

Nonparametric Survival Function for Pneumo Carcinoma Patients

Taha A. Taha¹, Sufyan Abdulraheem Mutar²

¹Directorate-General for Education of Anbar, Ministry of Education, Iraq
Email: [tahaanwar36\[at\]gmail.com](mailto:tahaanwar36[at]gmail.com)

²Anbar Secondary School for Distinguished Boys Directorate-General for Education of Anbar, Ministry of Education, Iraq
Email: [instructuree1947\[at\]gmail.com](mailto:instructuree1947[at]gmail.com)

Abstract: *In this paper, we use a variety of nonparametric estimation techniques, based on a sample from complete real data, to estimate the survival function of lung cancer patients. This function describes the time a patient with lung cancer is expected to live after receiving a diagnosis of the disease or entering a hospital. The mean squared error was used to compare the aforementioned estimation techniques, and it was found that the shrinkage approach produced the best survival function for lung cancer.*

Keywords: Non-parametric Estimation, Lung Carcinoid tumors Disease, Actual Data, Em-Estimator

1. Introduction

The success of in the area of medical statistics, analysis is one of the most important statistical methods that is often used. It's also important in many other fields, like medicine, biology, physics, economics, engineering, epidemiology, public health, and even event history analysis in sociology, which are all science fields. The modeling of time to event data is part of the survival analysis. In the writing on survival analysis, a death or loss is called an "event." Traditionally, each subject experiences only one event, after which the organism or mechanism is either broken or dead. That presumption is relaxed in models that take into account recurring or repeated events. The study of recurring occurrences is applicable not only to the reliability of systems but also to a great deal of research in the social sciences and the medical field. To clarify, the analyses of survivor functions always include modeling of the passage of time.

Lung carcinoid tumors: Lung carcinoid tumors make up less than 5% of lung tumors. The majority of these develop slowly.

"This is to state that the study of the patient's case began with the case diagnosis and continued until the event. In the literature of survival analyses in medical experiments, the occurrence represents death [6]". Cancer is a collection of diseases in which a cell or group of cells grows uncontrollably, invades, and sometimes spreads to other parts of the body via lymph or blood (metastasis).

In 2007, it was responsible for about 13%, human beings death and 7.6 million individuals of all ages were affected by it. There are many things that can cause cancer, but 90 – 95% of cancers are caused by lifestyle and the environment. The other 5–10% are caused by genes. The most common type of cancer in the world is lung cancer, which is caused by smoking. Most types of lung cancer are more likely to happen if you smoke a lot of cigarettes every day. High amounts of air pollution, radiation, and asbestos may also make lung cancer more likely. Lung cancer symptoms contain the following:

- 1) A cough that doesn't go away and gets worse as time goes on.
- 2) There is always chest pain.
- 3) The buildings' exits cough up blood.
- 4) Shortness of breath and whistling or hoarseness of the voice.
- 5) Injuries like asthma or bronchitis that keep coming back.
- 6) The neck and face swell up.
- 7) Losing your hunger and losing weight.
- 8) Fatigue or exhaustion.

And how he is treated for lung cancer depends on what state it is in. If lung cancer is rare and not too big, surgery may be enough to get rid of it. However, radiation treatment and chemotherapy may be needed to cure the cancer or at least slow its growth.

"One of the most significant variables in the pivotal study and the manner in which data is analyzed and results are evaluated [10] is "how well the individual understands statistics."



