Journal of the Nigerian Mathematical Society
Vol. 31, pp. 243-254, 2012
©Nigerian Mathematical Society

A SIMULATION STUDY ON KAPLAN MEIER NON-PARAMETRIC SURVIVAL METHODS

K. A. ADELEKE

ABSTRACT. Exploring a time to event data, especially time to failure (Death) assuming some data are censored and no tied observations. This article discusses the use of Kaplan Meier non-parametric approach on mortality data simulated over some period specifically 15-weeks follow up. Series of simulations were carried out and survival and hazards probabilities and Kaplan Meier graphs were obtained for different simulations. Law of large numbers as well as demographic stochasticity in the model were observed. The findings suggested that as the number of subjects at risk increases, the expected survival $\widehat{\mu}$which is approximately 6 weeks and unconditional probability of survival $S(t)$ are almost decreasing at the same rate indicating more mortality on daily basis while the hazards $h(t)$ are tending towards being constant

Keywords and phrases: Kaplan Meier, Censored data, Non-parametric Method,Unconditional probability
2010 Mathematical Subject Classification: 62N01, 62-09 and 62Nxx

1. INTRODUCTION

The actual computation of survival probability at a given time via a risk set can be carried out using the Kaplan-Meier (KM) method. This is a non-parametric or distribution free method which are quite easy to understand and apply (Lee and Wang (2003)). They are less efficient than parametric methods when survival time follows a theoretical distribution and more efficient when no suitable theoretical distribution is known. Kaplan and Meier (1958), developed a method called Product Limit (PL) of estimating survivorship function. This method is an alternative or special case of life-table where each interval contains only one observation. Therefore Kaplan Meier is based on individual survival time. In analyzing survival data, two functions that are dependent on time are of particular interest: the survival function and the hazard

Received by the editors April 12, 2012; Revised: June 11, 2012; Accepted: June 13, 2012

function. The survival function S(t) is defined as the probability of surviving at least to time t. The hazard function h(t) is the conditional probability of dying at time t having survived to that time (Bewick et al, 2004). This method has been applied in series of articles from different fields. Mark and Hills,(1988) applied a Bayesian nonparametric approach to a (right) censored data problem. He extended the posterior distribution of percentiles given by Hill (1968) to obtain predictive posterior probabilities for the survival of one or more new patients, using data from other individuals having the same disease and given the same treatment. Abeysekera and Sooriyarachchi (2009) worked on Cox Proportional Hazards (PH) model but used Kaplan-Meier curves, a preliminary analysis on the survival data. Akram *et al* (2007), compared the Kaplan-Meier method and Weibull model based on Anderson-Darling (1954) Goodness of Fit test and this was applied to the real life time data of cancer registry in Multan, Pakistan. He however concluded that there were different sex-superiority of survival pattern among different groups of cancer patients. Interestingly, Kaplan-Meier and Weibull model provided a very close estimate of the survival function. Kaplan-Meier method was used to estimate risk of childhood mortality by household environmental health hazards, although, it is a descriptive procedure for examining the distribution of time to an event (Olufunke A. Fayehun (2010)). A Kaplan Meier method among others was used in analyzing data from studies where the response variable is the length of time taken to reach a certain end-point, often death (Bewick *et al*, 2004). The rationale and interpretation behind the use of Kaplan Meier, log-rank and Cox models were explicitly discussed by (Clark et al, 2003), they further suggested alternative methods that can be applied when either the data or a given model is deficient. It is of my opinion to work on this model, using simulation approach. Programs and codes were generated to simulate random data which was used with the help of excel package and simulation methods by Rubinstein and Kroese (2008).

2. MATERIAL AND METHODS

A useful way of characterizing the survival in a homogeneous group of individuals is to compute and graph the empirical survival function. If there are no censored observations in the sample, the empirical survival function at time t is the ratio of survivors at time t and the sample size n. This step function decreases by $1/n$

just after each observed failure (for ease of presentation we assume no ties here). When dealing with censored data, a methodology for handling this with convenience is required. Let T (survival time,$T \geq 0 : t_1, t_2, \ldots, t_n$) be the survival time of n randomly sampled individual study Such that $t_1 \leq t_2 \leq t_3, \cdots, \leq t_n$ are of $T_1, T_2, ..., T_n$. Where $S(t) \sim b(n, p)$ and $P = P(T \geq t)$ then, $S_{n(t)} \sim N(p, p(1-p)/n)$.

