Comparison of Some Confidence Intervals in Two Parameter Pareto Distribution

Masood Ul Haq* and Muhammad Ajaz Rasheed**

ABSTRACT

In the two parameter Pareto Distribution; $f(x; a, v,) = v\, a^{v} x^{-v-1}$, $x > a, v > 0, a > 0$ the Confidence Intervals of log a and v have been developed. For the Confidence Interval of log a, two methods have been used;

(1) by using 2 n v log (g/a), which follows a chi- square distribution, and

(2) by using the distribution of $u = a/x_{(1)}$. The theory has been compared by computer simulation results (Table (1)).

For the Confidence Interval of the other parameter v, again two methods have been used;

(A) by using log-range, $R = \log X_{(n)} - \log X_{(1)}$ and

(B) by using the normality of MLE of v.

Later the theory has been compared by computer simulation results (Table (2)).

INTRODUCTION:

The two parameter Pareto Distribution is defined by:

$$f(x;a,v,) = v\, a^{v} x^{-v-1} \quad x > a\,,\ v > 0.$$

Several sampling distributions from the Pareto Distribution have been discussed by Malik H.J. [1], namely the distribution of the Geometric mean; the product of two minimum values from sample of unequal sizes; the product of 'k' minimum values from sample of equal sizes etc.. An extensive discussion has also been given by Muniruzzaman A.N.M. [2]; He has discussed the maximum likelihood estimates of Geometric Mean, H.M. and median of the sampling distribution of these estimates.

Pareto Distribution has found wide spread use in the statistical description of the upper half of the size distribution of such diverse things as employment incomes, mineral deposits, property values, city population sizes, size of firms and measurable human abilities.

In the present paper a study has been made regarding the confidence Interval of the parameters 'log a' and 'v'. Two methods have been employed to study the confidence interval of 'log a':

(i) By using the Sampling Distribution of the extreme order statistics.

(ii) By using the Sampling Distribution of the Geometric Mean.

The two confidence interval have been compared by examining the "rate of occurrence of the true value of the parameter 'log a', " inside the confidence interval. The comparison has been accomplished by a statistical simulation i.e. 150 samples are drawn from specified Pareto Distribution when each sample consisted of 10, 20, 30, 40 obbservations.

Jennings [5] has pointed out that simulation studies of confidence interval procedures often only report convergence rates. This is not sufficient to judge whether

KEY TERMS: Pareto Distribution, Log Range Distribution, Confidence Interval, Computer Simulation.

* Assoc. Prof., Dept. of Statistics, Univ. of Karachi (PAKISTAN).

** Lecturer, Dept. of Statistics, Univ. of Karachi (PAKISTAN)

the intervals are "unbiased" i.e. whether they are equally likely to be above or below the true values if they don't cover the true values; the user expects this because systematic percentile of the given sampling distribution are used in forming the confidence interval but this may not occur due to the skewness of the actual sampling distribution. As mentioned above, the sampling distribution of the extreme order statistics $x_{(1)}$ has been employed to obtain the confidence interval of 'log a', whis is skewed.

The other sampling distribution employed to obtain the confidence interval of 'log a' is chi-square which is right skewed for small valued of sample size and get symmetrical for large 'n'.

For the confidence interval of the parameter 'v' in the Pareto Distribution, the pivotal statistic is $w = v\{\log x_{(n)} - \log x_{(1)}\}$. The sampling distribution of w is independent of the parameters 'a' and 'v' (See section 3); the distribution is shown in Fig (1) for various values of sample size 'n'; It shows that the sampling distribution of w is skewed for small n, but gets symmetrical for large 'n'.

2. Confidence Interval of 'log a'.

Suppose $x_{(1)}, x_{(2)}, \ldots\ldots, x_{(n)}$ is a random sample from the two parameter Pareto Distribution:

$$f(x; a, v,) = v\, a^{v}\, x^{-v-1}, \; x > a, \; v > 0. \qquad (1)$$

The results that have been employed for the formulation of the Confidence Intervals of 'log a' are given below as theorem (1) and theorem (2).

