Aligarh Journal of Statistics
Vol. 35(2015), 107-116

ESTIMATING THE PARAMETERS OF A PARETO DISTRIBUTION THROUGH LOCAL FREQUENCY RATIO METHOD

Ch.Yugandhar and V.V.Haragopal

ABSTRACT

Most of the Estimation techniques are based on complete information of the sample and parameters are estimated by different Methods of Estimation. But if only partial information is used, how to estimate the parameter? If estimated, how far these estimators are good when comparing with the full information sample. In this study we apply Local Frequency Ratio Method to estimate the parameters and found that this method estimates the parameters effectively with less information.

1. INTRODUCTION

According to Hogg and Tanis (2001) estimation is a process of inputting numerical values (or a range for the values) to the parameter of interest based on a sample observation coming from a specified distribution. The function of the sample value (used for this purpose) is a statistic and its value (in case of point estimation) is taken as the parameter values of the distribution or a specified function of the parameter. The statistics so used, is called an estimator of the parameter and the particular value obtained from the data is called an estimate.

Most of the estimation techniques for estimation of parameters are based on complete Information of sample under study. However it is also possible and often necessary to construct estimators based on partial information from samples i.e. by information on sample values, which falls into two or few of their lines or bins in a frequency distribution ignoring the values falling into other region in the frequency distribution. Estimators are based not on global but on local information from the sample.

This approach is of course not entirely new. For instance, estimation problem in situation where sample observations are censored or truncated can obviously be claimed to belong to this category. But any detailed study of such estimation procedure and the properties of such estimators do not seem to have been reported so far. The present problem is an effort in this direction.

Our investigation aims to answer mainly the following questions.

1) Using only local information from different localities (locals) in the sample set, how good an estimator of the parameter can one hope to obtain?

2) How do these estimators compare with the usual, full global sample based estimators?

3) Particularly, when only partial data is considered how the local information estimator compares with the global estimators that takes into account the entire sample.

2. THE PARETO DISTRIBUTION

The Pareto distribution is a Continuous probability distribution named after the economist Vilfredo Pareto. The Pareto distribution has been used to represent the income distribution of a society. It is also used to model many phenomena such as city population sizes, occurrence of natural resources, stock price fluctuations, size of firms, and brightness of comets. The Pareto distribution is defined by the following functions:

$$PDF : f\left\langle x \mid \alpha, k \right\rangle = \frac{a\,k^{a}}{x^{a+1}}; k \le x \le \infty; a, k > 0$$

$$CDF : F\left\langle x \mid a.k \right\rangle = 1 - \left(\frac{k}{x}\right)^{a}; k \le x \le \infty; a, k > 0$$

The parameter k marks a lower bound on the possible values that a Pareto distributed random variable can take on. A few well known properties are:

$$E(X) = \frac{a\,k}{a-1}, a > 1; V(X) = \frac{a\,k^{2}}{\left[(a-1)^{2}(a-2)\right]}, a > 2$$

2.1 Parameter Estimation

We are interested in estimating the parameters of the Pareto distribution from which a random sample comes. We will outline a few parameter estimation methods.

2.1.1 Method of moments

Under this method, we equate the sample mean and variance with the distribution's theoretical expected value and variance. We obtain two equations and two unknowns:

$$\overline{x} = \frac{a\,k}{a-1}; s^{2} = \frac{a\,k^{2}}{\left[(a-1)^{2}(a-2)\right]}$$

Solving these equations yields the following estimators of *a* and *k:*

$$\hat{a} = 1 + \sqrt{\frac{1 + \overline{x}^{-2}}{s^2}}; \; \hat{k} = \frac{\overline{x}(a-1)}{a}$$

2.1.2 Maximum Likelihood Estimation

Let $x_1, x_2, \ldots x_n$ a random sample of n observations drawn from Pareto population. The likelihood function for the Pareto distribution has the form

$$L = \prod_{i=1}^{n} f\langle x_i | a, k \rangle = \prod_{i=1}^{n} \frac{a k^a}{x_i^{a+1}}; \qquad L = \frac{a^n k^{an}}{\prod_{i=1}^{n} x_i^{a+1}}$$

The maximum likelihood estimates for *k* and *a* are the values of *k* and *a* that makes L as large as possible given the data we have. The value of k cannot be larger than the smallest value in the sample generated, so the estimate of *k* can be taken as $\hat{k} = \min(x_i)$.