Where P =probability of success (survival) and $1 - p =$ probability of failure (death). Let $F(t)$ denote the cumulativae distribution of t with $f(t) \geq 0$ for all $t \geq 0$ and $f(t) = 0$ for all $t < 0$.

Then, $F(t) = [P(T \leq t) = \int_0^t f(x)dx]$

i.e probability of an individual surviving to a time t is

$$P(s) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx$$

Hence, $P(s) = S(t)$

Where

$$f(t) = \lim_{\Delta_t \to \infty} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

$$\frac{dF(t)}{dt} = \frac{-dS(t)}{dt}$$

Now let $Y_i \in (Y_i, \tau_i)$ Where $Y_i = \min(O_i, C_i)$, O_i =Observed and C_i =censored

Then,

$$\tau_i = \begin{cases} 1; \text{ if } O_i \leq C_i \\ 0; \text{ otherwise} \end{cases}$$

Recall

$$P(A \cap B) = P(B/A).P(A)$$

, We then introduce Kaplan Meier survival method as

$\widehat{S}(t_j) = \widehat{S}(t_{(j-1)} * \widehat{(Pr)}(T > t_j / T \geq t_j)$

Then $S(t) = P(T > t) = T/P_i$

We now relate this to a situation of a mortality case with time to death $t_i \leq t$,

$n_i =$ number of observation

and d_i number of death.

The probability of failure or death is given as

$$q = \frac{(d_i)}{(n_i)}$$

The conditional probability of survival

$$P_c = h(t) = 1 - q = 1 - d_i/n_i = \frac{(n_i - d_i)}{(n_i)}$$

Hence, an unconditional probability of survival is

$$P_u = \widehat{S}(t) = \prod(t_i \leq t)\widehat{P}_{c_i} = \prod(t_i \leq t) = (\frac{(n_i - d_i)}{(n_i)})$$

Expected survival which is the mean survival time

$$\widehat{\mu} = E(\widehat{S}(t)) = \int_0^\infty \widehat{S}(t)dt$$

or

$$\widehat{\mu} = \sum_{i=1}[\widehat{S}(t_{(i-1)})(t_i) - (t_{(i-1)})]$$

where $t^i = ith$ time when t is ranked for uncensored observation.

3. RESULTS

We use the above models to generate a simulated data which at the same time produced an estimates of the Survival probabilities as shown in the tables below. We set an initial probability of survival at 90 percents and probability of censored as well as loss to follow up to 10 percent. Tables (1-6) show a summary of the simulations output while Figures (1 - 6) show the Kaplan-Meier curves of the survival function used in this study of time to death with limited follow up of 15 weeks. We repeated each process at diferent sizes.

Table 1: Estimates of survival probabilities with $n = 200$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	200	17	23	-0.915	0.915	0.9	0.915
2	160	15	38	0.9063	0.8292	0.81	0.9106
3	107	6	16	0.9439	0.7827	0.729	0.9215
4	85	3	19	0.9647	0.7551	0.6561	0.9321
5	63	5	16	0.9206	0.6952	0.5905	0.9298
6	42	6	8	0.8571	0.5958	0.5314	0.9173
7	28	4	10	0.8571	0.5107	0.4783	0.9085
8	14	2	9	0.8571	0.4377	0.4305	0.9019
9	3	2	8	0.3333	0.1459	0.3874	0.8075
10	-7	3	4	1.4285	0.2084	0.3487	0.8549
11	-14	5	6	1.3571	0.2829	0.3138	0.8916
12	-25	4	12	1.16	0.3282	0.2824	0.9113
13	-41	2	9	1.0487	0.3442	0.2542	0.9212
14	-52	1	9	1.0192	0.3508	0.2288	0.9279
15	-62	1	7	1.0161	0.3565	0.2059	0.9335