Theorem (1): " If x is a Pareto variable, then

$$x_{(1)} = \text{Min}\{x_{(1)}, x_{(2)}, \ldots\ldots, x_{(n)}\},$$

is also Pareto distributed with parameters 'nv' and 'a'. "

The proof of the theorem is obtained by deriving the probability density function of $x_{(1)}$ which is:

$$f(x_{(1)}; a, v) = n\, v\, a^{nv}\, x_{(1)}^{-nv-1}, \; a < x_{(1)} < \infty \qquad (2)$$

and comparing it with the p.d.f. of the Pareto Distribution with parameters v and a, David [6] the distribution of $x_{(1)}$ is transformed by

$$u = a / x_{(1)}$$

to give the probability distribution :

$$f(u) = n\, v\, u^{nv-1}; \; 0 \le u \le 1 \qquad (3)$$

The variable 'u' acts as a pivotal statistic for determining the confidence interval of 'log a'. The confidence interval of 'log a' thus obtained is: (when $a > 1$; See Table (2))

$$\{\log x_{(1)} + \log u_1 ; \log x_{(1)} + \log u_2\} \qquad (4)$$

for the Confidence co-efficient $(1-\alpha)$. u_1 and u_2 are lower and upper limits of the u-distribution so that

$$P[u_1 < u < u_2] = 1 - \alpha.$$

Theorem (2) : " Let the sample geometric mean be g and $g = \exp(u/n)$ where $u = \Sigma \log x_i$, and x_i is pareto distributed for $i = 1, 2, \ldots\ldots, n$, as given in (1). Then $2nv \log(g/a)$ is distributed as chi-square with 2n degrees of freedom."

The proof of this theorem is given by Malik [1]. The variable $2nv \log(g/a)$ acts as a pivotal statistic for determining the confidence interval of 'log a'. The confidence interval of 'log a' for a given v, thus obtained is :

$$\left\{\log_e g - \frac{\chi^2_2}{2nv} ; \log_e g - \frac{\chi^2_1}{2nv}\right\} \qquad (5)$$

for the Confidence co-efficient $(1-\alpha)$. χ^2_1 and χ^2_2 are lower and upper limits of the chi-square (χ^2) with 2n degrees of freedom and at $(1-\alpha)$ level, so that

$$P[\chi^2_1 < \chi^2 < \chi^2_2] = 1 - \alpha$$

3. Confidence Interval of 'v'

In order to determine the confidence interval of the parameter 'v', the statistic $R = (\log x_{(n)} - \log x_{(1)})$ is employed,

where

$$x_{(n)} = \text{Max}\,(x_{(1)}, x_{(2)}, \ldots\ldots, x_{(n)})$$

and

$$x_{(1)} = \text{Min}\,(x_{(1)}, x_{(2)}, \ldots\ldots, x_{(n)})$$

the probability density of

$$w = vR = v\,(\log x_{(n)} - \log x_{(1)}) \qquad (6)$$

is

$$f(w) = (n-1)\,e^{-w}\,(1 - e^{-w})^{n-2}, \; 0 \le w \le \infty \qquad (7)$$

Thus 'w' would be used as a pivotal statistic. w_1 and w_2 for various values of n and $1-\alpha$ are obtained by solving the integral:

$$\int_0^{w_1} (n-1)\,e^{-w}\,(1 - e^{-w})^{n-2}\,dw = \alpha/2 \qquad (8)$$

$$\int_0^{w_2} (n-1)\,e^{-w}\,(1 - e^{-w})^{n-2}\,dw = 1 - \alpha/2 \qquad (9)$$

Thus $w_1 = \log_e \left\{ \dfrac{1}{1 - (\alpha/2)^{1/(n-1)}} \right\}$

The values of w_1 and w_2 for $\alpha = 0.05$ are given below:

Table (1)

w_1 and w_2 for $\alpha = 0.05$

n	w_1	w_2
10	1.08985	5.87488
20	1.73462	6.62135
30	2.12490	7.04398
40	2.40516	7.34013
50	2.62390	7.56833

Fig (1) shows that the distribution of w for small samples is positively skewed with a small left tail. As the size of the sample increases, the right and left tail becomes more pronounced and the probability curve becomes more and more symmetrical.

