Test for the Pareto Distribution Based on the Transformed Sample Lorenz Curve

Suk-Bok Kang[1] and Young-Suk Cho[2]

Abstract

A powerful and easily computed goodness-of-fit test for Pareto distribution which does not depend on the unknown location and scale parameters is proposed based on the transformed sample Lorenz curve. We compare the power of the proposed test statistic with the other goodness-of-fit tests for Pareto distribution against various alternatives through Monte Carlo methods.

1. Introduction

A continuous random variable X has the Pareto distribution with the location parameter a, the scale parameter b, and the shape parameter c if it has a cumulative distribution function (cdf) of the form

$$F(x) = 1 - [1 + (x-a)/b]^{-c}, \quad x \ge c, \quad b, \quad c > 0. \tag{1.1}$$

Likes (1969) derived the uniformly minimum variance unbiased estimators (UMVUE) of the parameters in the Pareto distribution. Malik (1970) derived distributions of the maximum likelihood estimators (MLE) of the parameters in the Pareto distribution. Kulldorff and Vannman (1973) studied estimation of the location and scale parameters of the Pareto distribution. Woo and Kang (1990) considered a more general class of UMVUE for the function of two parameters in the Pareto distribution. Kang and Cho (1996) obtained the jackknife estimator and the generalized jackknife estimator, the minimum risk estimator (MRE) of two parameters in the Pareto distribution.

Moothathu (1985) derived the MLEs of the Lorenz curve and the Gini index of a Pareto distribution, their exact and asymptotic distributions and moments. Moothathu (1990) also obtained the UMVUE and a strongly consistent asymptotically normal unbiased estimator (SCANUE) of the Lorenz curve, the Gini index and Theil entropy index of a Pareto distribution. Kang and Cho (1999) proposed the several estimators of the Lorenz curve in the Pareto distribution.

Use of the Pareto distribution for practical applications can be enhanced by an accurate

1 Professor, Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea

2 Adjunct Assistant Professor, Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea

method of determining whether a set of data comes from a population governed by the Pareto distribution. One class of goodness-of-fit tests that can be used for this purpose consists of tests based on the distance between the empirical distribution function (edf) and the hypothesized cdf. Three of the better known tests in this class Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), Cramer-von Mises (C-vM) are valid when there are no unknown parameters in the hypothesized distribution.

2. Goodness-of-fit Tests

Let $X_{(j)}\ (j=1,2,\ldots,n)$ be the j-th order statistic based on a random sample $X_1, X_2, \ldots, X_n$ from the Pareto distribution with cdf (1.1). Consider now the case when shape parameter c is known and both the location parameter a and the scale parameter b are unknown. The best linear unbiased estimates (BLUEs) $\hat{a}$ and $\hat{b}$ were proposed by Kulldorff and Vannman (1973).

For $c>2$, the BLUEs are

$$\hat{a}=x_{(1)}-\frac{Y}{(nc-1)(c-2)-ncD)} \tag{2.1}$$

and

$$\hat{b}=(x_{(1)}-\hat{a})(nc-1) \tag{2.2}$$

where

$$\begin{aligned} B_i &= \left(1-\frac{2}{c(n-i+1)}\right)B_{i-1}, \text{for } i=1,2,\ldots n, \\ B_0 &\equiv 1, \\ D &= (c+1)\sum_{i=1}^{n-1} B_i+(c-1)B_n, \\ Y &= (c+1)\sum_{i=1}^{n-1} B_i x_{(i)}+(c-1)B_n x_{(n)}-Dx_{(1)}. \end{aligned}$$

For $2/n < c \leq 2$ and $2/c$ is an integer, the BLUEs are

$$\hat{a}=x_{(1)}-\frac{\hat{b}}{nc-1} \tag{2.3}$$

and

$$\hat{b}=\frac{(c+1)(c+2)(nc-1)}{(nc-2)(nc-c-2)}\left[\sum_{i=1}^{n-2/c} B_i x_{(i)}-\frac{(nc-2)}{(c+2)}x_{(1)}\right]. \tag{2.4}$$

The BLUEs were used to find the hypothesized cumulative distribution function $P_i=F(x_{(i)}, \hat{a}, \hat{b}, c)$, for $i=1,2,\ldots,n$. Then the values of the three modified test statistics were calculated.

