International Mathematical Forum, 1, 2006, no. 38, 1861-1869

Asymptotic Efficiencies of the Survival Functions Estimators for the Exponential Distribution

Ahmed A. Abu-Taleb

Department of Mathematics and Statistics
Jordan University of Science and Technology
P.O. Box 3030, 11220 Irbid, Jordan
abutaleb@just.edu.jo

Mahmoud M. Smadi

Department of Mathematics and Physics
University of Qatar
P.O. Box 2713 Doha, Qatar

Abstract In medical statistics, the survival function is a relationship between proportion and time. Proportion is the proportion of subjects which are still surviving at time, t. The term is also used in other fields and is known as "units still operating" instead of subjects still alive. In this paper, an estimator of the survival time, X_i, for the i^{th} patient on a clinical trial with censoring time, T_i (dropping out of the trial) and its properties, when both survival and censoring time are exponentially distributed, considered. A simulation is carried out to determine the performance of the estimators for different combinations of parameters related to the survival and censoring times.

Keywords: Survival function, time censoring, root mean square errors (RMSE) simulation

1. **Introduction**

Let $X_1, X_2, \cdots, X_n$ be i.i.d. positive random variables with unknown survival function $S_1(x) = P_\Theta(X_1 > x)$. Also, let $T_1, T_2, \cdots, T_n$ be i.i.d. random variables with unknown survival function $S_2(t) = P_\Theta(T_1 > t)$. Assume that all the X_i's and T_i's are independent variables. A randomly censored data set consists of n i.i.d pairs (Y_i, D_i), where $Y_i = \min(X_i, T_i)$ and $D_i = I(X_i \leq T_i)$ for $i = 1, 2, 3, \cdots, n$. In the context of survival analysis and reliability, X_i refers to the survival time and T_i refers to the censoring time.

Several statistical approaches and scenarios for this problem can generally be used according to the model and type of information available, Collett [3]. Turnbull [6] provided a nonparametric estimation of a survival function with doubly censored data, i.e., some of the data are censored on the left and some on the right. Suzuki[4,5] studied the right censored case and in these papers both parametric and nonparametric estimates of survival are proposed. Bravo et al. [2] presented a semi-nonparametric estimation of a survival function when analyzing incomplete and doubly data. Abu-Taleb [1] provided a parametric estimation of the survival function when the follow ups are random.

In this paper, an estimator of the survival time, X_i, for the i^{th} patient on a clinical trial with censoring time, T_i, (dropping out of the trial) and it asymptotic properties, assuming exponential distribution for both variables are considered.

In Section 2 we give a full description of the problem, while in Section 3 we use the method of maximum likelihood to estimate the parameters. An estimator of the survival function along with a study of its asymptotic properties is given in section 4. A simulation study for the performance of the estimators for different combinations of parameters related to the survival and censoring times is finally given in Section 5.

2. **Description Of The Problem**

Let $X_1, X_2, \cdots, X_n$ be i.i.d. positive random variables with unknown survival function $S_1(x) = P_\Theta(X_1 > x)$. Also, let $T_1, T_2, \cdots, T_n$ be i.i.d. random variables with unknown survival function $S_2(t) = P_\Theta(T_1 > t)$. Assume that all the X_i's and T_i's are independent variables.

A randomly censored data set consists of n i.i.d pairs (Y_i, D_i), where $Y_i = \min(X_i, T_i)$ and $Di = I(X_i \leq T_i)$ for $i = 1, 2, 3, \cdots, n$. In the context of survival analysis and reliability, X_i refers to the survival time and T_i refers to the censoring time. Assume that X_i, the survival time, $i = 1, \cdots, n$, and T_i, censoring time, are independent exponentially distributed with probability density functions given by:

$$f_{X_i}(x_i, \theta) = \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \qquad x_i > 0, \quad \theta > 0,$$

and

$$f_{T_i}(t_i, \lambda) = \frac{1}{\lambda} e^{-\frac{t_i}{\lambda}} \qquad t_i > 0, \quad \lambda > 0,$$

respectively, and suppose we observe $[(Y, D); i = 1, \cdots, n]$, then the pdf of D_i is:

$$P(D_i = 1) = P(X_i \leq T_i) = \int_0^\infty \int_{x_i}^\infty \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \frac{1}{\lambda} e^{-\frac{t_i}{\lambda}} dt_i dx_i$$
$$= \frac{\lambda}{\lambda + \theta}.$$