For the Confidence co-efficient $(1 - \alpha)$, w_1 and w_2 can be formed as the lower and upper limit of w to give:

$$P\,[\,w_1 < w < w_2\,] = 1 - \alpha$$

Thus

$$P\left[\frac{w_1}{\log X_{(n)} - \log X_{(1)}} < v < \frac{w_2}{\log X_{(n)} - \log X_{(1)}}\right] = 1 - \alpha \qquad (10)$$

Another method to determine the confidence interval of 'v' is to use asymptotic property of the maximum likelihood estimator of v which follows a normal distribution with mean and variance respectivelt as v and v^2/n, since

$$E\left(\frac{\partial}{\partial v}\log(x; v)\right)^2 = \frac{1}{v^2} \text{ and } \frac{\partial}{\partial v}\log(x; v) = -\frac{1}{v}\log\left(\frac{x}{a}\right),$$

hence the Cramer-Rao lower bound is attained and $v(v) = v^2/n$. Thus the standard normal variate

$$z = \frac{(V - v)}{v/\sqrt{n}}$$

is employed as the pivotal statistic, where

$$V = [\,\log (g/x_{(1)})\,]^{-1} \qquad (11)$$

the confidence interval of 'v' with $1-\alpha$ as the Confidence co-efficient is obtained as

$$P\left[V - Z_{\alpha/2}\frac{v}{\sqrt{n}} < v < V + Z_{\alpha/2}\frac{v}{\sqrt{n}}\right] = 1 - \alpha \qquad (12)$$

4. Statistical Simulation of Confidence Intervals

Case I : Confidence Interval of Log a

The confidence interval of 'log a ' has been obtained by drawing 150 random samples from the Pareto Distribution each sample consisting of 10, 20, 30 and 40 observations. The Confidence co-efficient is 0.95. The number of times these confidence intervals fell below and above the true value 'log a' are recorded in Table No. 2. Ideally, we would expect the interval to fall below the true value about 2.5% times and above equally often. Method 1 in Table (2) refers to expression (5) and Method 2 to expression 4.

Table No. (2)

150 random samples for n =10, 20, 30 and 40 parameter 'a' in col (1);

f^1 = % of samples below log g - (χ^2_2/ 2 n v)

f^2 = % of samples above log g - (χ^2_1/ 2 n v)

f^3 = % of samples within the confidence interval

95% Confidence interval for log a, v=1.5 (fixed)

Chi Squared as a Pivot

Method 1

n = 10

a	f^1	f^2	f^3
1.1	2.67	1.33	96.00
1.2	2.00	2.00	96.00
1.3	2.00	2.67	95.33
1.4	2.00	2.00	96.00
1.5	3.33	2.67	94.00
1.6	3.33	2.67	94.00
1.7	3.33	3.33	96.00
1.8	2.67	1.33	96.00
1.9	2.00	2.00	96.00
2.0	2.00	2.67	95.33
2.1	3.33	3.33	93.33
2.2	3.33	2.67	94.00
2.3	3.33	2.67	94.00
2.4	3.33	3.33	93.33
2.5	2.67	1.33	96.00

n = 20

a	f^1	f^2	f^3
1.1	1.33	2.00	96.67
1.2	6.67	5.33	88.00
1.3	4.67	2.67	92.67
1.4	3.33	2.00	94.67
1.5	4.67	3.33	92.00
1.6	3.33	2.00	94.67
1.7	4.67	3.33	92.00
1.8	3.33	2.00	94.67
1.9	4.67	3.33	92.00
2.0	3.33	2.00	94.67
2.1	4.67	3.33	92.00
2.2	1.33	2.67	96.00
2.3	3.33	0.67	96.00
2.4	2.67	2.67	94.67
2.5	2.67	3.33	94.00

n = 30

a	f^1	f^2	f^3
1.1	1.33	4.00	94.67
1.2	1.33	1.33	97.33
1.3	3.33	4.00	92.67
1.4	4.67	2.00	93.33
1.5	3.33	1.33	95.33
1.6	2.00	3.33	94.67
1.7	4.00	4.00	92.00
1.8	2.00	4.00	94.00
1.9	4.00	4.00	92.00
2.0	2.00	3.33	94.67
2.1	2.67	0.67	96.67
2.2	2.00	3.33	94.67
2.3	2.67	3.33	94.00
2.4	2.67	4.00	93.33
2.5	3.33	2.00	94.67