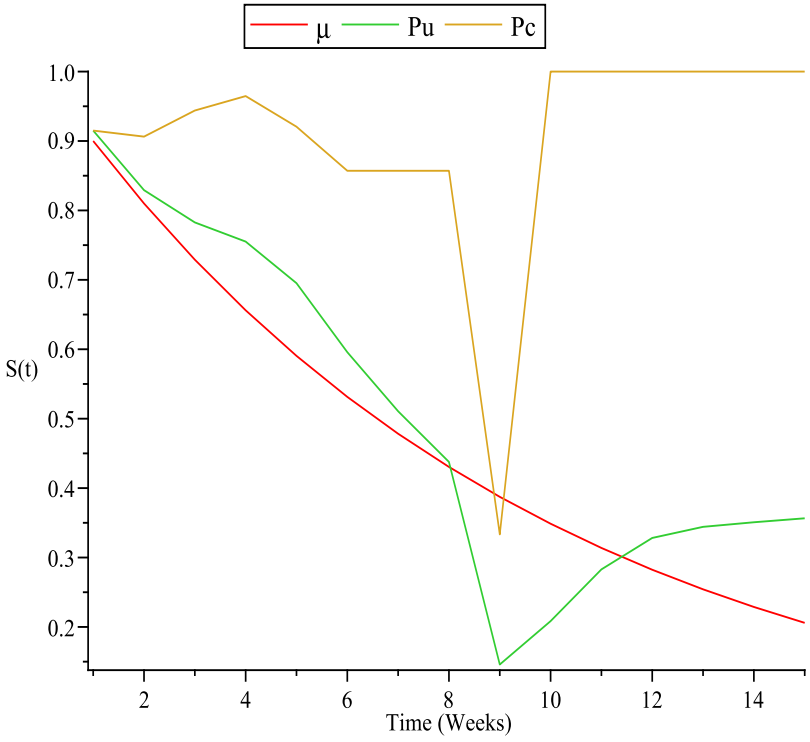

Fig. 1. Kaplan Meier Survival curve when n = 200.

Table 2: Estimates of survivals probabilities with $n = 500$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	500	45	56	0.9100	0.9100	0.9000	0.9100
2	399	34	42	0.9147	0.8324	0.8100	0.9124
3	323	29	23	0.9102	0.7577	0.7290	0.9116
4	271	23	26	0.9151	0.6930	0.6561	0.9125
5	222	18	23	0.9189	0.6371	0.5905	0.9138
6	181	18	20	0.9005	0.5738	0.5314	0.9116
7	143	15	17	0.8951	0.5136	0.4783	0.9092
8	111	18	11	0.8378	0.4303	0.4305	0.8999
9	82	11	13	0.8658	0.3726	0.3874	0.8961
10	58	11	8	0.8103	0.3019	0.3487	0.8871
11	39	6	5	0.8461	0.2555	0.3138	0.8833
12	28	10	8	0.6428	0.1642	0.2824	0.8602
13	10	7	7	0.727	0.1194	0.2542	0.8492
14	-4	6	9	0.9999	0.2985	0.2287	0.9173
15	-19	9	8	1.0000	0.4399	0.2059	0.9467

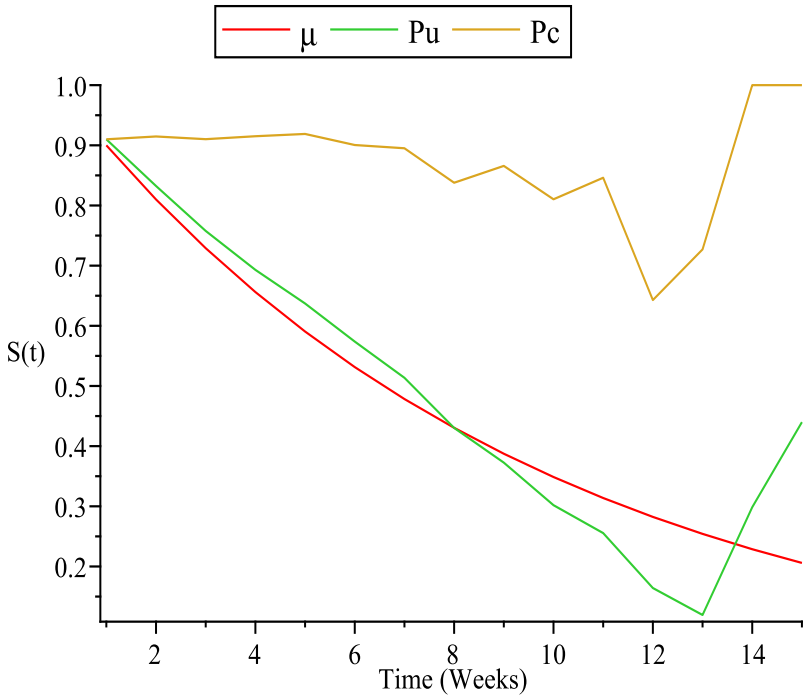

Fig. 2. Kaplan Meier Survival curve when n = 500.