And, hence,

$$P(D_i = 0) = \frac{\theta}{\lambda + \theta}.$$

It can be shown that the joint probability density function of (Y_i, D_i) is:

$$f_{Y_i,D_i}(y_i, d_i, \theta, \lambda) = \left(\frac{1}{\theta} + \frac{1}{\lambda}\right) e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y_i} \left(\frac{\lambda}{\lambda+\theta}\right)^{d_i} \left(\frac{\theta}{\lambda+\theta}\right)^{1-d_i}, \; y_i > 0;\; \theta, \lambda > 0.$$

Recalling that D_i and Y_i are independent, which also clear the joint probability density function, based on the joint probability density function of (Yi, Di), we see that $(\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i)$ are joint sufficient statistics for $\Theta = (\theta, \lambda)$. Also Y_i is exponentially distributed with parameter $\frac{1}{\frac{1}{\theta}+\frac{1}{\lambda}}$, we have $\sum_{i=1}^n Y_i$ has a Gamma distribution with parameters n and $\frac{1}{\frac{1}{\theta}+\frac{1}{\lambda}}$. Also, Di has a Bernoulli distribution with parameter $\frac{\lambda}{\lambda+\theta}$, therefore $\sum_{i=1}^n D_i$ has a Binomial distribution with parameters n and $\frac{\lambda}{\lambda+\theta}$.

The joint probability density function of $(\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i)$ can be written as

$$f_{\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i}(s, u) = \frac{s^{n-1} e^{-s\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)}}{n!} \left(\frac{1}{\theta} + \frac{1}{\lambda}\right)^n \binom{n}{u} \left(\frac{\lambda}{\lambda+\theta}\right)^u \left(\frac{\theta}{\lambda+\theta}\right)^{n-u}.$$

The logarithm of the joint probability density function of $(\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i)$ is equal to

$$\ln f_{\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i}(s, u) = (n-1)\ln s - s\left(\frac{1}{\theta} + \frac{1}{\lambda}\right) - n\ln\lambda + u\ln\left(\frac{\lambda}{\theta}\right)$$
$$+ \ln\binom{n}{u} - \ln n!$$

which constitutes an exponential family form. Therefore, $(\sum_{i=1}^n Y_i, \sum_{i=1}^n D_i)$ are complete statistics for $\Theta = (\theta, \lambda)$.

3. **Estimation Of The Parameters**

Based on the notations and assumptions listed in Section 1, the likelihood function, and log-likelihood functions are given by

$$L(\theta,\lambda)=\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)\sum_{i=1}^{n}y_i}\left(\frac{\lambda}{\theta}\right)^{\sum_{i=1}^{n}d_i}\left(\frac{\theta}{\theta+\lambda}\right)^{n},$$

and

$$L^*(\Theta)=\ln L(\Theta)=-n\ln\lambda-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)\sum_{i=1}^{n}y_i+\left(\sum_{i=1}^{n}d_i\right)\ln\lambda-\left(\sum_{i=1}^{n}d_i\right)\ln\theta,$$

respectively. Therefore the ML equations:

$$\frac{\partial}{\partial\theta}\ln L(\theta,\lambda)=\frac{1}{\theta^2}\sum_{i=1}^{n}y_i-\frac{1}{\theta}\sum_{i=1}^{n}d_i=0$$

$$\frac{\partial}{\partial\lambda}\ln L(\theta,\lambda)=-\frac{n}{\lambda}+\frac{1}{\lambda^2}\sum_{i=1}^{n}y_i-\frac{1}{\lambda}\sum_{i=1}^{n}d_i=0.$$