n = 40

a	f^1	f^2	f^3
1.1	3.33	1.33	95.33
1.2	0.67	4.67	94.67
1.3	2.67	2.00	95.33
1.4	1.33	3.33	95.33
1.5	3.33	4.00	92.67
1.6	1.33	2.67	96.00
1.7	2.00	5.33	92.67
1.8	3.33	0.67	96.00
1.9	4.00	3.33	92.67
2.0	4.00	2.00	94.00
2.1	2.67	2.67	94.67
2.2	3.33	3.33	93.33
2.3	4.67	2.00	93.33
2.4	1.33	2.67	96.00
2.5	5.33	2.00	92.67

n = 20

a	f^1	f^2	f^3
1.1	2.00	2.67	95.33
1.2	0.67	3.33	96.00
1.3	1.33	2.00	96.67
1.4	2.00	2.00	96.00
1.5	0.67	2.67	96.67
1.6	4.00	4.00	92.00
1.7	0.00	3.33	96.67
1.8	3.33	3.33	93.33
1.9	2.67	2.00	95.33
2.0	2.67	0.67	96.67
2.1	2.67	3.33	94.00
2.2	2.67	0.67	96.67
2.3	2.67	3.33	94.00
2.4	0.67	2.67	96.67
2.5	3.33	2.67	94.00

95% Confidence interval for log a , v = 1.5

Method 2

n = 10

a	f^1	f^2	f^3
1.1	1.33	1.33	97.33
1.2	1.33	2.67	96.00
1.3	3.33	0.67	96.00
1.4	2.67	2.00	95.33
1.5	0.00	2.00	98.00
1.6	4.00	2.00	94.00
1.7	4.00	1.33	94.67
1.8	2.67	3.33	94.00
1.9	2.67	4.67	92.67
2.0	4.67	0.67	94.67
2.1	3.33	2.67	94.00
2.2	8.67	2.67	88.67
2.3	2.00	2.00	96.00
2.4	0.67	0.67	98.67
2.5	1.33	1.33	97.33

n = 30

a	f^1	f^2	f^3
1.1	4.00	2.67	93.33
1.2	3.33	0.67	96.00
1.5	0.67	2.00	97.33
1.6	4.00	4.00	92.00
1.7	3.33	2.00	94.67
1.8	2.67	0.67	96.67
1.9	6.00	4.67	89.33
2.0	3.33	2.00	94.67
2.1	2.00	2.67	95.33
2.2	2.67	1.33	96.00
2.3	2.00	3.33	94.67
2.4	2.00	1.33	96.67
2.5	2.67	1.33	96.00

n = 40

a	f^1	f^2	f^3
1.1	0.67	1.33	98.00
1.2	0.67	2.00	97.33
1.3	2.67	4.00	93.33
1.4	0.67	4.00	95.33
1.5	3.33	3.33	93.33
1.6	3.33	2.67	94.00
1.7	4.00	3.33	92.67
1.8	0.00	2.67	97.33
1.9	1.33	3.33	95.33
2.0	2.67	2.00	95.33
2.1	1.33	2.67	96.00
2.2	2.67	3.33	94.00
2.3	1.33	3.33	95.33
2.4	1.33	1.33	97.33
2.5	3.33	0.67	96.00

Conclusions :

The table presented above is quite brief, since only one level of significance (i.e. 5%) has been shown; The total No. of samples drawn from the given Pareto Distribution in each case is 150. The Tables show that the performance of the Confidence Intervals are quite good, except for starred samples (*).

Case II : Confidence Interval of 'v'

The Confidence Interval of 'v' has been obtained by drawing 150 random samples from pareto distribution, each sample consisting of 10,20,30 and 40 observations. The Confidence Co-efficient (1- α) is 0.95. The number of times these Confidence Interval fell below and above 'v' for a given value of 'a' are recorded in the table No. 3. The table shows the performance of the two methods.