Table 3. Estimates of survivals probabilities with $n = 1000$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	1000	87	101	0.913	0.913	0.900	0.913
2	812	84	95	0.897	0.819	0.8100	0.905
3	633	79	103	0.875	0.716	0.7290	0.895
4	451	85	84	0.812	0.581	0.6561	0.873
5	282	84	99	0.702	0.408	0.5905	0.836
6	99	97	95	0.020	0.008	0.5314	0.449
7	-93	90	107	1.967	0.016	0.4783	0.555
8	-290	93	107	1.321	0.021	0.4305	0.619
9	-490	94	93	1.192	0.026	0.3874	0.665
10	-677	100	101	1.147	0.029	0.3487	0.703
11	-878	81	102	1.0923	0.032	0.3138	0.731
12	-1061	88	100	1.083	0.035	0.282	0.756
13	-1249	76	107	1.061	0.037	0.2542	0.776
14	-1432	97	81	1.068	0.039	0.2288	0.7935
15	-1610	90	101	1.0559	0.0415	0.2058	0.8088

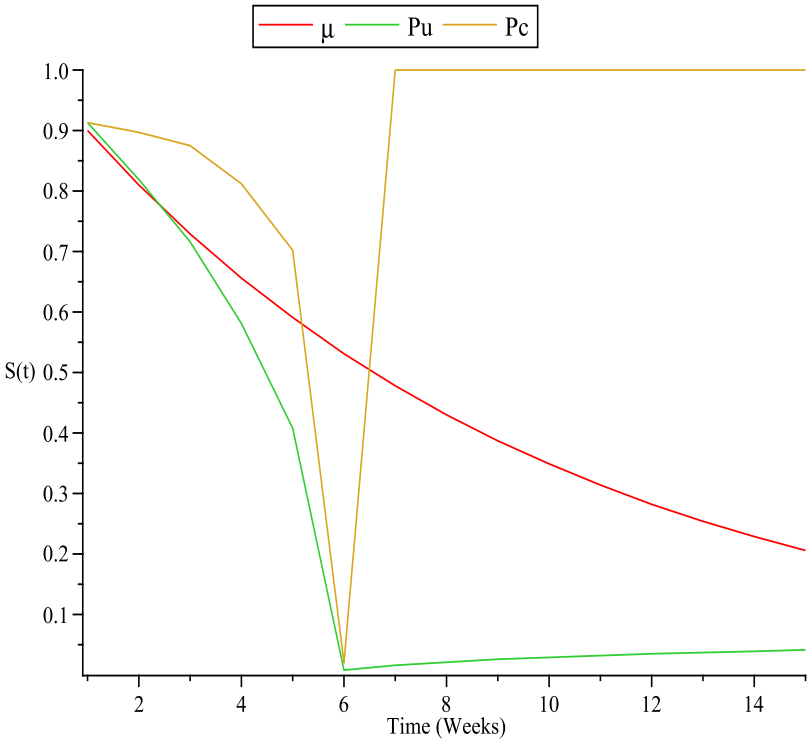

Fig. 3. Kaplan Meier Survival curve when n = 1000.

Table 4. Estimates of survivals probabilities with $n = 5000$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	5000	462	504	0.907	0.907	0.9	0.907
2	4034	342	419	0.915	0.831	0.81	0.911
3	3273	290	345	0.911	0.757	0.729	0.911
4	2638	245	272	0.907	0.687	0.656	0.91
5	2121	183	230	0.914	0.627	0.591	0.911
6	1708	169	158	0.901	0.565	0.531	0.909
7	1381	115	127	0.917	0.518	0.478	0.91
8	1139	89	121	0.922	0.478	0.431	0.912
9	929	76	101	0.918	0.439	0.387	0.913
10	752	64	91	0.914	0.401	0.349	0.913
11	597	70	71	0.883	0.354	0.314	0.909
12	456	54	57	0.882	0.312	0.282	0.908
13	345	44	40	0.872	0.273	0.254	0.905
14	261	38	27	0.854	0.233	0.229	0.901
15	196	32	32	0.838	0.195	0.206	0.897