Solving the above equations for $\hat{\theta}$ and $\hat{\lambda}$, one obtains:

$$\hat{\theta}=\frac{\sum_{i=1}^{n}Y_i}{\sum_{i=1}^{n}D_i}\quad\text{and}\quad\hat{\lambda}=\frac{\sum_{i=1}^{n}Y_i}{n-\sum_{i=1}^{n}D_i},$$

which are the maximum likelihood estimators of θ and λ.

The information matrix, whose entries are found from the second derivatives of the log-likelihood function can be written as:

$$I(\Theta)=\begin{bmatrix}\frac{\lambda}{\theta^2(\lambda+\theta)} & 0\\ 0 & \frac{\theta}{\lambda^2(\lambda+\theta)}\end{bmatrix}.$$

Therefore, the asymptotic variance-covariance matrix of the parameters, i.e. the inverse of the information matrix, can be expressed as:

$$\Sigma=I^{-1}(\Theta)=\begin{bmatrix}\frac{\theta^2(\lambda+\theta)}{\lambda} & 0\\ 0 & \frac{\lambda^2(\lambda+\theta)}{\theta}\end{bmatrix}.$$

4. **Estimation Of The Survival Functions And Their Asymptotic Properties**

The maximum likelihood estimators of the survival functions, $S_1(x)$ and $S_2(y)$, which are defined as the probability that the survival time is greater than or equal to x and the censoring time which is greater than or equals to y, respectively. For fixed x and y, let

$$H(\Theta)=[S_1(x),S_2(y)]=[P_\Theta(X_1>x),P_\Theta(Y_1>y)].$$

By the invariance property of the ML, the ML estimators of $S_1(x)$ and $S_2(y)$ are:

$$\hat{P}_\Theta(X_1 > x) = \int_x^\infty \frac{1}{\theta} e^{-\frac{z}{\theta}} dz = e^{-\frac{x}{\theta}},$$

and

$$\hat{P}_\Theta(Y_1 > y) = \int_y^\infty \left(\frac{1}{\theta} + \frac{1}{\lambda}\right) e^{-z\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)} dz = e^{-y\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)}.$$

Thus the ML maximum likelihood estimator of $H(\Theta)$ can be expressed as:

$$\hat{H}(\Theta) = \left[e^{-\frac{x}{\theta}}, e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y}\right].$$

Now, we derive the asymptotic distribution of the estimator $H(\Theta)$. Under the regularity conditions, we have (Zaks [7]):

$$\sqrt{n}\left[H\left(\hat{\Theta}\right) - H(\Theta)\right] \to N\left(0, \frac{\partial H^T(\Theta)}{\partial \Theta} \Sigma \frac{\partial H(\Theta)}{\partial \Theta^T}\right).$$

A partial derivation of $H(\Theta)$ with respect to θ and λ, respectively, gives:

$$\frac{\partial}{\partial \theta} \hat{H}(\Theta) = \left[\frac{x}{\theta^2} e^{-\frac{x}{\theta}}, \frac{y}{\theta^2} e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y}\right]$$
$$\frac{\partial}{\partial \lambda} \hat{H}(\Theta) = \left[0, \frac{y}{\lambda^2} e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y}\right],$$

so,

$$\frac{\partial}{\partial \Theta} \hat{H}(\Theta) = \begin{bmatrix} \frac{x}{\theta^2} e^{-\frac{x}{\theta}} & \frac{y}{\theta^2} e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y} \\ 0 & \frac{y}{\lambda^2} e^{-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y} \end{bmatrix}.$$