Method A: It is shown in expression (10). The pivotal statistic in this method is

$$w = v \{ \log x_{(n)} - \log x_{(1)} \} \text{ and}$$

$$P [w_1 \leq w \leq w_2] = 1 - \alpha$$

Method B : In this method the pivotal statistic is

$$Z = (v - v) / (v/\sqrt{n})$$

where

$$V = [\log (g/ X_{(1)})]^{-1}$$

The author has examined the frequency table of:

$$Z = (v - v) / (v/\sqrt{n})$$

for various values of v and n, through computer simulation. It is observed that the normality of v is good when n is approximately 100 and a = 1.5 giving β_1 = 0.257 and β_2 = 3.09. However the frequency table is unimodal and two tailed even for smaller sample sizes.

Table No. (3)

150 random samples for n = 10, 20, 30 and 40 parameter 'v' in col (1); parameter a = 1.5.

f_1 = % of samples below $W_1 / \{\log X_{(n)} - \log X_{(1)} \}$

f_2 = % of samples above $W_2 / \{\log X_{(n)} - \log X_{(1)} \}$

f_3 = % of samples within the confidence interval

95% Confidence interval for 'v', a = 1.5

w1 = 1.0 w2 = 5.874878

Method A

n = 10

v	f^1	f^2	f^3
1.10	2.00	2.00	96.00
1.20	1.33	1.33	97.33
1.30	1.33	4.00	94.67
1.40	2.00	2.67	95.33
1.50	2.00	2.00	97.33
1.60	0.67	3.33	90.00
1.70	6.67	0.67	94.67
1.80	4.67	3.33	92.00
1.90	4.67	0.67	96.67

2.00	2.67	0.67	96.67
2.10	2.67	2.67	94.67
2.20	1.33	4.00	94.67
2.30	0.00	3.33	96.67
2.40	2.67	1.33	96.00
2.50	3.33	1.33	95.33

2.00	3.33	4.00	92.67
2.10	2.00	1.33	96.67
2.20	2.00	1.33	96.67
2.30	2.67	2.67	94.67
2.40	3.33	0.00	96.67
2.50	3.33	1.33	95.33

n = 20

v	f^1	f^2	f^3
1.10	2.00	3.33	94.67
1.20	2.67	0.67	96.67
1.30	4.00	3.33	92.67
1.40	2.00	4.00	94.00
1.50	2.00	2.67	95.33
1.60	2.00	2.00	96.00
1.70	4.67	0.67	94.67
1.80	1.33	2.00	96.67
1.90	4.00	2.67	93.33
2.00	2.67	1.33	96.00
2.10	1.33	1.33	97.33
2.20	2.00	0.67	97.33
2.30	0.67	2.67	96.67
2.40	0.67	2.67	96.67
2.50	2.00	1.33	96.67

n = 40

v	f^1	f^2	f^3
1.10	4.67	2.00	93.33
1.20	2.00	2.67	95.33
1.30	4.00	2.00	94.00
1.40	4.67	0.67	94.67
1.50	2.67	2.67	94.67
1.60	3.33	2.67	94.00
1.70	2.00	3.33	94.67
1.80	2.00	3.33	94.67
1.90	4.00	1.33	94.67
2.00	2.67	1.33	96.00
2.10	0.67	1.33	98.00
2.20	0.67	3.33	96.00
2.30	0.67	2.67	96.67
2.40	1.33	2.00	96.67
2.50	4.67	4.67	90.67

n = 30

v	f^1	f^2	f^3
1.10	2.67	2.00	95.33
1.20	2.67	4.00	93.33
1.30	0.67	2.67	96.67
1.40	2.00	2.00	96.00
1.50	2.00	5.33	92.67
1.60	0.67	2.00	97.33
1.70	1.33	2.00	96.67
1.80	1.33	1.33	97.33
1.90	1.33	1.33	97.33

95% Confidence interval for 'v' , a = 1.5

Method B

n = 10

v	f^1	f^2	f^3
1.10	1.33	1.33	97.33
1.20	2.00	2.00	96.00
1.30	2.00	1.33	96.67
1.40	0.67	3.33	96.00
1.50	1.33	1.33	97.33
1.60	0.00	0.67	99.33
1.70	1.33	1.33	97.33

