広島工業大学紀要研究編
第45巻（2011）pp.281-289
論　　　　　文

A Comparison of Different Procedures Relating to Parameter Estimation in Extreme Value Type I Distribution Through Simulation

Satyabrata PAL*, Sanpei KAGEYAMA** and Subhababa PAL***

(Received Aug. 31, 2010)

Abstract

The extreme value (EV) distribution is widely used for fitting and analyzing a long range of data emanated from different real-life situations. Annual maximum water-flow data collected from a river over a considerable period of time (say, about 30 years) are found to follow the EV distribution very satisfactorily. Also the distribution of annual maximum rainfall data in a region closely follows the EV distribution. Thus these distributions have important applications in the works related to the flood frequency estimation/estimation of return period flood. The estimation of parameters from these types of data (data following the EV type I with two parameters, EV I(2), distribution) has important use in the field of hydrology to understand the situation more deeply. Here an investigation is made through simulation for several pairs of values of a and b (parameters of the EV I(2) distribution) using different sample sizes in order to compare the accuracy of maximum likelihood method as well as method of moments in estimating the original parameters (the pairs of values of a and b for which the simulation is done). A module is developed in SAS/IML which deals with a random number generation in the EV I(2) distribution and the subsequent estimation of parameters with method of moments and maximum likelihood method. It is found that when the parameter value of a is less than or equal to 0.5, the method of moments produce more accurate estimators of the parameters than the corresponding maximum likelihood estimators for large samples (250 or more). An error distribution investigation with respect to the parameters is also made with CAPABILITY and QQPLOT procedures in SAS/QC. It is found that when the sample size is small (10, 100, etc.) the error distribution follows near normality but with the increase in sample sizes (200, 300, etc.) the error distribution gradually deviates much away from normality. Some procedures in SAS/BASE and SAS/STAT are used in the calculation leading to the final results.

Key Words: EV I(2) distribution, Moment estimator, MLE, Simulation, Normality.

1. Introduction

The generalized extreme value (GEV) distribution is a flexible three-parameter model that combines the Gumbel, Fréchet and Weibull maximum extreme value distributions, which has a probability density function (pdf) as

* Swami Vivekananda Institute of Management and Computer Science, Karbala, Sonarpur, Kolkata 700103, West Bengal, India
** Department of Environmental Design, Hiroshima Institute of Technology, Hiroshima 731-5193, Japan
*** Department of e-Learning, Sikkim Manipal University, Bangalore, Karnataka, India

$$f(x)=\begin{cases}\dfrac{1}{\sigma}(1+kz)^{-\frac{1}{k}-1}\exp\left[-(1+kz)^{-\frac{1}{k}}\right] & when \quad k\neq 0,\\ \dfrac{1}{\sigma}\exp[-z-\exp(-z)] & when \quad k=0,\end{cases}$$

where $z=(x-\mu)/\sigma$, and k, σ, μ are the shape, scale, location parameters respectively (cf. Gumbel, 1958). The scale parameter σ must be positive, the shape and location parameters can take on any real value. The range of definition of the GEV distribution depends on k and $1+kz>0$. In what follows, σ is replaced by a, and μ is replaced by b. Thus a pdf of the GEV distribution is of the form

$$f(x)=\frac{1}{a}\left[1+k\left(\frac{x-b}{a}\right)\right]^{-\frac{1}{k}-1}\exp\left\{-\left[1+k\left(\frac{x-b}{a}\right)\right]^{-\frac{1}{k}}\right\}.$$

The range of the variable x depends on the sign of the parameter k. The GEV distribution function is of the form

$$F(x)=\exp\left\{-\left[1+k\left(\frac{x-b}{a}\right)\right]^{\frac{1}{k}}\right\},\ \text{a>0, b>0.}$$

When $k=0$, we have a pdf of the GEV distribution given by

$$f(x)=\frac{1}{a}\exp\left\{-\left(\frac{x-b}{a}\right)-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\},\ -\infty<x<\infty,$$

which is here called a pdf of the extreme value type I (with two parameters) distribution, denoted by EV I(2) distribution. The distribution function of the EV I(2) distribution is shown by

$$F(x)=\exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}.$$

The random samples are generated through simulation from the EV I(2) distribution considering the values of the parameters, a and b, lying in different ranges. To achieve the above, random numbers from the EV I(2) distribution are drawn (in case of sample sizes of 10, 50, 100, 150, 200, 250, and 300 when simulated samples are created 200 times against each size) for the above parameters values of which lie in different ranges. Based on the above sets of samples, the estimators obtained by the method of moments and by the method of maximum likelihood are found out and subsequently, the differences of the estimated values of the parameters from the original values of parameters are calculated. Finally, appropriate conclusion in regard to the accuracy of the different parameter estimation procedures is presented with reference to an important feature revealed when the parameters lie in the range of specific values. Table 1 presents a complete list of parameter values and sample sizes used for the simulation purpose.