In order to find the maximum likelihood estimate for *a*, we take logarithm of L, since L is nonnegative.

$$LogL = n \log a - an \log k - (a+1)\sum_{i=1}^{n} \log x_i$$

The likelihood equation is $\frac{\partial \log L}{\partial a} = 0$, $\frac{n}{a} - n \log k - \sum_{i=1}^{n} \log x_i = 0$

On simplification, we get $\hat{a} = \frac{n}{n \log k - \sum_{i=1}^{n} \log x_i} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{k}\right)}$

For example; we generate 50 random samples, each of size 1000 from Pareto distribution by taking (K=1; a=3). For each sample we estimate parameters k and a by using above procedures. The Mean, Standard Error , $\sqrt{\beta_1}, \beta_2$ of these 50 estimates were computed. The Estimated bias was calculated as the mean minus the true value of the parameter. The Mean Squared Error (MSE) was calculated as the bias squared plus the variance. The results are shown in the following table.

Table 1: Descriptive for both methods

	Method of Moments		Method of MLE	
	a	k	a	k
Mean	3.2064	1.0278	2.9937	1.0004

SE	0.3585	0.0482	0.1010	0.0004
$\sqrt{\beta_1}$	0.4106	1.1852	0.4220	1.4256
β_2	2.9936	4.3673	3.2226	4.4269
Bias	0.2064	0.0278	-0.0063	0.0004
MSE	0.1711	0.0031	0.0102	0.0000

2.1.3 Frequency Ratio Method of Estimation

Let $y_1, y_2 \ldots\ldots y_n$ be a random sample from a distribution. From this sample a frequency distribution is constructed with an appropriate bin width 'h'. The midpoint points of these bins are denoted by x_i , i = 1,2…….k (number of bins).The corresponding frequencies are denoted by f_i , i = 1,2….k. Thus $\frac{f_i \times h}{n}$ is an estimate of the probability of y falling in the corresponding bin 'i' and is an estimate of the probability lying in the interval. Thus $f(x_i) \times h$ can be estimated by $\frac{f_i}{n}$ using the ratios of $f(x_i)$'s and equating them with corresponding observed frequency ratios gives a way of estimating the parameters similar to the moments method of estimation. Let f_1 and f_2 are the frequency densities at the points x_1 and x_2 given by

$$f_1 = \frac{a\,k^{\,a}}{x_1^{\,a+1}}; \quad f_2 = \frac{a\,k^{\,a}}{x_2^{\,a+1}}$$

The ratio of the frequencies is

$$\frac{f_1}{f_2} = \frac{x_2^{\,a+1}}{x_1^{\,a+1}} = \left(\frac{x_2}{x_1}\right)^{a+1}$$

Taking logarithms on both sides, we get

$$\log\left(\frac{f_1}{f_2}\right) = (a+1)\log\left(\frac{x_2}{x_1}\right)$$

On simplification, $\hat{a} = \dfrac{\log\left(\frac{f_1}{f_2}\right)}{\log\left(\frac{x_2}{x_1}\right)} - 1$ and $\hat{k} = \min(x_i)$

Illustration:

As explained earlier, we generate a sample of size 1000 from a Pareto (a=3, k=1) distribution using MATLAB function. For the generated data the following frequency distribution is obtained.

x	1.2501	1.7501	2.2501	2.7501	3.2501
f	677	193	72	28	0

Using these values in the above formula, the estimated value of a is

$$\hat{a} = \frac{\log\left(\frac{677}{193}\right)}{\log\left(\frac{1.7501}{1.2501}\right)} - 1 = 2.73$$

The above procedure is repeated for 50 samples. The mean, Standard Error, $\sqrt{\beta_1}$, $\beta_2$bias of these 50 estimates were computed. The estimated bias was calculated as the mean minus the true value of the parameter. The Mean Squared Error (MSE) was calculated as the bias squared plus the variance. The following results are obtained:

Table 2: Descriptive for LFR Method

	a	k
Mean	3.1496	1.0004
SE	0.2469	0.0004
$\sqrt{\beta_1}$	0.3798	1.4256
β_2	3.3016	4.4269
Bias	0.1496	0.0004
MSE	0.0834	0.0000

From the above tables, we notice that the actual values of (a, k) and the mean estimated values of (a, k) under the frequency ratio method and other methods are almost same. Therefore, it can be taken as a good estimator. Similar procedure is followed for different sample sizes and different values of (a, k) and the results are tabulated in the following tables.