Thus, the asymptotic variance-covariance matrix of $H(\Theta)$ is given by:

$$\begin{bmatrix} \frac{x^2}{\theta}\left(\frac{1}{\theta}+\frac{1}{\lambda}\right) e^{-\frac{2x}{\theta}} & \frac{xy}{\theta}\left(\frac{1}{\theta}+\frac{1}{\lambda}\right) e^{\frac{x}{\theta}-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y} \\ \frac{xy}{\theta}\left(\frac{1}{\theta}+\frac{1}{\lambda}\right) e^{\frac{x}{\theta}-\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y} & y^4\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)^2 e^{-2\left(\frac{1}{\theta}+\frac{1}{\lambda}\right)y} \end{bmatrix}.$$

Based on the asymptotic distribution of the survival functions estimators further statistical inference such as confidence intervals and hypothesis testing could be performed for the survival functions and their parameters.

5. **Simulation Study**

In order to study the performance of the estimators of θ, λ, $P_\Theta(X_1 > x_s)$, and $P_\Theta(Y_1 > y_s)$ discussed in this paper, a simulation was carried out. Different sample sizes are considered for each of different combinations of parameter values (θ, λ). The means and root mean square errors (RMSE) of the maximum likelihood estimates $\hat{\theta}$ and $\hat{\lambda}$ of θ and λ are calculated. The results are shown in Tables (1)-(3). The probabilities $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ are estimated by $\hat{P}_\Theta(X_1 > x_s)$ and $\hat{P}_\Theta(Y_1 > y_s)$ respectively . The values of x_s

and y_s are taken respectively, $x_s = \theta$, and $y_s = \frac{\theta+\lambda}{\theta\lambda}$ (i.e. x_s = the means of survival time, X_i and y_s = the mean of $Y_i = \min(X_i, T_i)$). The RMSE of the maximum likelihood estimates $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ of $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ are calculated. The results are shown in Tables (4)-(6). The simulation results are based on 1000 replicates. Also sample sizes of 50, 150, and 300 are considered for different combinations of parameter.

As one expects the performance of the maximum likelihood estimators depends on the two intensity rates θ and λ, which are the means of the survival and censoring times. We note as the mean of the survival and censoring times increases, the corresponding MSE increases. On the other hand we note that if the mean of survival time increases and the mean of the censoring time decreases, the MSE of the MLE of the survival time increases and the MSE of the MLE of the censoring time decreases and visa versa.

Concerning the survival functions $S_1(x) = P_\Theta(X_1 > x)$ and $S_2(t) = P_\Theta(T_1 > t)$ for different parameters values, we note that the MSE of the MLE of the survival function for the survival time $S_1(x) = P_\Theta(X_1 > x)$ is smaller than the MLE of the survival function for the censoring time $S_2(t) = P_\Theta(T_1 > t)$. Finally, as we expect the MSE of the maximum likelihood estimators of survival times and survival functions decreases as the sample size increases values (θ, λ).

Table 1. Means and RMSE of MLE's of θ and λ, $n = 50$.

θ	$\hat{\theta}$	RMSE	λ	$\hat{\lambda}$	RMSE
0.1	0.10111	0.015542	0.5	0.567478	0.264757
0.5	0.562456	0.297004	0.1	0.100449	0.015444
0.5	0.511913	0.109627	0.5	0.511484	0.106877
1	1.002304	0.148033	4	4.399987	1.729204
4	4.414233	1.787153	1	1.007165	0.16433
4	4.048353	0.8549	4	4.070881	0.827237
4	4.090776	0.862862	4	4.049434	0.826238
4	4.045578	0.661026	10	10.40202	3.011581
10	10.55578	3.300795	4	4.003439	0.669263
10	10.19181	2.178379	10	10.10456	2.123669
10	10.15783	1.63826	40	43.46129	17.31276
40	43.56587	17.06	10	10.16452	1.582584
40	40.7813	8.063063	40	41.03706	8.585889

TABLE 2. Means and RMSE of MLE's of θ and λ, $n = 150$.