Figure (1): Lungs tumor "X-Ray" (marked by arrow)

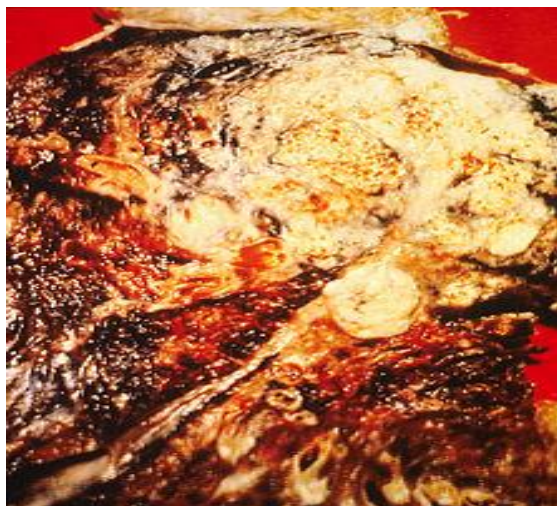


Figure (2): Transverse section of lung of human: The black zones are discoloration because the smoking, Cancer is the light area in the upper part.

In this study, we use real data from patient's lung cancer. The sample size was 100, and there were 68 men and 32 women. The years 2012 and 2013 may have killed more people with other types of cancer .

The purpose of current exploration is to calculate the survivor function for all of the aforementioned real-world results. The shrinkage technique was found to be the most accurate estimation method when compared to the others using the statistical measure of mean square error. For predicting how long a person will be able to endure lung disease.

"If we know the proportion of failures at time "t" that have already occurred, we can estimate the conditional probability of failure at time t and thus determine the underlying survival distribution S, as proposed by "Kaplan", E.L., and "Meier", P. [6]." (t).By comparing the characteristics of the various estimates, they determined that the most consistent and normal estimate was the most probable. Nelson, W. "[11]" talks about the uses and theory of a simple diagramming method called "hazard plotting" for studying "multiply censored" data, which is made up of the failure times of failed units mixed in with the running times of units that didn't fail. The technique is employed for biological data from matched comparisons with multiple censoring on both sides, tool service life data with multiple censoring events, and strength data with multiple possible failure modes. The hazard drawing technique relies on a distribution's hazard function. The reasoning behind it depends on the properties of order "statistics for Type II samples with multiple censors. To demonstrate the consistency characteristic of the Kaplan-Meier estimator

and to propose an estimator for the cumulative hazard function, Peterson, A.V .[12]" .

"The tools used in the 14 Ramadan workshops on tissues were evaluated for their reliability using Non-parameter Kaplan and Meier techniques, as described by Haifa, K. [5]". She found "no major differences between the two estimates when comparing the Kaplan-Meier technique and the reliability function when failure data followed an exponential distribution. Without access to the theoretical probability distributions, Al-Qurashi, I.K.[1] "proposed two methods for estimating the reliability function for datasets of any size. He then went on to evaluate the suggested formulas against both parametric and nonparametric alternatives".

The Shrunken Kaplan-Meier survival function was made by "Borkowf", C. B. [3]. It is based on the Kaplan-Meier survival function. The Shrunken Kaplan-Meier survival function for the study had n cases, which showed that these estimators did better than Greenwood's and Peto's. Borkowf focused solely on the variance estimators in his study .

"Mei-C. W. [9] provided a short overview of Survival analysis in biostatistics and demonstrated some non-Parametric estimation techniques that can be used in this context."

2. Basic Concepts

2.1 Survival Function C(t)

The goal is primarily the survival function, indicated by the symbol C and defined as [7]:

$$C(t) = \Pr(T > t) \quad (1)$$

Where T is a r.v., t is the death time.

The survival function $C(t)$ is the probability that the patient will survive until time t.

Survival probability is usually supposed to approach zero as age is not decreasing) i.e.;

- 1) $C(0) = 1$.
- 2) $\lim_{t \rightarrow \infty} C(t) = 0$.
- 3) $C(t)$ is non-increasing and continuous from the right side.

"The duration of survival can't be adverse, a further feature of the data on survival. [13]" .