3. COMPARISON OF METHODS OF MOMENTS, MLE AND FREQUENCY RATIO FOR DIFFERENT SAMPLE SIZES AND DIFFERENT PARAMETERS

Table 3: Simulation statistics for Pareto (3, 1, 100)

a=3; k=1	n_s = 100					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	**k**	**a**	**k**	**a**	**k**
Mean	3.2670	1.0377	3.0017	1.0003	3.1478	1.0003
SE	0.3981	0.0580	0.0979	0.0003	0.2748	0.0003
$\sqrt{\beta_1}$	0.6866	2.0851	0.1785	1.2986	0.4619	1.2986
β_2	3.6152	8.7724	2.9173	4.1652	2.8449	4.1652
Bias	0.2670	0.0327	0.0017	0.0003	0.1478	0.0003
MSE	0.2298	0.0044	0.0096	0.000	0.0974	0.000

Table 4: Simulation statistics for Pareto (3, 1, 200)

a=3; k=1	n_s = 200					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	k
Mean	3.2657	1.0317	3.0098	1.0003	3.1985	1.0003
SE	0.3979	0.0528	0.0966	0.0003	0.2415	0.0003
$\sqrt{\beta_1}$	0.5345	1.7159	0.3035	1.6101	0.1710	1.6101
β_2	3.1307	6.5251	3.2694	5.6326	2.6308	5.6326
Bias	0.2657	0.0317	0.0098	0.0003	0.1985	0.0003
MSE	0.2289	0.0038	0.0094	0.0000	0.0977	0.0000

Table 5: Simulation statistics for Pareto (5, 10, 50)

a=5; k=10	n_s = 50					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	k
Mean	4.9345	9.9662	4.9835	10.002	5.4030	10.002
SE	0.3443	0.1376	0.1563	0.0017	0.3534	0.0017

$\sqrt{\beta_1}$	0.0087	-0.5740	0.0651	1.1698	0.4410	1.1698
β_2	2.5867	2.7132	2.7175	4.2560	2.8721	4.2560
Bias	-0.065	-0.0338	-0.0165	0.0020	0.4030	0.0020
MSE	0.0044	0.0012	0.0003	3.93e-006	0.1625	3.93e-006

Table 6: Simulation statistics for Pareto (5, 10, 100)

a=5; k=10	n_s = 100					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	k
Mean	4.9345	9.9662	4.9835	10.0019	5.2030	10.0019
SE	0.3443	0.1376	0.1563	0.0019	0.3534	0.0019
$\sqrt{\beta_1}$	0.0087	-0.574	0.0651	1.8182	0.4410	1.8182
β_2	2.5867	2.7132	2.7175	6.8251	2.8721	6.8251
Bias	-0.065	-0.033	-0.0165	0.0019	0.4358	0.0019
MSE	0.0044	0.0012	0.0003	3.76e-006	0.1901	3.76e-006

Table 7 : Simulation statistics for Pareto (5, 10, 200)

a=5; k=10	n_s = 200					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	k
Mean	2.9221	3.9224	3.0154	4.0018	3.1798	4.0018
SE	0.2424	0.1480	0.0849	0.0018	0.2453	0.0018
$\sqrt{\beta_1}$	-0.134	-0.3781	0.1308	1.1341	0.3441	1.1341
β_2	1.9689	2.2147	2.4638	3.3688	2.5521	3.3988
Bias	-0.078	-0.0779	0.0154	0.0018	0.1798	0.0018
MSE	0.0061	0.0061	0.0002	3.34e-006	0.0324	3.34e-006