θ	$\hat{\theta}$	RMSE	λ	$\hat{\lambda}$	RMSE
0.1	0.100542	0.008815	0.5	0.515962	0.1061
0.5	0.51845	0.11821	0.1	0.10021	0.009341
0.5	0.503781	0.060934	0.5	0.503526	0.056763
1	1.002878	0.093249	4	4.062795	0.744112
4	4.111007	0.811753	1	0.999543	0.089084
4	4.01707	0.460751	4	4.021614	0.458327
4	4.020148	0.466896	4	4.028791	0.490258
4	4.002785	0.373985	10	10.1524	1.638401
10	10.2323	1.639423	4	4.001404	0.400689
10	10.09034	1.231656	10	10.01969	1.157952
10	10.06355	0.935365	40	40.84993	7.74527
40	40.92254	7.871092	10	10.00206	0.927289
40	40.3951	4.764727	40	40.55375	4.888066

TABLE 3. Means and RMSE of MLE's of θ and λ, $n = 300$.

θ	$\hat{\theta}$	RMSE	λ	$\hat{\lambda}$	RMSE
0.1	0.100065	0.006409	0.5	0.507198	0.073456
0.5	0.510063	0.076157	0.1	0.100095	0.006371
0.5	0.500872	0.041033	0.5	0.500238	0.041417
1	1.001582	0.066177	4	4.041907	0.533885
4	4.061445	0.533951	1	1.000656	0.066544
4	4.010444	0.320553	4	4.01876	0.33547
4	4.000656	0.333633	4	4.031202	0.33439
4	4.000358	0.284218	10	10.09713	1.109816
10	10.07806	1.113677	4	4.009069	0.267618
10	10.01877	0.810227	10	10.01472	0.81808
10	10.0058	0.663954	40	40.73375	5.522239
40	40.4094	5.33873	10	10.01759	0.624763
40	40.09864	3.329273	40	40.08545	3.295385

Table 4. Means and RMSE of MLE's of $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ of different θ and λ values, $n = 50$.

θ	λ	$P_\Theta(X_1 > x_s)$	$\hat{P}_\Theta(X_1 > x_s)$	RMSE	$P_\Theta(Y_1 > y_s)$	$\hat{P}_\Theta(Y_1 > y_s)$	RMSE
0.1	0.5	0.36788	0.367507	0.057217	0.0000000	0.0000000	0.0000000
0.5	0.1	0.36788	0.382223	0.133084	0.0000000	0.0000000	0.0000000
0.5	0.5	0.36788	0.365734	0.074618	0.0000001	0.0000007	0.0000039
1	4	0.36788	0.366515	0.054966	0.2096114	0.2087356	0.0432931
4	1	0.36788	0.383007	0.121161	0.2096114	0.2076473	0.0447864
4	4	0.36788	0.366895	0.071317	0.7788008	0.7749553	0.0290675
4	4	0.36788	0.365567	0.073083	0.7788008	0.7741493	0.0284945
4	10	0.36788	0.364418	0.061914	0.8847059	0.8822215	0.0158969
10	4	0.36788	0.374988	0.098438	0.8847059	0.8826497	0.0157815
10	10	0.36788	0.369659	0.072198	0.9607894	0.9603564	0.005624
10	40	0.36788	0.364759	0.054973	0.9844964	0.9842027	0.0021299
40	10	0.36788	0.38283	0.12082	0.9844964	0.9841439	0.0022181

Table 5. Means and RMSE of MLE's of $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ of different θ and λ values, $n = 150$.