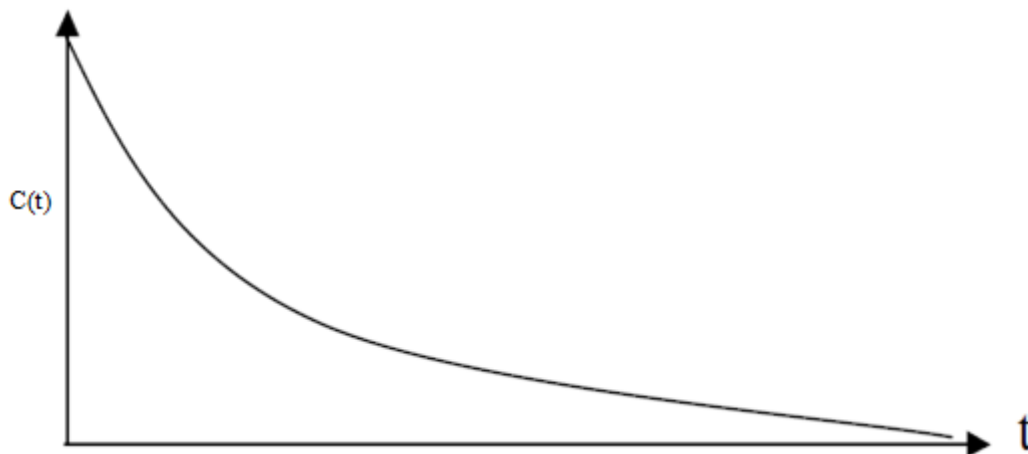


Figure 3: Refers to the survival function curve

2.2 Nonparametric approach

"Compared to parametric methods, nonparametric methods are often very easy to understand. Also, nonparametric methods are used more often when it's not clear what the exact shape of the distribution is [13]. In the present paper, we have been utilized nonparametric processes such as "Borkowf" (BE), "Empirical Function" (EM), "Thompson-e" (TB), and "Nelson" (NE).

3. Estimation method

In this section, we discuss four non- parametric estimation methods as follows:

3.1 Empirical Function (EM)

Let $D(t)$ denote the distribution of life for a specific item type, we need to evaluate the distribution function $D(t)$ and the survivor function $C(t) = 1 - D(t)$ by a complete data set of n independent lifetimes. Let $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(n)}$ be the data set organized in ascending order. The empirical distribution function is known as:-

$$D(t) = \frac{\text{Number of life time} \leq t}{n} \tag{2}$$

When we assume (that) in the data set there are no ties, we can write down the actual distribution function.

$$D(t) = \begin{cases} 0 & \text{for } t < t_{(1)} \\ \frac{i}{n} & \text{for } t_i \leq t \leq t_{(i+1)} \quad i = 1, \dots, \dots, n \\ 1 & \text{for } t_n \leq t \end{cases} \tag{3}$$

The corresponding empirical survivor function is"

$$"C(t) = 1 - D(t) = \frac{\text{Number of life time} > t}{n} "$$

(4)

When there are no ties set in the data you can also use the empirical survival function.

$$" \hat{C}(t)_{EM} = \begin{cases} 1 & \text{for } t < t_{(1)} \\ 1 - \frac{i}{n} & \text{for } t_i \leq t \leq t_{(i+1)} \quad i = 1, \dots, \dots, n \\ 0 & \text{for } t_n \leq t \end{cases} "$$

(5)

The variance of (the) empirical survivor function is"

$$" \text{Var}(\hat{C}(t))_{EM} = \frac{\hat{C}(t)_{EM}(1 - \hat{C}(t)_{EM})}{n} "$$

(6)

"If each observation is unique, $C(t)$ is a step function that decreases by $1/n$ just before each observed failure time[6]. Any ties in the data can be taken into account with an easy change. $C(t)$ is shown to be a function of t , so we have

$$" \hat{C}(t)_{EM} = 1 - \frac{i}{n} \quad i = 1, \dots, \dots, n "$$

(7)