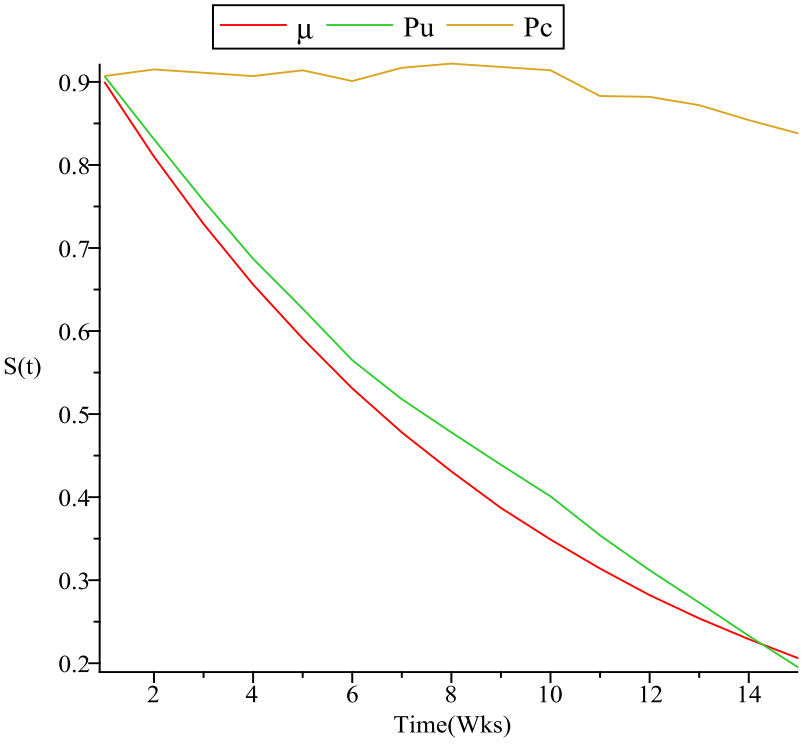

Fig. 4. Kaplan Meier Survival curve when n = 5000.

Table 5. Estimates of survivals probabilities with $n = 10000$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	10000	892	996	0.911	0.911	0.9	0.911
2	8112	703	801	0.913	0.832	0.81	0.913
3	6608	612	629	0.907‘	0.755‘	0.729	0.91
4	5367	472	553	0.912	0.688	0.656	0.911
5	4342	372	474	0.914	0.629	0.591	0.912
6	3496	326	363	0.907	0.571	0.531	0.911
7	2807	250	273	0.911	0.52	0.478	0.911
8	2284	201	248	0.912	0.474	0.43	0.911
9	1835	154	208	0.916	0.434	0.387	0.911
10	1473	126	146	0.914	0.397	0.349	0.912
11	1201	107	121	0.911	0.361	0.314	0.912
12	973	90	114	0.908	0.328	0.282	0.911
13	769	87	85	0.887	0.291	0.254	0.909
14	597	54	71	0.91	0.265	0.229	0.909
15	472	47	50	0.9	0.239	0.206	0.909