2. Materials and methods

2.1 Parameter estimation by method of moments

The first moment of the EV I(2) distribution is given by $\mu_1'=b+0.5772157a$ (cf. Kite, 1977), where the value 0.5772157 is an approximation of Euler's constant. The second central moment is given by

$$\mu_2=(\frac{\pi^2}{6})a^2.$$

The parameter estimates are obtained by replacing μ_1' and μ_2 by their corresponding sample estimates m_1' and m_2. Thus the expressions of the above estimators reduce respectively to

$$\hat{a}_{MOM}=\frac{\sqrt{6}}{\pi}\sqrt{m_2}=0.7797\sqrt{m_2}, \tag{1}$$

Table 1: Detailed list of parameters and sample sizes for simulation

Study	Parameters		Sample sizes							
	a	b	10	50	100	150	200	250	300	
1	2	5	200	200	200	200	200	200	200	Number of repetitions of the simulation with fixed parameters (i.e., fixed values of a and b) and with a specified sample size (i.e., generating a fixed number of random samples from the EV1(2) distribution with specified values of the parameters repeatedly)
2	0.5	10	200	200	200	200	200	200	200	
3	2	0.5	200	200	200	200	200	200	200	
4	0.5	5	200	200	200	200	200	200	200	
5	0.4	0.5	200	200	200	200	200	200	200	
6	5	0.8	200	200	200	200	200	200	200	
7	0.9	2	200	200	200	200	200	200	200	
8	1	2	200	200	200	200	200	200	200	
9	0.7	2	200	200	200	200	200	200	200	
10	0.8	2	200	200	200	200	200	200	200	
11	0.6	2	200	200	200	200	200	200	200	
12	0.55	2	200	200	200	200	200	200	200	
13	0.53	2	200	200	200	200	200	200	200	
14	0.52	2	200	200	200	200	200	200	200	
15	0.51	2	200	200	200	200	200	200	200	
16	0.505	2	200	200	200	200	200	200	200	
17	0.5	2	200	200	200	200	200	200	200	
18	0.495	2	200	200	200	200	200	200	200	
19	0.49	2	200	200	200	200	200	200	200	
20	0.48	2	200	200	200	200	200	200	200	
21	0.45	2	200	200	200	200	200	200	200	
22	0.4	2	200	200	200	200	200	200	200	
23	0.4	10	200	200	200	200	200	200	200	
24	0.35	2	200	200	200	200	200	200	200	

$$\hat{b}_{MOM} = m_1' - 0.45005\sqrt{m_2}. \tag{2}$$

The estimates of the parameters obtained by the method of moments from the simulated random samples generated from the EV I(2) distribution are calculated using (1) and (2). Random samples are generated by an SAS/IML module and the method of moments estimates are also estimated by the same module in SAS/IML.

2.2 Parameter estimation by maximum likelihood method

The likelihood function for a sample of size n from an EV I(2) distribution is given by

$$L = \frac{1}{a^n}\exp\left\{-\sum_{i=1}^{n}\left(\frac{x_i-b}{a}\right)-\sum_{i=1}^{n}\exp\left[-\left(\frac{x_i-b}{a}\right)\right]\right\}$$

which leads to the following

$$\log L = -n\log a - \frac{1}{a}\sum_{i=1}^{n}(x_i-b)-\sum_{i=1}^{n}e^{-\left(\frac{x_i-b}{a}\right)}. \tag{3}$$