3.2 Borkowf (BE)

According to the framework of the empirical survival function, "Borkowf" suggested a survival function[3].The expression for the Borkowf empirical survival function with n number of cases in the investigation is:

$$" \hat{C}(t)_{BE} = \frac{(n - 1)\hat{S}(t)_{EM}}{n} + \frac{1}{2n} "$$

(8)

"Borkowf" demonstrated that the variation of the suggested estimator, $C(t)$, is less than the variance of the Empirical Survival Function, $C(t)$, by using the Greenwood statistic. Greenwood's method is commonly used to calculate the standard deviation of $C(t)$. Borkowf suggested a survival function with a variance of

$$" \text{Var}(\hat{C}(t)_{BE}) = \left(\frac{n - 1}{n}\right)^2 \frac{\hat{C}(t)_{EM}^2(1 - \hat{C}(t)_{EM})}{n} "$$

(9)

4. Thompson (TE)

The method for estimating shrinkage is the Bayesian approach, which is based on prior information with regard to the value of the specific parameter gleaned from previous

experiences or studies. In this section, we must estimate C(t) given prior knowledge of S(t) as the initial value C₀(t).

Thus, "Thompson- type shrinkage estimator has the following form [15]"

$$\hat{C}(t)_{TE} = \xi \hat{C}(t)_{EM} + (1 - \xi)C_0(t), 0 \leq \xi \leq 1 \quad (10)$$

Where ξ is a shrinkage factor. $0 < \xi < 1$. Here, $C_0(t)$ is selected based on (the) Wald test statistic for $H_0 : C(t) = C_0(t)$, against $H_A: C(t) \neq C_0(t)$ with Level of significance equal to 0.05.

In this paper, we put forward the shrinkage weight function ξ as $\text{Exp}(-10/n)$.

2.4 Nelson_Aalen (NA)

Alternative estimates of the survivor function can be calculated using the individual event times and the cumulative hazard rate H(t) at time t, as in the Nelson-Aalen Estimator[11]:-

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n} \text{ for } t > 0 \quad (11)$$

Suppose that there are n individuals with observed survival times t_1, t_2, \dots, t_n . The ordered death times $t_{(i)}$, $i=1,2,\dots,n$. Where d_i is the number of individuals who die at time $t_{(i)}$.

$$H(t) = \int_0^t h(X)dx = -\text{Ln}C(t)$$

Hence, we can write the estimation of survival Nelson as follows:

$$\hat{C}(t)_{NA} = \text{Exp}(-\hat{H}(t)) \quad (12)$$

5. Methods of Estimation of Survival Function

The results of (the) estimating (of) the Survival Function (are by) using the MATLAB (2018a) program [8,14] and the four methods described above with complete data are shown in Table(1).

Table 1: Estimated Values for the Survival Function

No.	Time/d	\hat{C}_{EM}	\hat{C}_{BE}	\hat{S}_{TE}	\hat{S}_{NA}
1	3	0.9915	0.9874	0.9917	0.9916
2	37	0.9831	0.979	0.9833	0.9832
3	72	0.9746	0.9706	0.9748	0.9749
4	75	0.9661	0.9622	0.9663	0.9667
5	91	0.9576	0.9537	0.9578	0.9585
6	100	0.9492	0.9453	0.9494	0.9504
7	103	0.9407	0.9369	0.9409	0.9424
8	121	0.9322	0.9285	0.9324	0.9345
9	127	0.9237	0.9201	0.924	0.9266
10	140	0.9153	0.9117	0.9155	0.9187
11	154	0.9068	0.9033	0.907	0.911
12	156	0.8983	0.8949	0.8985	0.9033
13	164	0.8898	0.8865	0.8901	0.8957
14	186	0.8814	0.8781	0.8816	0.8881
15	211	0.8729	0.8697	0.8731	0.8806
16	212	0.8644	0.8613	0.8646	0.8732

