Applications and Applied Mathematics: An International Journal (AAM)

Volume 10 | Issue 1 Article 5

6-2015

Estimation of the parameters of Exponential distribution using top-K-lists

M. Ahsanullah
Rider University

Mohamed Habibullah
Northeastern University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Vol. 10, Issue 1 (June 2015), pp. 71 - 80

Applications and Applied Mathematics:
An International Journal
(AAM)

Estimation of the parameters of Exponential distribution using top-*K*-lists

M. Ahsanullah
Department of Management Sciences
Rider University
Lawrenceville, New Jersey, USA

Mohamed Habibullah
Department of Supply Chain and Information Management
Northeastern University
Boston, Massachusetts, USA

Received: March 28, 2014; Accepted: March 16, 2015

Abstract

This paper deals with the estimation of location and scale parameters of the exponential distribution based on top-*k*-list of a sequence of observations from a two parameter exponential distribution. The minimum variance unbiased estimates of the location and scale parameters are given. Some comparisons of the variances of these estimates with respect to that of the k^{th} record values are given.

Keywords: Top-*k*-list; k^{th} records; Order statistics; Best linear unbiased estimates; Exponential distribution; Location parameter; Scale parameter

MSC 2010: 62H12, 62 G 30

1. Introduction

The top-*k*- lists are used in many areas such as sports, technology and social sciences. Lopez-Blazquez and Wesolowski (2007) introduced the top-*k*-list of sequence of random vectors and elaborated the usefulness of such data. They developed the distributions of top-*k*- lists and their properties arising from different specific probability distributions, such as, exponential distribution and uniform distribution on (0,1). Top-*k*-lists are used in the field of information retrieval (Fagin et al. (2003)), in sports and meteorology (Kozan and Tanil (2013)). Tanil (2009) also studied the joint and marginal distributions of top- k -lists obtained from a sequence of independent and identically distributed (iid) random variables with cumulative distribution function (cdf) F. In this paper, we use the top-*k*-list of a sequence of two parameter exponential random variables to estimate the location and scale parameters of the exponential distribution. Ahsanullah et al. (2012) used top-*k*-scores for characterization of distributions. Razmkhah and Ahmadi (2012) studied frequent inference based on top-*k*- scores.

Suppose $X_1, X_2, \ldots, X_k$ be k iid random variables with cdf $F(x)$ and the probability density function (pdf) as $f(x)$. Let $X_{1,k.} < X_{2,k} < \ldots < X_{k,k}$ be corresponding ordered set. We write *ord* $(X_1, X_2 \ldots X_k) = (X_{1,k.} < X_{2,k.} \ \ldots, < X_{k,k.}) = L_0^k$. Dziubdziela and Kopocinsky (1976) (see also Ahsanullah (2004)) introduced the *k*-record times as follows:

$$T_0^{(k)} = k,$$

$$T_{n+1}^{(k)} = \{ j \mid j > T_n^{(k)} \text{ and } X_j > X_{T_{n-k+1}^{(k)}, T_n^{(k)}} \}, \ n \geq 0.$$

The random variable $R_n^{(k)} = X_{T_{n-k+1,}^{(k)} T_n^{(k)}}$ is the n^{th} *k*-record. The pdf $f_{R_n^{(k)}}(x)$ of $R_n^{(k)}$ is defined as

$$f_{R_n^{(k)}}(x) = \frac{k^{n+1}}{\Gamma(n+1)} [-ln(1-F(x))]^n (1-F(x))^{k-1} f(x) \tag{1}$$

Lopez-Blaquez and Wesolowski (2007) defined the top-*k*-list $L_n^{(k)}, n \geq 0$, where n is the number of times the lowest value of the top-*k*-values is replaced as follows. Let $\{X_i, i = 1,2, \ldots\}$ be a sequence of iid random variables with cdf $F(x)$ and pdf $f(x)$. The n^{th} Top-*k*-list is defined as follows:

$$L_n^{(k)} = \{Y_{1,n}, Y_{2,n}, \ldots, Y_{k,n}\}, \ n \geq 0,$$

where

$$Y_{j,n} = X_{T_{n-k+j,}^{(k)} T_n^{(k)}}, \ j = 1, 2, 3, \ldots, k \text{ and } n = 0, 1, 2, \ldots.$$

Note that $Y_{1,n} = R_n^{(k)}$.