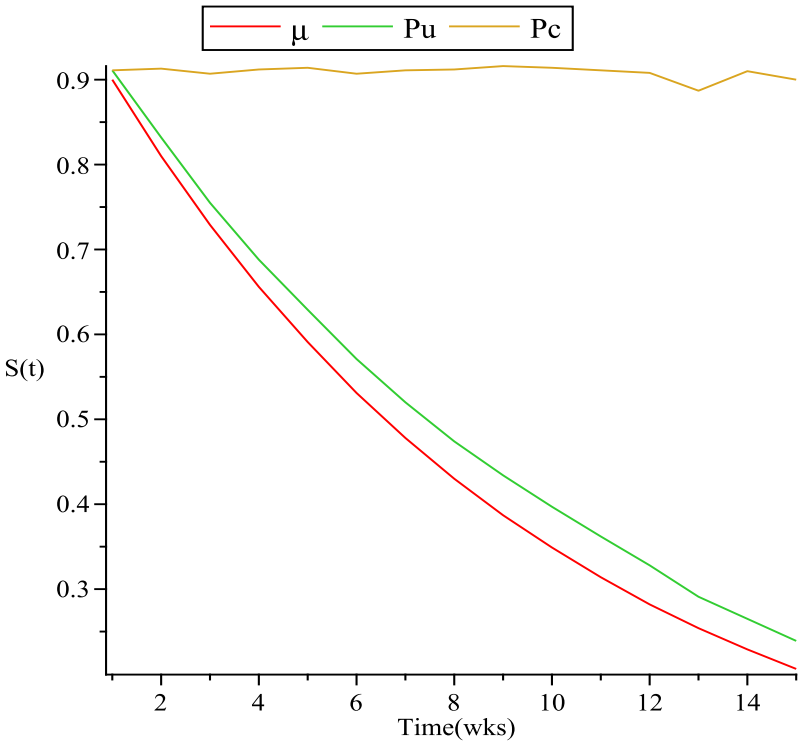

Fig. 5. Kaplan Meier Survival curve when n = 10000.

Table 6. Estimates of survivals probabilities with $n = 20000$.

Twk	R	D	C	P_c	P_u	$\widehat{\mu}$	α
1	20000	1751	2077	0.9124	0.9124	0.9	0.9124
2	16172	1483	1626	0.9083	0.8288	0.81	0.9103
3	13063	1251	1226	0.9042	0.7494	0.729	0.9083
4	10586	920	1062	0.9131	0.6843	0.6561	0.9095
5	8604	767	853	0.9109	0.6232	0.5904	0.9098
6	6984	628	720	0.9101	0.5672	0.5314	0.9098
7	5636	528	563	0.9063	0.5141	0.4783	0.90933
8	4545	426	459	0.9063	0.4659	0.4305	0.9089
9	3660	367	353	0.8997	0.4192	0.3874	0.9079
10	2940	278	297	0.9054	0.3796	0.3487	0.9077
11	2365	191	242	0.9192	0.3488	0.3138	0.9087
12	1932	185	196	0.9042	0.3155	0.2824	0.9083
13	1551	122	168	0.9213	0.2907	0.2542	0.9093
14	1261	115	129	0.9088	0.2642	0.2288	0.9093
15	1017	110	125	0.8918	0.2356	0.2059	0.9081

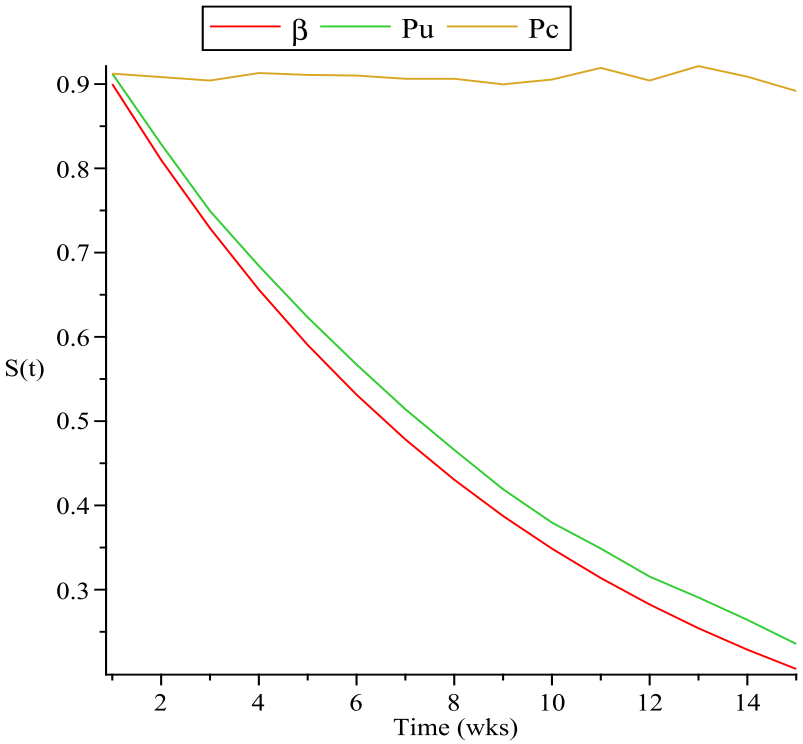

Fig. 6. Kaplan Meier Survival curve when n = 20000.