17	213	0.8559	0.8529	0.8562	0.8658
18	217	0.8475	0.8445	0.8477	0.8585
19	218	0.839	0.8361	0.8392	0.8513
20	221	0.8305	0.8277	0.8308	0.8441
21	221	0.822	0.8193	0.8223	0.837
22	233	0.8136	0.8109	0.8138	0.8299
23	240	0.8051	0.8025	0.8053	0.8229
24	241	0.7966	0.7941	0.7969	0.816
25	243	0.7881	0.7857	0.7884	0.8091
26	249	0.7797	0.7773	0.7799	0.8022
27	254	0.7712	0.7689	0.7715	0.7955
28	266	0.7627	0.7605	0.763	0.7888
29	273	0.7542	0.7521	0.7545	0.7821
30	276	0.7458	0.7437	0.746	0.7755
31	277	0.7373	0.7353	0.7376	0.769
32	278	0.7288	0.7269	0.7291	0.7625
33	281	0.7203	0.7185	0.7206	0.756
34	290	0.7119	0.7101	0.7121	0.7497
35	301	0.7034	0.7017	0.7037	0.7433
36	301	0.6949	0.6933	0.6952	0.7371
37	301	0.6864	0.6849	0.6867	0.7308
38	302	0.678	0.6765	0.6783	0.7247
39	304	0.6695	0.6681	0.6698	0.7186
40	304	0.661	0.6597	0.6613	0.7125
41	306	0.6525	0.6512	0.6528	0.7065
42	307	0.6441	0.6428	0.6444	0.7005
43	307	0.6356	0.6344	0.6359	0.6946
44	308	0.6271	0.626	0.6274	0.6887
45	313	0.6186	0.6176	0.619	0.6829
46	313	0.6102	0.6092	0.6105	0.6772
47	314	0.6017	0.6008	0.602	0.6715
48	318	0.5932	0.5924	0.5935	0.6658
49	330	0.5847	0.584	0.5851	0.6602
50	331	0.5763	0.5756	0.5766	0.6546
51	332	0.5678	0.5672	0.5681	0.6491
52	332	0.5593	0.5588	0.5596	0.6436
53	334	0.5508	0.5504	0.5512	0.6382
54	334	0.5424	0.542	0.5427	0.6328
55	335	0.5339	0.5336	0.5342	0.6274
56	335	0.5254	0.5252	0.5258	0.6221
57	335	0.5169	0.5168	0.5173	0.6169
58	335	0.5085	0.5084	0.5088	0.6117
59	338	0.5	0.5	0.5003	0.6065
60	341	0.4915	0.4916	0.4919	0.6014
61	342	0.4831	0.4832	0.4834	0.5963
62	345	0.4746	0.4748	0.4749	0.5913
63	349	0.4661	0.4664	0.4665	0.5863
64	354	0.4576	0.458	0.458	0.5814
65	357	0.4492	0.4496	0.4495	0.5765
66	363	0.4407	0.4412	0.441	0.5716
67	364	0.4322	0.4328	0.4326	0.5668
68	364	0.4237	0.4244	0.4241	0.562
69	366	0.4153	0.416	0.4156	0.5572
70	367	0.4068	0.4076	0.4071	0.5525
71	368	0.3983	0.3992	0.3987	0.5479
72	371	0.3898	0.3908	0.3902	0.5433
73	373	0.3814	0.3824	0.3817	0.5387
74	374	0.3729	0.374	0.3733	0.5341
75	380	0.3644	0.3656	0.3648	0.5296
76	387	0.3559	0.3572	0.3563	0.5252
77	392	0.3475	0.3488	0.3478	0.5207
78	393	0.339	0.3403	0.3394	0.5163
79	397	0.3305	0.3319	0.3309	0.512
80	399	0.322	0.3235	0.3224	0.5076
81	400	0.3136	0.3151	0.314	0.5034
82	400	0.3051	0.3067	0.3055	0.4991
83	401	0.2966	0.2983	0.297	0.4949