Differentiating (3) with respect to a and b and then equating to zero, we get the following equations

$$\frac{d\log L}{da} = -\frac{n}{a}+\frac{1}{a^2}\sum_{i=1}^{n}(x_i-b)-\frac{1}{a^2}\sum_{i=1}^{n}(x_i-b)\exp\left[-\left(\frac{x_i-b}{a}\right)\right]=0,$$

$$\frac{d\log L}{db} = \frac{n}{a}-\frac{1}{a}\sum_{i=1}^{n}\exp\left[-\left(\frac{x_i-b}{a}\right)\right]=0,$$

which ultimately reduce to the following equation

$$G(a)=\sum_{i=1}^{n}x_i e^{-\frac{x_i}{a}}-\left(\frac{1}{n}\sum_{i=1}^{n}x_i-a\right)\sum_{i=1}^{n}e^{-\frac{x_i}{a}}=0. \tag{4}$$

Equation (4) on a cannot be solved analytically and thus it is to be solved iteratively. An initial value of a is required and $\hat{a}_{MOM}$ is taken as the initial value of a. The value of a is updated by the following formula

$$a_{k+1} = a_k - G(a_k)/G'(a_k) \tag{5}$$

where $G'(a)$ is the derivative of $G(a)$ in Equation (4). The iteration in Equation (5) is repeated until $G(a)$ is sufficiently close to zero. Hence $\hat{a}_{ML}$ is calculated after several iterations making G(a) sufficiently close to zero. After $\hat{a}_{ML}$ is obtained, $\hat{b}_{ML}$ is calculated from the following equation

$$\hat{b}_{ML} = \hat{a}_{ML} \log\left[\frac{n}{\sum_{i=1}^{n} e^{-\frac{x_i}{\hat{a}_{ML}}}}\right].$$

In fact, $\hat{a}_{ML}$ and $\hat{b}_{ML}$ are also calculated by the same SAS/IML module.

Briefly, following the above procedure the random samples are generated from the EV I(2) distribution at first and from the samples the estimates $\hat{a}_{MOM}$, $\hat{b}_{MOM}$, $\hat{a}_{ML}$ and $\hat{b}_{ML}$ subsequently, defining the modules in SAS/IML through coding, are obtained.

The following estimates are generated as outcomes:

$\hat{a}_{MOM(k)}$ = method of moments estimate of a at the k-th sample drawing, k = 1 (1) 200,

$\hat{b}_{MOM(k)}$ = method of moments estimate of b at the k-th sample drawing, k = 1 (1) 200,

$\hat{a}_{ML(k)}$ = maximum likelihood estimate of a at the k-th sample drawing, k = 1 (1) 200,

$\hat{b}_{ML(k)}$ = maximum likelihood estimate of b at the k-th sample drawing, k = 1 (1) 200,

and based on the above estimates, mean squared errors (MSE's) in case of the method of moments estimates and of the maximum likelihood estimates of a and b are obtained respectively as

$$MSE(a)_{MOM} = \frac{1}{200}\sum_{k=1}^{200}(\hat{a}_{MOM(k)} - a_0)^2\ ,\ MSE(b)_{MOM} = \frac{1}{200}\sum_{k=1}^{200}(\hat{b}_{MOM(k)} - b_0)^2,$$

$$MSE(a)_{ML} = \frac{1}{200}\sum_{k=1}^{200}(\hat{a}_{ML(k)} - a_0)^2\ ,\ MSE(b)_{ML} = \frac{1}{200}\sum_{k=1}^{200}(\hat{b}_{ML(k)} - b_0)^2.$$

Again, these MSE's are calculated with the help of SAS/BASE and SAS/STAT procedures. The final conclusion is achieved on examining the MSE values obtained by adopting the two estimation procedures on the above samples generated through simulation. The error distributions are generated in case of both methods with the help of CAPABILITY and QQPLOT procedures in SAS/QC.

3. Results and discussion

Tables 2 to 7 contain the results which reveal distinctive features in respect of the objective (comparison of the two methods) of the investigation. These tables are the selected results of the 24 studies. In the study mentioned in Table 1, the initial values were taken around a = 1 and after observing the outcomes, values of a were chosen in the order a = 0.9, a = 0.7, a = 0.6, a = 0.5 and a = 0.4, etc. A striking feature (contrast) is noticed at this stage, which is mentioned below.

In the range $a \leq 0.5$, the method of moments estimator is found to be a better estimator than the competing maximum likelihood estimator with the increase in sample size (as found in Table 6, the contrast exists for a = 0.5, as in case of any value of $a < 0.5$). The same feature does not exist for $a > 0.5$. The above conclusion is evident on examining Table 2 ($a > 0.5$), Table 3 ($a < 0.5$), Table 4 ($a > 0.5$), Table 5 ($a < 0.5$), Table 6 (a = 0.5) and Table 7 ($a > 0.5$), that is, $a \leq 0.5$ for Tables 3, 5 and 6, while $a > 0.5$ for Tables 2, 4 and 7.