The evolution of the list is as follows:

On the 0^{th} top-k-list, $L_0^k = (X_{1.k}, X_{2,k}, \ldots, X_{k,k})$. The list remains unaltered until time T_1^k. At this moment, the first element of the top-*k*-list L_0^k is removed and the random variable $X_{T_1^{(k)}}$ enters the list. Then,

$$L_0^k = (X_{1.k}, X_{2,k}, ..., X_{k,k}).$$

The process continues in a similar way.

For *n* > *1*, an (*n*-1)*st* top-*k*-list L_{n-1}^k remains unaltered until the n^{th} k^{th} record time $T_n^{(k)}$ occurs. Then, we get,

$$L_n^{(k)} = ord\{X_{T_n^{(k)}}, Y_{2,n-1}, ..., Y_{k,n-1}\}$$

and the process is continued.

Example: In order to clarify top- *k*-lists, let us consider the following sequence of observations.

0.51 0.49 0.94 0.55 0.68 0.54 0.92 0.88 0.35 1.22 0.88 0.35 1.42 0.81

Table 1. The top *k* values *k*= 1, 2, 3

n	*k*=1	*k*=2	*k*=3
0	0.51	0.49, 0.51	0.49, 0.51, 0.94
1	0.94	0.51, 0.94	0.51, 0.55, 0.94
2	1.22	0.55, 0.94	0.55, 0.68, 0.94
3	1.42	0.68, 0.94	0.68 ,0.92, 0.94
4		0.92, 0.94	0.88, 0.92, 0.94
5		0.94, 1.22	0.92, 0.94, 1.22
6		1.22, 1.42	0.94, 1.22, 1.42

Let us consider the top-*k*-list from the exponential distribution. We then estimate the location and scale parameters of the exponential distribution using n^{th} top-*k*-observations. We show that the estimates obtained by using top-*k*-list of random variables, are unbiased and the variance of the estimate of the location parameter is less than the variance of the location parameter obtained by k^{th} order statistics for a range of values of *k* and *n*.

2. Main Results

The random variable *X* is distributed as a two parameter exponential if it's cdf *F* is given by

$$F(x) = 1 - e^{-(x-\mu)/\sigma}, x > \mu, \sigma > 0. \quad (2)$$

We denote by $X \sim E(\mu, \sigma)$, if the cdf of *X* is as given in (2).

Let $L_n^{(k)} = \{Y_{1,n}, Y_{2,n}, ..., Y_{k,n}\}$ be the n^{th} top-k-list in a sequence of $\{X_i,\ i=1,\ 2,\ \ldots\}$ iid random variables. Then, the joint pdf of $Y_{1,n}, Y_{2,n}, ..., Y_{k,n}$ is given by Lopez-Blazquez and Wesolowski (2007) as

$$f_{L_n^{(k)}}(y_{1,}\ y_2, \ldots, y_k) = \frac{k!k^n}{\Gamma(n+1)}\left(\frac{y_1-\mu}{\sigma}\right)^n \frac{1}{\sigma^k} e^{-\sum_{i=1}^{k}\left(\frac{y_i-\mu}{\sigma}\right)}. \quad (3)$$

With μ=0 and σ=1, the pdf of $Y_{j,n}$ is given by

$$f_{Y_{j,n}}(y) = \frac{k^n\, k}{\Gamma(n)\ \Gamma(j)\ \Gamma(k-j+1)}\ e^{-(k-j+1)y} A,$$

where

$$A = \int_0^y x^{n-1}(e^{-x} - e^{-y})^{j-1}\ dx.$$

It is also known (Lopez- Blazquez and Wesolowski (2007)) that

$$Y_{j,n} \stackrel{d}{=} R_n^{(k)} + W_{j,k}, \quad (4)$$

where $W_{j,k}$, is the jth order statistics, j=1,2,..., k, of k iid random variables from $E(\mu,\sigma)$, $R_n^{(k)}$ is the n^{th} k record from iid sequence of random variables from $E(\mu,\sigma)$ and $\stackrel{d}{=}$ denotes the equality in distribution. Ahsanullah and Nevzorov (2005) showed that

$$\frac{W_{j,k}-\mu}{\sigma} \stackrel{d}{=} \frac{V_1}{k} + \frac{V_2}{k-1} + \cdots + \frac{V_j}{k-j+1}, \quad j=1,\ 2,\ldots,k,\ k \geq 1, \quad (5)$$

where $V_1, V_2, .., V_k$ are iid random variables from $E(0,1)$ and that

$$\frac{R_j^{(k)}-\mu}{\sigma} \stackrel{d}{=} \frac{V_1 + V_2 + \cdots + V_j}{k}. \quad (6)$$