4. CONCLUDING REMARKS

Kaplan-Meier survival analysis is a non parametric statistical method, also known as the Kaplan-Meier product limit estimate or the Kaplan-Meier survival curve, can be used to estimate survival. The method has been applied broadly to measure how long it takes for any specific event to occur. Such as the time it takes until death, the time until a cancer patient recovers from a treatment, the time until an infection appears, the time until pollination occurs, and so on. The estimated survivor curves obtained are step functions however; and the graphs are interpreted the same way. Note that the expected P_u is a straight line because we set the weekly survival probability as a constant over time. Sharp drops in the P_u line indicate more mortality on a given week, and shallow drops in a line indicate fewer deaths occurring during a particular interval. We can see that as the number of subjects at risk increases (i.e sample size), the expected survival $\widehat{\mu}$ and unconditional probability of survival ($\widehat{S}(t)$) are almost dropping at the same rate indicating more mortality weekly. The estimated median survival time which is a better measure of central tendency for skewed data than arithmetic mean, and this is at t(50), i.e at S(t(50)) = 0.5 which is the time beyond which 50 percents of individuals under study are expected to survive or equivalently, the time within which 50 percents of the patients are expected to die. Going through all the graphs, we discovered that expected median survival time is approximately 6 weeks. Fig (2.0) is a scenario in which mortality rate increases from week 1 till 13th week and is almost constant for the remaining weeks under study, with an expected median survival time of six (6) weeks. Fig (3.0) is a worst scenario in which the mortality rate increases via a sharp drop in the survival probability towards zero and this remains constant for the rest of the study period, the expected median survival time is also six (6) weeks. Fig 4.0 to 6.0 where we have 5000, 10,000 and 20,000 samples of simulated trials, satisfy the law of large numbers and there is demographic stochasticity in the model in which the conditional probability of survival h(t) is constant while the unconditional and expected probability of survival ($\widehat{S}(t)$)and $\widehat{\mu}$ are decreasing almost at the same rate indicating that the mortality rates obtained for such populations are constant over the study period as simulation sizes increases and expected median survival time remain the same as six weeks.

ACKNOWLEDGEMENTS

The author would like to thank the anonymous referees whose comments improved the original version of this manuscript.

NOMENCLATURE

T	Event Time in weeks
$\widehat{\mu}$	Expected Survival Probability
α	Actual Survival Probability
P_u	Unconditional Probability
P_c	Conditional Probability
R	Number at Risk
D	Number of death
C	Number Censored

REFERENCES

[1] T. W. Anderson and D. A. Darling, *A test of goodness of fit. J*, Amer. Stat. Assoc., **49** (1): 765 - 769, 1954.

[2] E. T. Lee and J. W. Wang : *Statistical Methods for Survival Data Analysis*, John Wiley Sons, Inc., Hoboken, New Jersey, 2003.

[3] Fayehun, Olufunke A. *Household Environmental Health Hazards and Child Survival in Sub-Saharan Africa*, DHS Working Papers No. 74. Calverton, Maryland, USA: ICF Macro 2010.

[4] B. M. Hill , *Posterior Distribution of Percentiles: Bayes' Theorem for Sampling from a Finite Population*, Journal of the American Statistical Association **63** 677 - 697, 1968.

[5] E. L. Kaplan and P. Meier *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, **53**, 457 - 481, 1958.

[6] M. Akram, M. A. Ullah and R. Taj *Survival analysis of cancer patients using parametric and non-parametric approaches*, Pakistan Vet. J,**27** (4) 194 - 198, 2007.

[7] Mark Berliner L. and Bruce M. Hill, *Journal of the American Statistical Association,* **83**, (403) 772 - 779, 1988.

[8] Rubinstein and Kroese *Simulation and the Monte carlo method*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.

[9] T.G. Clark, M.J. Bradburn, S.B. Love and D.G. Altman *Tutorial Paper and Further concepts and methods in Survival Analysis.* British Journal of Cancer *89*, 781 - 786, 2003.

[10] V. Bewick1, L. Cheek1 and J. Ball; BioMed Central Ltd, http: ccforum.com/content/8/5/389, 8:389-394 (DOI 10.1186/cc2955) 2004.

[11] W. W. M. Abeysekera and M.R. Sooriyarachchi, Journal of the National Science Foundation of Sri Lanka **37**, 2009.

DEPARTMENT OF MATHEMATICS, OBAFEMI AWOLOWO UNIVERSITY, ILE - IFE, NIGERIA

E-mail address: kay7bright2002@yahoo.com