Table 2

$a = 2$		$b = 5$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.292506	0.386844	0.163807	0.386032	71	129
50	0.086835	0.083953	0.053938	0.082367	74	126
100	0.063344	0.051452	0.051601	0.037089	60	140
150	0.031392	0.036124	0.028461	0.029944	64	134
200	0.030698	0.020222	0.029706	0.014581	60	140
250	0.03047	0.020716	0.029624	0.010578	62	138
300	0.027172	0.010962	0.024307	0.011545	51	149

Table 3

$a = 0.4$		$b = 0.5$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.054437	0.043247	0.055646	0.046346	88	112
50	0.043496	0.039187	0.045742	0.041645	97	103
100	0.03358	0.0331278	0.034612	0.03459	104	96
150	0.02438	0.025821	0.025513	0.026921	126	74
200	0.014015	0.013446	0.015472	0.015041	144	56
250	0.009543	0.008347	0.011534	0.008399	160	40
300	0.002375	0.000153	0.003222	0.000275	171	29

Table 4

$a = 0.52$		$b = 2$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.143213	0.321942	0.113142	0.304219	93	107
50	0.104962	0.303216	0.094312	0.294219	92	108
100	0.094312	0.294216	0.082142	0.264219	94	106
150	0.073242	0.243216	0.064232	0.234162	90	110
200	0.064312	0.201312	0.059422	0.193219	84	116
250	0.043212	0.18419	0.043012	0.174236	80	120
300	0.032114	0.104862	0.032104	0.093612	75	125

Table 5

$a = 0.45$		$b = 2$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.316432	0.543162	0.308269	0.531682	65	135
50	0.246819	0.396219	0.231942	0.364198	80	120
100	0.144312	0.261483	0.141362	0.251362	89	111
150	0.093168	0.134231	0.094368	0.121368	100	100
200	0.054263	0.032162	0.064193	0.031492	115	85
250	0.014986	0.014269	0.019269	0.011268	132	68
300	0.009216	0.009421	0.009643	0.009341	147	53

Table 6

$a = 0.5$		$b = 2$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.24321	0.543216	0.043251	0.544216	84	116
50	0.132196	0.489326	0.035192	0.496231	92	108
100	0.094312	0.396214	0.024196	0.398242	96	104
150	0.023196	0.294361	0.233421	0.316431	101	99
200	0.019432	0.194321	0.219632	0.194319	107	93
250	0.014163	0.101642	0.141796	0.101649	110	90
300	0.009421	0.094321	0.095219	0.094412	117	83

Table 7

$a = 0.7$		$b = 2$			$n = 200$	
Sample size	$MSE(a)_{MOM}$	$MSE(b)_{MOM}$	$MSE(a)_{ML}$	$MSE(b)_{ML}$	Number of times the method of moments estimate of a is near to the original parameters	Number of times the maximum likelihood estimate of a is near to the original parameters
10	0.094212	0.364921	0.094209	0.364212	78	128
50	0.061212	0.271316	0.061021	0.263214	73	127
100	0.054121	0.143213	0.049131	0.123121	70	130
150	0.048316	0.084131	0.043212	0.071231	68	132
200	0.0321293	0.074212	0.029834	0.070012	67	133
250	0.021941	0.0694212	0.014131	0.068421	63	137
300	0.012131	0.059421	0.009832	0.052129	59	141

4. Error distribution

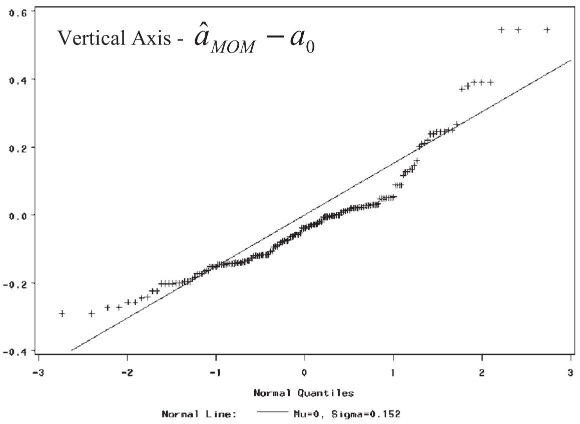
Samples generated for one set of parameter values (i.e., $a = 0.5$ and $b = 10$), using different sample sizes, are used for the examination of error distribution. Here our objective is to investigate the underlying pattern (i.e., whether the error distribution deviates from normality or not) of the error distribution in the cases of the four estimated parameters (for a and b using the methods of maximum likelihood and of moments) with the help of the Q-Q plot as well as with CAPABILITY procedure in SAS/QC. In what follows, we present the results in the cases of the four parameters for the sample sizes 10 and 300 only (though intermediate values 100 and 200 were also considered).