Since $R_n^{(k)}$ and $W_{j,k}$ are independent, using (5) we obtain

$$E(Y_{i,n}) = \mu + \sigma\left[\frac{n}{k} + \sum_{l=1}^{i}\frac{1}{k-l+1}\right], \quad i=1,\ 2,\ \ldots,\ k,$$

$$Var(Y_{i,n}) = \sigma^2\left[\frac{n}{k^2} + \sum_{l=1}^{i}\frac{1}{(k-l+1)^2}\right], \quad i=1, 2, ..., k$$

and

$$Cov(Y_{i,n}, Y_{j,n}) = \sigma^2\left[\frac{n}{k^2} + \sum_{l=1}^{i}\frac{1}{(k-l+1)^2}\right], \ 1 \le i < j \le k.$$

Let $X=(Y_{1,n}, Y_{2,n}, \ldots, Y_{k.n})$. Then, $Var(X) = \sigma^2 V$, $V= (V_{i,j})$ and $V^{-1} = (V^{i,j})$ for $1 \le i \le k$ and $1 \le j \le k$.

It can be shown that

$$V^{1,1} = \frac{k^2}{n+1} + (k-1)^2,$$

$$V^{i,j} = 0, \text{ for } |i-j| \ge 2,$$

$$V^{i,j} = V^{j,i}, \text{ for all } i \text{ and } j.$$

$$V^{i,i} = (k-i+1)^2 + (k-i)^2, \ i>1,$$

$$V^{i,i+1} = -(k-i)^2, \ i = 1, 2, \ldots, k\text{-}1, \ j > i,$$

and

$$V^{k,k} = 1, \ V^{k-1,k} = -1 = V^{k,k-1}.$$

3. Best Linear Unbiased Estimates (BLUE) of μ and σ

Let $E(X) = A\theta$, where

$$\theta = \binom{\mu}{\sigma}, \ A' = \binom{L'}{B'}, \quad L' = (1,1,\ldots,1), \ 1 \times k \text{ vector},$$

$$B' = \left(\frac{n}{k} + \frac{1}{k}, \frac{n}{k} + \frac{1}{k} + \frac{1}{k-1}, \cdots, \frac{n}{k} + \frac{1}{k} + \frac{1}{k-1} + \cdots + 1\right), \ 1 \times k \text{ vector},$$

and

$$A' = \begin{bmatrix} L' \\ B' \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \frac{n}{k} + \frac{1}{k} & \frac{n}{k} + \frac{1}{k} + \frac{1}{k-1} & \cdots & \frac{n}{k} + \frac{1}{k} + \frac{1}{k-1} + \cdots + 1 \end{bmatrix}.$$

Then, $E(X) = A\theta$, where A is a $k \times 2$ matrix, θ is 2×1 matrix and

$$A=\begin{bmatrix} 1 & \frac{n}{k}+\frac{1}{k} \\ 1 & \frac{n}{k}+\frac{1}{k}+\frac{1}{k-1} \\ \vdots & \vdots \\ 1 & \frac{n}{k}+\frac{1}{k}+\frac{1}{k-1}+\cdots+1 \end{bmatrix}, \quad A'V^{-1}=\begin{bmatrix} \frac{k^2}{n+1} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \end{bmatrix},$$

$$A'V^{-1}A=\begin{bmatrix} \frac{k^2}{n+1} & k \\ k & n+k \end{bmatrix}, \quad \text{and} \quad (A'V^{-1}A)^{-1}=\frac{n+1}{k^2(k-1)}\begin{bmatrix} n+k & -k \\ -k & \frac{k^2}{n+1} \end{bmatrix}.$$

We have

$$A'V^{-1}X=\begin{bmatrix} \frac{k^2}{n+1}Y_{1,n} \\ \sum Y_{j,n} \end{bmatrix}, \qquad \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (A'V^{-1}A)^{-1}\, A'V^{-1}X \,.$$