84	402	0.2881	0.2899	0.2885	0.4907
85	407	0.2797	0.2815	0.2801	0.4866
86	409	0.2712	0.2731	0.2716	0.4825
87	419	0.2627	0.2647	0.2631	0.4784
88	421	0.2542	0.2563	0.2546	0.4744
89	421	0.2458	0.2479	0.2462	0.4704
90	422	0.2373	0.2395	0.2377	0.4664
91	422	0.2288	0.2311	0.2292	0.4625
92	423	0.2203	0.2227	0.2208	0.4586
93	427	0.2119	0.2143	0.2123	0.4547
94	428	0.2034	0.2059	0.2038	0.4509
95	430	0.1949	0.1975	0.1953	0.4471
96	446	0.1864	0.1891	0.1869	0.4433
97	450	0.178	0.1807	0.1784	0.4395
98	454	0.1695	0.1723	0.1699	0.4358
99	461	0.161	0.1639	0.1615	0.4321
100	463	0.1525	0.1555	0.153	0.4285
101	470	0.1441	0.1471	0.1445	0.4249
102	477	0.1356	0.1387	0.136	0.4213
103	481	0.1271	0.1303	0.1276	0.4177
104	481	0.1186	0.1219	0.1191	0.4142
105	483	0.1102	0.1135	0.1106	0.4107
106	483	0.1017	0.1051	0.1021	0.4073
107	497	0.0932	0.0967	0.0937	0.4038
108	511	0.0847	0.0883	0.0852	0.4004
109	512	0.0763	0.0799	0.0767	0.397
110	512	0.0678	0.0715	0.0683	0.3937
111	516	0.0593	0.0631	0.0598	0.3904
112	517	0.0508	0.0547	0.0513	0.3871
113	519	0.0424	0.0463	0.0428	0.3838
114	533	0.0339	0.0378	0.0344	0.3806
115	534	0.0254	0.0294	0.0259	0.3774
116	535	0.0169	0.021	0.0174	0.3742
117	540	0.0085	0.0126	0.009	0.371
118	550	0	0.0042	0.0005	0.3679

6. Numerical results and Conclusions

1) As expected, the values of the survival function of all estimation methods suggested in this paper have been

going down gradually as the failure times of lung cancer patients have gone up. This means that failure times and survival function are related to the opposite way. "This shows that the value of the patients' survival function was high when they were still living in the hospital and low when they died.[14] "

2) The table (2) below shows the mean squares error [14], for suggested methods of the estimating of the survival function .

Table 2: Comparing between four Non-parametric Methods

Methods	MS [$\hat{C}(t)$]
EM	0.000018020
NA	0.02920232
BK	0.000019121
TE	0.000015565

Where;

$$MSE[\hat{C}(t_i)] = \frac{\sum_{i=0}^n [\hat{C}(t_i) - C(t_i)]^2}{n} \quad (13)$$

Where $C(t_i)$ is the Median rank survival function, $\hat{C}(t_i)$ is the specific estimated survival function and n refers to the sample size of the patient.

- 3) As a consequence, the computations of the mentioned statistical indicators which are shown in the table (2) above, lead to the result that the mean squares error(MSE) for Thompson estimator (TE) method are less than those of the EM, BE and NA methods, so the shrinkage Method is the best estimation method.
- 4) By observing the figure (4) below, one can note the matching of the proposed estimation methods in this paper and the extent of convergence resulting accuracy of these methods, especially to real Median rank survival function methods C(t).

