey
[9] Lawrence M. Leemis (1995). Reliability Probabilistic Model and Statistical Methods
[10] Omwonylee, J.O., Abdou KA, D. and Odongo, L.O. (2014), Modelling Internally Displaced Persons' (IDPs) Time to Resuming their Ancestral Homes after IDPs' Camps in Northern Uganda Using Parametric Methods, *International Journal of Science and Research,* **3**(5), 125-131, May 2014 http://www.ijsr.net/archive/v3i5/MDIwMTMxODE5
[11] Saleem, M., Mahmud, Z. and Khan, K.H. (2012). Survival Analysis of CABG Patients by Parametric Estimations In Modifiable Risk Factors - Hypertension and Diabetes. *American Journal of Mathematics and Statistics*; 2**(5),** 120-128