4.1 Sample size (n = 10)

The error distribution of $\hat{a}_{MOM}$

Variance = 0.0231 Skewness = 1.3314
Kurtosis = 2.5973
Test for location

Test	Statistic		p Value	
Student's t	t	–2.5680	Pr > |t|	0.0110
Sign	M	–31	Pr ≥ |M|	<0.0001
Signed rank	S	–3448	Pr ≥ |S|	<0.0001

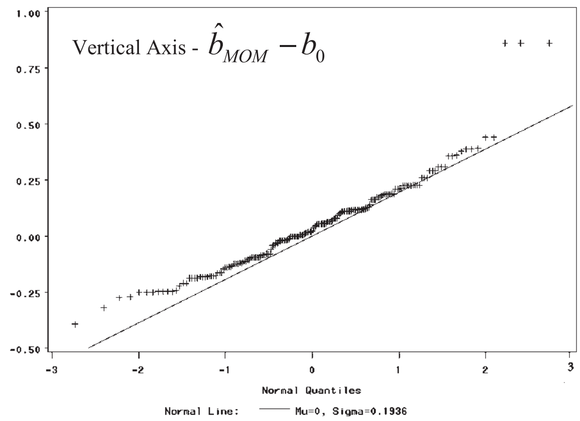


Graph 1: Q-Q Plot for error distribution of method of moments estimates of a

The error distribution of $\hat{b}_{MOM}$

Variance = 0.3749 Skewness = 1.0892
Kurtosis = 3.3117
Test for location

Test	Statistic		p Value	
Student's t	t	3.1078	Pr > |t|	0.0022
Sign	M	12	Pr ≥ |M|	0.1036
Signed rank	S	2057	Pr ≥ |S|	0.0117

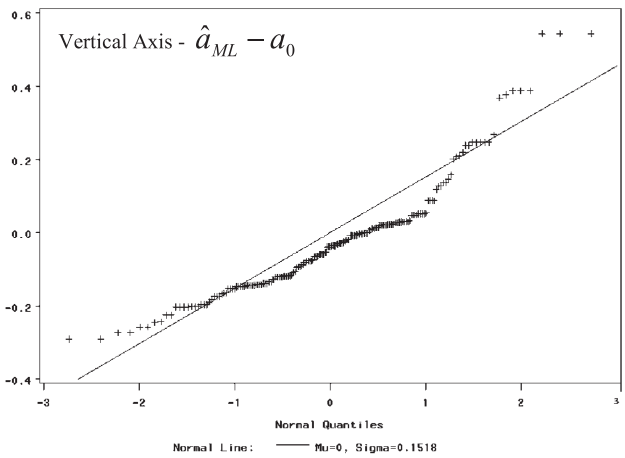


Graph 2: Q-Q Plot for error distribution of method of moments estimates of b

The error distribution of $\hat{a}_{ML}$

Variance = 0.0230 Skewness = 1.3257
Kurtosis = 2.5765
Test for location

Test	Statistic		p Value	
Student's t	t	–2.5684	Pr > |t|	0.0109
Sign	M	–27	Pr≥|M|	0.0002
Signed rank	S	–3429	Pr ≥ |S|	<0.0001