1.80	2.67	4.00	93.33
1.90	3.33	3.33	93.33
2.00	3.33	2.00	94.67
2.10	0.00	3.33	96.67
2.20	2.67	1.33	96.00
2.30	2.00	3.33	94.67
2.40	1.33	2.00	96.67
2.50	2.67	2.67	94.67

n = 20

v	f^1	f^2	f^3
1.10	2.00	2.67	95.33
1.20	2.67	2.67	94.67
1.30	1.33	2.67	96.00
1.40	3.33	0.00	96.67
1.50	0.67	2.00	97.33
1.60	2.67	2.67	94.67
1.70	1.33	2.67	96.00
1.80	1.33	3.33	95.33
1.90	2.00	2.00	96.00
2.00	0.67	3.33	96.00
2.10	2.67	2.67	94.67
2.20	1.33	3.33	95.33
2.30	2.00	2.00	96.00
2.40	2.67	2.67	94.67
2.50	3.33	1.33	95.33

n = 30

v	f^1	f^2	f^3
1.10	3.33	4.00	92.67
1.20	1.33	0.67	98.00
1.30	4.00	1.33	94.67
1.40	2.00	2.00	96.00
1.50	2.00	2.67	95.33
1.60	2.67	3.33	94.00
1.70	0.67	2.67	96.67
1.80	3.33	1.33	95.33
1.90	2.00	2.67	95.33
2.00	2.67	2.00	95.33
2.10	3.33	1.33	95.33
2.20	3.33	1.33	95.33
2.30	4.00	1.33	94.67
2.40	4.00	1.33	94.67
2.50	4.00	1.33	94.67

n = 40

V	f^1	f^2	f^3
1.10	1.33	1.33	97.33
1.20	0.67	3.33	96.00
1.30	2.67	1.33	96.00
1.40	0.67	1.33	98.00
1.50	1.33	2.67	96.00
1.60	2.00	1.33	96.67
1.70	1.33	4.67	94.00
1.80	2.00	0.67	97.33
1.90	1.33	2.00	96.67
2.00	2.67	2.00	95.33
2.10	1.33	1.33	97.33
2.20	0.67	3.33	96.00
2.30	2.00	1.33	96.67
2.40	1.33	1.33	97.33
2.50	2.67	2.00	95.33

5. Comparison of Confidence Intervals of LOG_a

The confidence intervals of $\log_a$ as determined by expressions (4) and (5) show equally good performance when these are compared on the basis of simulation in Table No. (2). However, it must be observed that in expression (4), the confidence interval of loga is obtained by the use of the 1st ordered variable, $x_{(1)}$, and the (n - 1) remaining observations are not used, and thus this procedure is simple. The confidence interval in expres-

sion (5) is obtained by the use of the sample geometric mean "g", which requires all the observations of the sample; thus this method is comparatively lengthy. Both methods are based on the exact probability "α" (the level of confidence).

Comparison of Confidence Intervals of v:

The confidence intervals of "v" as determined by expression (10) is an exact confidence interval, since it is based on the exact distribution of

$$w = v (\log x_{(n)} - \log x_{(1)}),$$

whereas the confidence interval as determined by expression (12) is an asymptotic result, since it is based on the asymptotic property of the maximim likelihood estimator of v. In this sense the confidence interval from (10) is preferable to that obtained from (12). Besides method A is based on extreme ordered statistics, so that it does not use the (n - 2) remaining observations, whereas method B requires the sample geometric mean and requires the whole sample.

References :

[1] Malik, H.J. Distribution of product statistics from Pareto Distribution, Matrika, Vol 15, 1976. Fasc 1/2.