On simplification, we obtain

$$\begin{aligned}\hat{\mu} &= \frac{n+1}{k^2(k-1)}\left(\frac{(n+k)k^2}{n+1}Y_{1,n} - k\sum_{j=1}^{k}Y_{j,n}\right) \\ &= \frac{n+k}{(k-1)}Y_{1,n} - \frac{n+1}{k(k-1)}\sum_{j=1}^{k}Y_{j,n}\,, \end{aligned} \tag{7}$$

and

$$\begin{aligned}\hat{\sigma} &= \frac{n+1}{k(k-1)}\left(\frac{k}{n+1}\sum_{j=1}^{k}Y_{j,n} - \frac{k^2}{n+1}Y_{1,n}\right) \\ &= \frac{1}{(k-1)}\sum_{j=1}^{k}Y_{j,n} - \frac{k}{k-1}Y_{1,n}\,, \end{aligned} \tag{8}$$

as estimates of μ and σ, respectively. It can be shown that

$$E(Y_{1,n}) = \mu + \frac{n+1}{k}\sigma \text{ and } \sum_{j=1}^{k} E(Y_{j,n}) = k\left[\mu + \sigma\frac{n+k}{k}\right].$$

The expected values, variances and covariances of the estimators are

$$E(\hat{\mu}) = \frac{n+1}{k^2(k-1)}\left\{\frac{(n+k)k^2}{n+1}(\mu + \frac{n+1}{k}\sigma) - k^2(\mu + \frac{n+k}{k}\sigma)\right\} = \mu, \tag{9}$$

$$E(\hat{\sigma}) = \frac{n+1}{k(k-1)}\left\{\frac{k^2}{n+1}(\mu + \frac{n+k}{k}\sigma) - \frac{k^2}{n+1}(\mu + \frac{n+1}{k}\sigma)\right\} = \sigma, \tag{10}$$

$$Var\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} = \sigma^2(A'V^{-1}A)^{-1} = \sigma^2\frac{n+1}{k^2(k-1)}\begin{bmatrix} n+k & -k \\ -k & \dfrac{k^2}{n+1}\end{bmatrix}, \tag{11}$$

$$Var(\hat{\mu}) = \frac{(n+1)(n+k)}{k^2(k-1)}\sigma^2,$$

$$Var(\hat{\sigma}) = \frac{\sigma^2}{k-1},$$

and

$$Cov(\hat{\mu},\hat{\sigma}) = -\frac{(n+1)\sigma^2}{k(k-1)}.$$

Table 2. $Var(\hat{\mu})/\sigma^2$ based on top-*k*- statistics

n *k* →	2	3	4	5	6	7	8	9
5	10.5	2.667	1.125					
6	14	3.5	1.4583	.77				
7	18	4.444	1.833	.96	.5778			
8	22.5	5.5	2.25	1.17	.70	.45918		
9	27.5	6.6667	2.7083	1.4	.8333	.5442	.3794	

For fixed *k*, the variance of $\hat{\mu}$ increases as *n* increases and for fixed *n*, the variance of $\hat{\mu}$ is decreasing as *k* increases and *n* represents the number of times the lowest value of the top-*k*-values is replaced.

It can be noted that the best linear unbiased estimates $\hat{\mu}_{kRv}$, $\hat{\sigma}_{kRv}$ of μ and σ respectively obtained by using ordinary upper k records, $X(1)$, $X(2)$, …, $X(k)$. Note also that $X(i)$, i=1,2, …, k is the ordinary upper k records (see Ahsanullah (1980)).
The estimates are given by

$$\hat{\mu}_{kRv} = \frac{kX(1) - X(k)}{k-1} \text{ and } \hat{\sigma}_{kRv} = \frac{X(k) - X(1)}{k-1},$$

and the corresponding variances and covariance are

$$Var(\hat{\mu}_{kRv}) = \frac{k}{k-1}\sigma^2, \; Var(\hat{\sigma}_{kRv}) = \frac{\sigma^2}{k-1} \text{ and } Cov(\hat{\mu}_{kRv}\hat{\sigma}_{kRv}) = -\frac{\sigma^2}{k-1}.$$