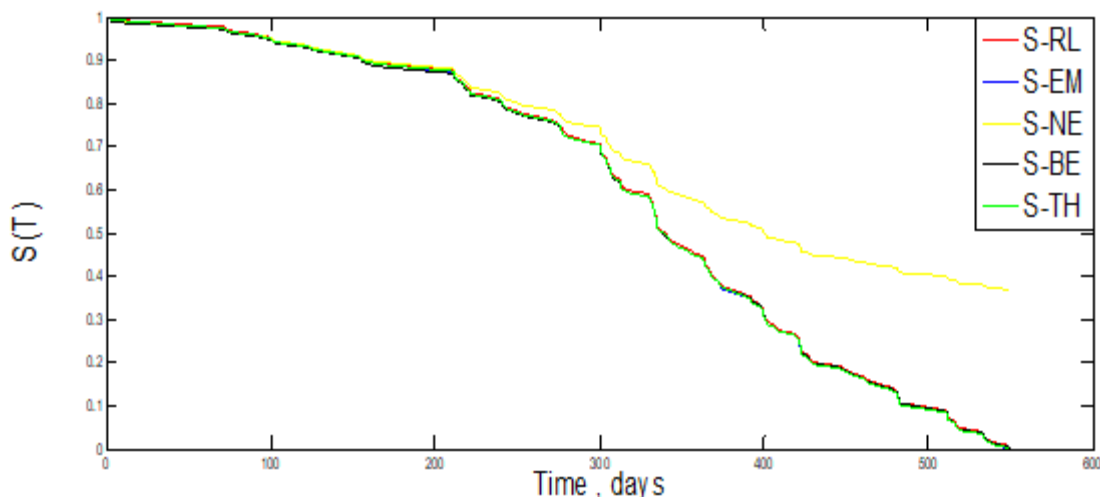


Figure 4: Shows the curve of four used estimation methods for the survival function

References

[1] AL- Qurashi , A . K. (2001). Estimate survival function for nonparametric methods Ph.D. thesis in statistics submitted to the Faculty of Management and Economics at the University of Mustansiriya.

[2] American Cancer Society (December 2007). "Report sees 7.6 million global 2007 cancer deaths". Reuters. Retrieved 2008-08-07.

[3] Borkowf, C. B. (2005). A simple hybrid variance estimator for the Kaplan-Meier survival function. *Statistics in Medicine*, Vol. 24; 827-851.

- [4] Basher, F. M.(2010).Some Of The Parametric Methods And Nonparametric to Estimate the reliability function With The Practical Application. Master thesis, Baghdad University, College of Administration and Economics
- [5] Haifa, K. (1987).Use Non-Parametric Method to Find Reliability Function For the Tools of 14 Ramadan factory - Department of textile . Master Thesis, Baghdad University, College of Administration and Economics.
- [6] Kaplan, E.L. and Meier, P. (1958). Non Parametric Estimation from Incomplete Observations .*Journal of the American Statistical Association*. 53. 457-481.
- [7] Marvin. R and Arnljot .H. (2004) . System Reliability Theory Models, Statistical Methods and Applications . Second Edition.
- [8] Mathews J. H. And Fink K. D. (2003)," Numerical Method Using MATLAB", Third Edition, Prentice Hall, USA.
- [9] Mei-C,W. (2006) .Summary Notes for Survival Analysis. Department of Biostatistics. Johns Hopkins University.
- [10] National Collaborating Centre for Cancer (2011), " The diagnosis and treatment of lung cancer (update) ". <http://www.nice.org.uk/nicemedia/live/13465/54199/54199.pdf>
- [11] Nelson, W.(1972).Theory and application of hazard plotting for censored failure Data ." *Techno metrics* 14:945-966.
- [12] Peterson , A.V.(1977) . Expressing the Kaplan-Meier estimator as a function of empirical sub-survival function . *JASA*.72,854-858.
- [13] Qamruz, Z. and Karl, P. (2011) ,Survival Analysis Medical Research. <http://interstat.statjournals.net/YEAR/2011/abstracts/1105005.php>.
- [14] Taha , A. T (2013)," Estimate the Parameters and Related Probability Functions for Data of the Patients of Lymph Glands Cancer via Birnbaum- Saunders Model", M.Sc., Baghdad University, Education College for Pure Sciences(Ibn Al-Haitham) .
- [15] Thompson , J.R. (1968) . Some Shrinkage Techniques for Estimating the Mean . *J. Amer.*