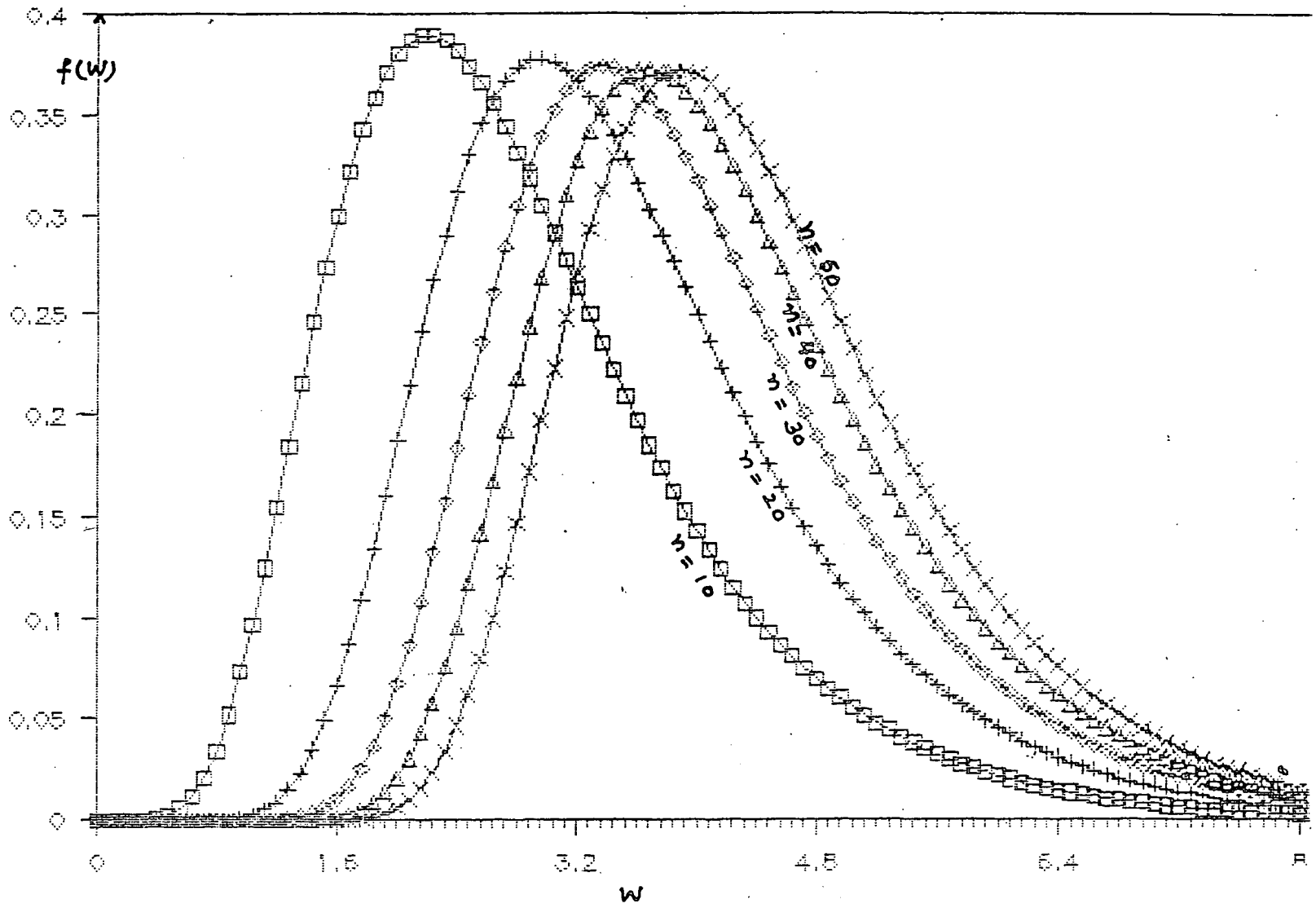
[2] Munir-uzzaman, A.M.N. On Measure of Location an Dispersion and tests of Hypothesis in a Pareto Population, Culcutta Statistical Association, Vol 17, No. 7, 1957.

[3] Malik, H.J. A Characterization of the Pareto Distribution. Skand. Aktuar Tidskr 1970.

[4] Malik, H.J. Exact moments of order statistics from Pareto Distribution, Skand. Aktuar Tidskr 1966.

[5] Jennings, D.E. How do we judge Confidence Intervals. The American Statistician, Vol 41, No. 4. 1987.

[6] David, H. A., Order Statistics, John Wiley & Sons N. Y. (1981).



fig(1)