4. Simulation

We have done a simulation study to see the closeness of the estimates by top-k-lists to their actual values. We used an exponential distribution with μ =2 and σ =3. We generated 100 random numbers and took the first 10 random numbers. Here k=10 and ordered these 10 numbers from highest to the lowest. This set of ten ordered numbers correspond to stage n = 0. We replaced the lowest of this set by the next higher number from the sequence of 100 numbers. Thus we have a new set of ten numbers which we order from the highest to the lowest again, and call this set as our sample corresponding to stage n=1. We replace the lowest number of this set by the next higher number from the sequence and we repeat it for stage n = 14. Thus we have a 15th top-10-list of random numbers from $Exp(\mu =2, \sigma =3)$. Using these ordered observations at the 15th stage, we estimated $\hat{\mu}$ and $\hat{\sigma}$ by applying the Equation (7) and Equation (8). We iterated this process 500 times and calculated the means of $\hat{\mu}$ and $\hat{\sigma}$. This process was repeated ten times. Then we took the average of these ten estimates to get the estimates based on 5000 iterations. The following table gives the means of $\hat{\mu}$ and $\hat{\sigma}$ based on 500 iterations and also their averages based on ten repetitions.

Table 3. Means of $\hat{\mu}$ and $\hat{\sigma}$ based on 500 iterations

Replications	Means of $\hat{\mu}$	Means of $\hat{\sigma}$
1	2.071350163	2.971460732
2	1.891786691	3.070190976
3	2.016901412	2.979510834
4	2.039152659	3.002757395
5	1.978798158	3.01242891
6	2.116496357	2.948089071
7	2.006137579	3.026332456
8	1.889094056	3.109977774
9	2.071745848	3.006487822
10	1.923312803	3.059529718
Average	2.000477573	3.018676569

Based on the 5000 iterations, we see that the estimate of μ is $\hat{\mu}$ = 2.000477573 and the estimate of σ is $\hat{\sigma}$ = 3.018676569 which are very close to μ(= 2) and σ (= 3), respectively.

5. Conclusion

We compared the $Var(\hat{\mu})$ based on top-*k*-values with the $Var(\hat{\mu}_{kRv})$ based on *k*- record values for some values of *n* and *k*. It is observed that $Var(\hat{\mu}) < Var(\hat{\mu}_{kRv})$, if $k^3 > (n+1)(n+k)$.
The following Table 4 gives the maximum value of *n* for some selected values of *k* that satisfy this inequality.

Table 4. The Maximum value of *n* for some selected values of *k* that satisfies the inequality $Var(\hat{\mu}) < Var(\hat{\mu}_{kRv})$

k	4	5	6	7	8	9	10
n	5	8	11	14	18	22	26

This table shows that if we decide to take a sample based on top-10 observations, we can replace the lowest value of the top-*k*-values maximum of *n* = 26 times that will guarantee $Var(\hat{\mu}) < Var(\hat{\mu}_{kRv})$.

Acknowledgements

The authors thank the referees and the editor for helpful comments and recommendations that improved the paper.

REFERENCES

Ahsanullah, M. (1980). Linear Prediction of Record values for the Two Parameter Exponential Distribution, Annals of the Institute of Statistical Mathematics, Vol.32, no. 3A, 363-368.
Ahsanullah, M. (2004). Record Values-Theory and Methods, University Press of America Inc., Lanham, MD, USA, 2003.
Ahsanullah, M., Hamedani, G.G. and Wesolowski, J. (2012). Linearity of regression inside top-*k* lists and related characterization, Studi Sci Math, Hungarica, vol. 49 (4), 436-453.
Ahsanullah, M. and Nevzorov, V.B. (2005). Order Statistics-Examples and Exercises, Nova Publishers Inc., New York, NY, USA.
Dziubdziela, W. and Kopocinsky, B. (1976). Limiting properties of the *k*-record values, Zastos, Mat. 15, No. 2, 87-189.
Lopes- Blazquez. F. and Wesolowski, I. (2007). Top k- List, Metrika, 65, 69-82.
Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top *k*-Lists, Siam J. Discrete Mathematics, 17(1): 134-160.

Kozan, A. and Tanil, H. (2013). Bottom–*k*-Lists, Journal of the Turkish Statistical Association, 6(2): 73-79.

Razmkhah, M. and Ahmadi, J. (2012). Statistical Inference based on the top scores, Metron, vol. LXX, 89-107.

Tanil, H. (2009). An order statistics based on the list of top m scores after lth change, Journal of Statistical Planning and Inference, 139, 2189-2195.