Communications of the Korean Statistical Society
2011, Vol. 18, No. 6, 817–835

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.6.817

Efficient Estimation of the Parameters of the Pareto Distribution in the Presence of Outliers

Dixit, U. J.[1,a], Jabbari Nooghabi, M.[a]

[a]Department of Statistics, University of Mumbai

Abstract

The moment(MM) and least squares(LS) estimations of the parameters are derived for the Pareto distribution in the presence of outliers. Further, we have derived a mixture method(MIX) of estimations with MM and LS that shows that the MIX is more efficient. In the final section we have given an example of actual data from a medical insurance company.

Keywords: Pareto distribution, maximum likelihood estimator, moment estimator, least squares estimation, mixture method, outliers, insurance.

1. Introduction

The Pareto distribution is used as a model for insurance, business, economics, engineering, reliability, hydrology, and mineralogy. These models have been studied by Quandt (1996), Malik (1970), Asrabadi (1990), Hossain and Zimmer (2000), and Nadeau and Teorey (2003). The Pareto distribution was originally used to describe the allocation of wealth among individuals since it seemed to adequately show the way that a larger portion of the wealth in society is owned by a smaller percentage of the people in that society. It can be shown that from a probability density function(pdf), graph of the population $f(x)$, the probability or fraction of $f(x)$ that own a small amount of wealth per person is high. The probability then steadily decreases as wealth increases.

Another application of this distribution is for On-Line Analytical Processing(OLAP) view size estimation. Nadeau and Teorey (2003) used Pareto distribution for OLAP provides useful information quickly from large amounts of data residing in a data warehouse. To improve the quickness of response to queries, pre-aggregation is a useful strategy. However, it is usually impossible to pre-aggregate along all combinations of the dimensions. The multi-dimensional aspects of the data lead to combinatorial explosion in the number and potential storage size of the aggregates. Nadeau and Teorey (2003) suggested to selectively pre-aggregate. Cost/benefit analysis involves estimating the storage requirements of the aggregates in question. They (Nadeau and Teorey, 2003) presented an original algorithm to estimate the number of rows in an aggregate based on the Pareto distribution model. They also tested the Pareto model algorithm empirically against four published algorithms, and concluded that the Pareto model algorithm is consistently the best of these algorithms to estimate view size.

Hossain and Zimmer (2000) compared methods of estimation for the parameters of the Pareto distribution to determine which method provides the better estimates when the observations are censored.

[1] Corresponding author: Professor, Department of Statistics, University of Mumbai, Mumbai, India.
E-mail: Jabbarinm@yahoo.com; Jabbarinm@um.ac.ir

They used unweighted least squares(LS), the maximum likelihood estimate(MLE) and modified likelihood estimate(MML) for censored and uncensored data. They proposed that the LS method be generally preferred over the ML and MML methods to estimate the parameters of the Pareto distribution for complete samples.

Dixit and Jabbari Nooghabi (2011) have considered the Pareto distribution in the presence of outliers when the parameter α is unknown and the parameters β and θ are known. ML and uniformly minimum variance unbiased estimator(UMVUE) of α, the pdf, cumulative distribution function(cdf) and the r^{th} moment are derived. These estimators are compared empirically for their mean square error(MSE) and are investigated with the help of numerical technique. They have shown that MLE of pdf and cdf are better than the UMVUEs. In addition, it is shown that the expectation of the MLE of r^{th} moment does not exist. Finally, they have illustrated these methods with the help of real data from an insurance company.

Let a set of random variables $(X_1, X_2, \ldots, X_n)$ represent the claim amounts of a medical cure insurance company. It is assumed that the claims of some of passengers are β times higher than claims of the passengers who have normal injuries.

Hence, we assume that the random variables $(X_1, X_2, \ldots, X_n)$ are such that any k of them are distributed with pdf

$$f_2(x; \alpha, \beta, \theta) = \frac{\alpha(\beta\theta)^\alpha}{x^{\alpha+1}}, \quad 0 < \beta\theta \le x,\ \alpha > 0,\ \beta > 1,\ \theta > 0, \tag{1.1}$$

and the remaining $(n-k)$ random variables are distributed as

$$f_1(x; \alpha, \theta) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, \quad 0 < \theta \le x,\ \alpha > 0. \tag{1.2}$$

In this paper, we assume that all three parameters are unknown and we have derived the moment(MM), LS and mixture method(MIX) of MM and LS estimators of the parameters of the Pareto distribution in the presence of outliers. We have shown that the MLE of these parameters does not exist. In addition, it is shown that the MIX estimator of the parameters are more efficient than their MM. Finally, we give an example of claims for a medical insurance company.

2. Joint Distribution of $(X_1, X_2, \ldots, X_n)$ with k Outliers

The joint pdf of $(X_1, X_2, \ldots, X_n)$ in the presence of k outliers is given by

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \alpha, \beta, \theta) &= \frac{\alpha^n \theta^{n\alpha} \beta^{k\alpha}}{C(n,k)} \left(\prod_{i=1}^{n} x_i \right)^{-(\alpha+1)} \\ &\times \sum_{A_1=1}^{n-k+1} \sum_{A_2=A_1+1}^{n-k+2} \cdots \sum_{A_k=A_{k-1}+1}^{n} \prod_{j=1}^{k} \mathbf{I}(x_{A_j} - \beta\theta) \mathbf{I}(x_{A_j} - \theta), \end{aligned} \tag{2.1}$$

where $C(n,k) = n!/\{k!(n-k)!\}$ and $\mathbf{I}$ is the indicator function defined as

$$\mathbf{I}(y) = \begin{cases} 1, & y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that marginal distribution of X_i is

$$f(x_i; \alpha, \beta, \theta) = b\frac{\alpha(\beta\theta)^\alpha}{x_i^{\alpha+1}} \mathbf{I}(x_i - \beta\theta) + \bar{b}\frac{\alpha\theta^\alpha}{x_i^{\alpha+1}} \mathbf{I}(x_i - \theta), \quad \alpha > 0,\ \beta > 1,\ \theta > 0, \tag{2.2}$$