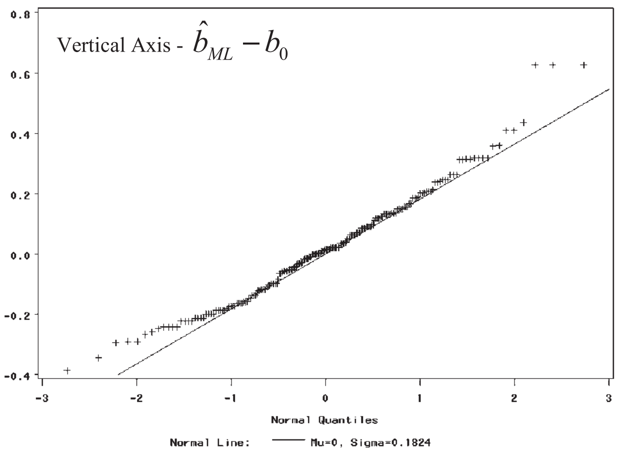


Graph 3: Q-Q Plot for error distribution of maximum likelihood estimates of a

The error distribution of $\hat{b}_{ML}$

Variance = 0.0333 Skewness = 0.5650
Kurtosis = 0.6277
Test for location

Test	Statistic		p Value	
Student's t	t	1.5976	Pr > |t|	0.1117
Sign	M	8	Pr≥|M|	0.2888
Signed rank	S	873	Pr ≥ |S|	0.2879



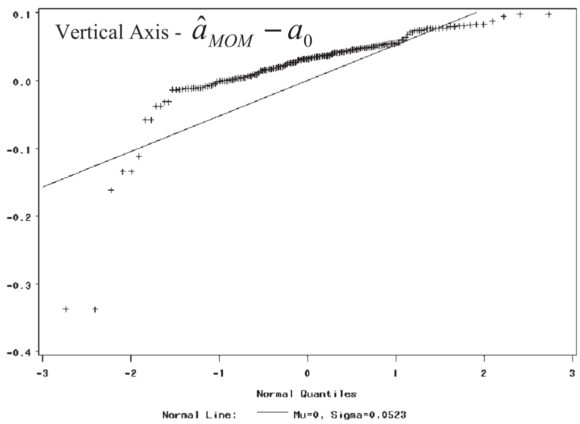
Graph 4: Q-Q Plot for error distribution of maximum likelihood estimates of b

4.2 Sample size (n = 300)

The error distribution of $\hat{a}_{MOM}$

Variance = 0.0027 Skewness = –3.8523
Kurtosis = 22.8482
Test for location

Test	Statistic		p Value	
Student's t	t	6.5010	Pr > $\|t\|$	<.0001
Sign	M	63	Pr≥ $\|M\|$	<.0001
Signed rank	S	7471	Pr ≥ $\|S\|$	<.0001

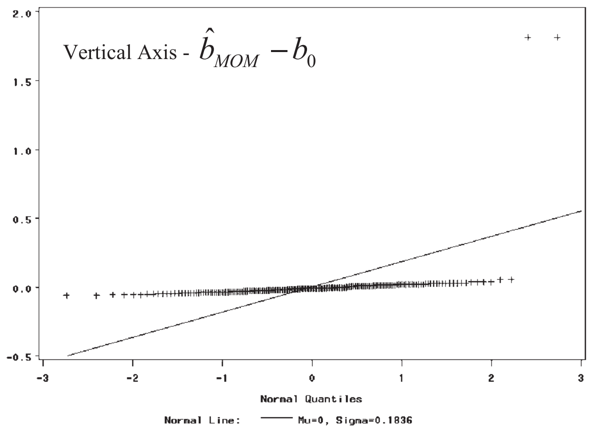


Graph 5: Q-Q Plot for error distribution of method of moments estimates of a

The error distribution of $\hat{b}_{MOM}$

Variance = 0.0337 Skewness = 9.6596
Kurtosis = 93.9748
Test for location

Test	Statistic		p Value	
Student's t	t	0.6873	Pr > $\|t\|$	0.4927
Sign	M	–30	Pr≥ $\|M\|$	<0.0001
Signed rank	S	–3695	Pr ≥ $\|S\|$	<0.0001