Table 8 : Simulation statistics for Pareto (3, 4, 50)

a=3; k=4	n_s = 50					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	k
Mean	4.9161	9.9523	4.9934	10.0019	5.4156	10.0019
SE	0.3390	0.1554	0.1618	0.0019	0.3240	0.0019
$\sqrt{\beta_1}$	-0.345	-0.9485	0.1258	1.5016	0.2530	1.5016
β_2	3.4192	4.7027	3.1781	5.1698	3.2370	5.1698
Bias	-0.084	-0.0477	-0.007	0.0019	0.4156	0.0019
MSE	0.0072	0.0023	0.0001	0.0000	0.1729	0.0000

Table 9 : Simulation statistics for Pareto (3, 4, 100)

a=3; k=4	n_s = 100					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	K
Mean	2.9445	3.9346	3.0000	4.0014	3.1930	4.0014
SE	0.2717	0.1683	0.1006	0.0013	0.2626	0.0013
$\sqrt{\beta_1}$	-0.171	-1.1393	0.0270	1.4220	-0.2190	1.4220
β_2	3.0482	5.1894	2.3433	6.0989	2.4703	6.0989
Bias	-0.055	-0.0654	4.15e-005	0.0014	0.1930	0.0014
MSE	0.0030	0.0043	1.01e-005	1.86e-006	0.0373	1.86e-006

Table 10: Simulation statistics for Pareto (3, 4, 200)

a=3; k=4	n_s = 200					
	Method of moments		Maximum Likelihood		Frequency Ratio Method	
	a	k	a	k	a	K
Mean	2.8898	3.8911	3.0047	4.0013	3.1855	4.0013
SE	0.3320	0.2376	0.0958	0.0011	0.2590	0.0011

$\sqrt{\beta_1}$	-0.631	-1.7891	-0.0263	1.1865	0.3217	1.1865
β_2	3.4086	7.4399	2.8725	4.0392	3.2658	4.0392
Bias	-0.110	-0.1089	0.005	0.0013	0.1855	0.0013
MSE	0.0123	0.0119	3.11e-005	1.58e-006	0.0345	1.58e-006

4. CONCLUSIONS

From the empirical study of the type of distribution, the estimates computed using the various estimation procedures including the one based on full information is reported for which the statistical distributions are summarized by its mean, standard error, $\sqrt{\beta_1}$, β_2 ,Bias and Mean square error computed from the simulated data.

We observe that the mean estimated values based on Method of moments/Method of Maximum likelihood estimation with full data and the Local frequency ratio method is nearly equal to the true value of the parameter. However, the standard errors of Local Frequency Ratio method are slightly more than that of the estimator based on full information sample. But in the particular case where out liars may affect the estimation procedure based on global information, this aspect is insignificant. Thus, when full information is available the local information based estimators are effectively as good as the corresponding Method of moments /Method of Maximum likelihood estimation with full information.

REFERENCES

Aris Spanos(1999). Probability Theory and Statistical Inference: United Kingdom at the University Press, Cambridge.

Hogg and Tanis (2001). Probability and Statistical Inference Sixth addition-Prentice-Hall publications, Newjersy.

Joseph Lee Peterson: Estimating the Parameters of a Pareto Distribution by a Quantile Regression method, **www.math.umt.edu/gideon/pareto.pdf**

Rao,C.R(1973). Linear Statistical Inference and its Applications-John Wiley and sons, New York.

Robert V. Hogg and Allent T. Craig (2002). Introduction to Mathematical Statistics-Pearson Publications, Singapore.

Rudra Pratap (2004). Getting Started with MATLAB-Oxford University press, New York.

Vijay K.Rohatgi and Ak Md Ehsames Saleh(2002). An Introduction to Probability and Statistics- John Wiley Inter science Publication.

Yugandhar.Ch , V.V.Haragopal, S.N.N.Pandit (2011). Local Information Based Parameter Estimation for Exponential distribution, ANU Journal of Physical Sciences,Vol.3,No 1&2,June-December 2011.

Received: 25.11.2013

Revised: 25.04.2015

Ch.Yugandhar[2] , V.V.Haragopal[1]

[1]Department of Statistics & Director, Center for Quantitative Methods, Osmania University, Hyderabad -500 007 (A.P.), India.

[2]Department of Statistics, St. Francis College for Women, Hyderabad – 500 016 (A.P.), India

E-mail*: haragopalvajjha@gmail.com, yug_0203@rediffmail.com*