θ	λ	$P_\Theta(X_1 > x_s)$	$\hat{P}_\Theta(X_1 > x_s)$	RMSE	$P_\Theta(Y_1 > y_s)$	$\hat{P}_\Theta(Y_1 > y_s)$	RMSE
0.1	0.5	0.36788	0.367262	0.033405	0.0000000	0.0000000	0.0000000
0.5	0.1	0.36788	0.379396	0.075424	0.0000000	0.0000000	0.0000000
0.5	0.5	0.36788	0.368011	0.044049	0.0000001	0.0000002	0.0000004
1	4	0.36788	0.367295	0.033488	0.2096114	0.2095786	0.0269176
4	1	0.36788	0.374325	0.06891	0.2096114	0.2094073	0.0266363
4	4	0.36788	0.36847	0.041586	0.7788008	0.7779495	0.0159655
4	4	0.36788	0.367072	0.042497	0.7788008	0.7774342	0.0162354
4	10	0.36788	0.366867	0.03489	0.8847059	0.8842716	0.0086566
10	4	0.36788	0.367446	0.056532	0.8847059	0.8834462	0.0094427
10	10	0.36788	0.368676	0.043685	0.9607894	0.9606651	0.0032383
10	40	0.36788	0.366243	0.033075	0.9844964	0.9843694	0.0012701
40	10	0.36788	0.369692	0.067789	0.9844964	0.9843791	0.0012982

References

[1] A. A. Abu-Taleb,. Estimation of lifetime distribution from incomplete data and follow ups. Developments in Applied Statistics, Metodoloski zvezki, 19, (2003), 85-95.

[2] J. I. Bravo, I. D. Fuentes, and A. J. Fernández, A semi-parametric estimation of a survival function from incomplete and doubly-censored data, Communications in Statistics-Theory and Methods, 24 (11):(1995), 2735-2752.

[3] D. Collett, Modelling Survival Data in Medical Research, Chapman and Hall, London, 1994.

[4] K. Suzuki, Nonparametric estimation of lifetime distributions from a record of failures and follow-ups, Journal of the American Statistical Association, 80,(1985a), 68-72.

[5] K. Suzuki,. Estimation of lifetime parameters from incomplete field data. Technometrics, 27, (3), (1985b), 263-271.

TABLE 6. Means and RMSE of MLE's of $P_\Theta(X_1 > x_s)$ and $P_\Theta(Y_1 > y_s)$ of different θ and λ values, $n = 300$.

θ	λ	$P_\Theta(X_1 > x_s)$	$\hat{P}_\Theta(X_1 > x_s)$	RMSE	$P_\Theta(Y_1 > y_s)$	$\hat{P}_\Theta(Y_1 > y_s)$	RMSE
0.1	0.5	0.36788	0.367003	0.023185	0.0000000	0.0000000	0.0000000
0.5	0.1	0.36788	0.372664	0.052679	0.0000000	0.0000000	0.0000000
0.5	0.5	0.36788	0.36611	0.029951	0.0000001	0.0000002	0.0000002
1	4	0.36788	0.368404	0.024648	0.2096114	0.2099557	0.0198531
4	1	0.36788	0.369668	0.047731	0.2096114	0.2101963	0.0188624
4	4	0.36788	0.367301	0.029755	0.7788008	0.7777269	0.0111647
4	4	0.36788	0.368126	0.029129	0.7788008	0.7785916	0.0107288
4	10	0.36788	0.36694	0.02553	0.8847059	0.884172	0.0065905
10	4	0.36788	0.370859	0.03839	0.8847059	0.8847547	0.0061888
10	10	0.36788	0.36851	0.030616	0.9607894	0.9607708	0.0022444
10	40	0.36788	0.36784	0.023863	0.9844964	0.9844325	0.0009108
40	10	0.36788	0.3694	0.048289	0.9844964	0.9844344	0.0009031

[6] B. W. Turnbull, Nonparametric estimation of a eurvivorship eunction with doubly censored data, Journal of the American Statistical Association, 69,(345),(1974), 169-173.

[7] S. Zacks, The Theory of Statistical Inference, John Wiely New York 1971.

Received: March 24, 2006