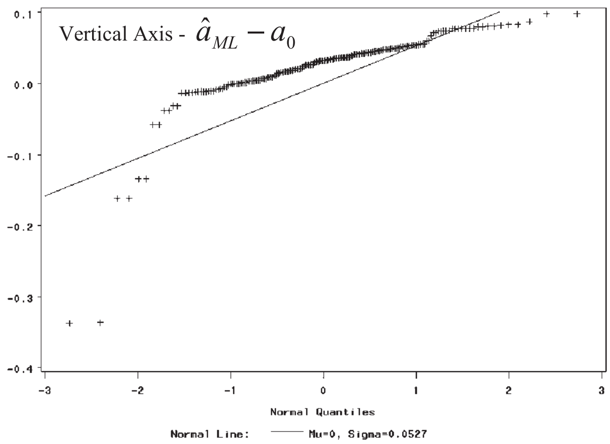


Graph 6: Q-Q Plot for error distribution of method of moments estimates of b

The error distribution of $\hat{a}_{ML}$

Variance = 0.0028 Skewness = –3.8476
Kurtosis = 22.2221
Test for location

Test	Statistic		p Value	
Student's t	t	6.1618	Pr > $\|t\|$	<.0001
Sign	M	63	Pr≥ $\|M\|$	<.0001
Signed rank	S	7431	Pr ≥ $\|S\|$	<.0001

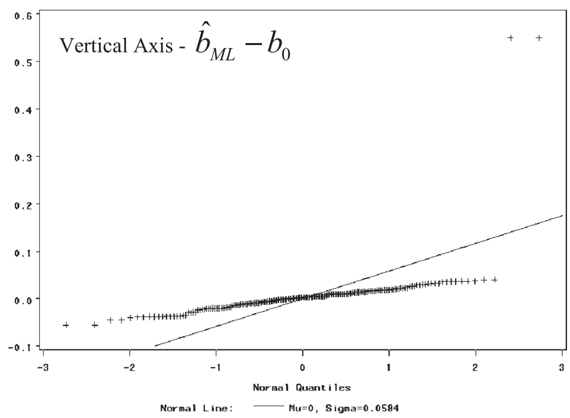


Graph 7: Q-Q Plot for error distribution of maximum likelihood estimates of a

The error distribution of $\hat{b}_{ML}$

Variance = 0.0034 Skewness = 8.1770
Kurtosis = 75.2613
Test for location

Test	Statistic		p Value	
Student's t	t	1.4176	Pr > $\|t\|$	0.1579
Sign	M	12	Pr≥ $\|M\|$	0.1036
Signed rank	S	846	Pr ≥ $\|S\|$	0.3031



Graph 8: Q-Q Plot for error distribution of maximum likelihood estimates of b

It can be seen from the above Q-Q plots that when n = 10 (the first four graphs), the points are lying more or less parallel (sometimes coincident) to the straight line, which indicate that the error distributions are near normal. But when $n = 300$, last four graphs, the points do no longer lie parallel to the straight line revealing that the error distributions deviate away from normality, which is also an interesting finding in this paper. The existence of the above pattern is also observed for values of $n = 100$ and 200. Thus, as the sample size increases, the error distributions deviate much away from normality.

5. Findings from the investigation

1) For $a < 0.5$, when the sample size increases the MSE in case of method of moments estimator is less than the MSE in case of the maximum likelihood estimator.
2) The value of the parameter b does not play any significant role.
3) When $a < 0.5$, the number of times the method of moments estimator of the parameter a is near to the original parameter is much more than that in case of the maximum likelihood estimator as the sample size increases.
4) The investigation has been carried out assigning very small values to 'a and b', at first. Some parameter values were taken in the range less than 1 (also greater than 1) for a, and after examining the outcomes, the latter parameter set values were taken and it was found that a sharp feature (contrast) is visible in the range around 0.5. In the range, $a \leq 0.5$, the method of moments estimators are found to be better than the maximum likelihood estimators as the sample size increases, as was mentioned in Section 3 also. This is evident while examining the tables for values of $a \leq 0.5$ as well as $a > 0.5$.
5) When the sample size is small, the error distributions (obtained by using both methods and for both a and b) follow near normality.
6) As the sample size increases, the error distributions deviate much away from normality.
7) The observations (graph plots) on the error distribution which have been included for $a = 0.5$ and $b = 10$, can be obtained for other parameter set values also. Similar patterns will follow in case of increase in sample sizes for other parameter set values.

6. Conclusion in case of EV I(2) distribution

1) Method of moments yields more accurate estimators of the parameters than the maximum likelihood estimators when the values of the parameter a is less than or equal to 0.5.
2) When the sample size is small, the error distribution in parameter estimation follows normality, but it gradually deviates away from normality with the increase in sample sizes.

References

Gumbel, E. J. (1958). Statistics of Extremes, Columbia Univ. Press, New York, USA.

Kite, G. W. (1977). Frequency and Risk Analysis in Hydrology, Water Resource Publications, Fort Collins, Colorado, USA.