The K-S statistic was computed from :

$$D = \max\{D^+, D^-\},$$

$$D^+ = \sup_{1 \le i \le n}\left[P_i - \frac{i-1}{n}\right],$$

$$D^- = \sup_{1 \le i \le n}\left[\frac{i}{n} - P_i\right].$$

The A-D statistic was computed from

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\log P_i + \log(1 - P_{n+1-i})).$$

The C-vM was computed from

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(P_i - \frac{2i-1}{2n}\right)^2.$$

This procedure was repeated 5,000 times for each sample size n, each shape parameter c with $a = b = 1$, and for all three tests. Critical values are contained in table by Porter III, Coleman, and Moore (1992). The null hypothesis that a set of sample data follows a Pareto distribution with specified shape parameter c is rejected at the desired significance level if the calculated value of the test statistic exceeds the table value.

The Lorenz curve is extensively used in the study of inequality distribution and used to be a powerful tool for the analysis of a variety of scientific problems. The Lorenz curve is given by

$$L(y) = \int_0^y x\, dF(x)/E(Y)$$

where Y is a nonnegative income variable for which the mathematical expectation $\mu = E(Y)$ exists.

Assume that $X_1, X_2, \ldots, X_n$ are positive random variables with order statistics $X_{(1)} < \ldots < X_{(n)}$. Let $r = [np]$ denote the greatest integer less than or equal to np. Then the sample Lorenz curve (Gail and Gastwirth (1978)) is defined by

$$L_n(p) = \frac{\sum_{i=1}^{r=[np]} X_{(i)}}{\sum_{i=1}^{n} X_{(i)}}.$$

Cho *et al.* (1999) proposed the transformed Lorenz curve that can be used in the study of symmetric distribution. The transformed Lorenz curve is defined by

$$TL(p) \equiv L(p) - p + 1.$$

To test $H_0 : X \sim F(x)$, Kang and Cho (2001) proposed Normalized Sample Lorenz Curve (*NSLC*) . The *NSLC* is defined by

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)}, \quad p = i/n, \quad i = 1,2,\cdots,n$$

where

$$TSL(p) \;=\; \frac{\sum_{j=1}^{i}(X_{j:n} - X_{1:n})}{\sum_{j=1}^{n}(X_{j:n} - X_{1:n})} - p + 1,$$

$$TSL_F(p) \;=\; \frac{\sum_{j=1}^{i}(F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))}{\sum_{j=1}^{n}(F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))} - p + 1.$$

We propose test statistic based on $NSLC$ for the pareto distribution as follows.

$$TS = NSLC_{par}(0.5)$$

where

$$NSLC_{par}(p) = \frac{TSL(p)}{TSL_{par}(p)}, \quad p = i/n, \quad i = 1,2,\cdots,n$$

$$TSL(p) \;=\; \frac{\sum_{j=1}^{i}(X_{j:n} - X_{1:n})}{\sum_{j=1}^{n}(X_{j:n} - X_{1:n})} - p + 1,$$

$$TSL_{par}(p) \;=\; \frac{\sum_{j=1}^{i}((1 - j/(n+1))^{1/c} - (1 - 1/(n+1))^{1/c})}{\sum_{j=1}^{n}((1 - j/(n+1))^{1/c} - (1 - 1/(n+1))^{1/c})} - p + 1.$$

References

[1] Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999). A study on distribution based on the Transformed Lorez Curve. *The Korean Journal of Applied Statistics,* Vol. 12(1), 153-163.

[2] Gail, M. H. and Gastwirth, J. L. (1978). A Scale-Free Goodness-of-Fit test for the exponential distribution based on Lorenz curve. *Journal of American Statistical Association*, Vol. 73, 787-793.

[3] Kang, S. B. and Cho, Y. S. (1996). Estimation of the Parameters in a Pareto Distribution by Jackknife and Bootstrap Method. *Journal of Information & Optimization Sciences,* Vol., 18(2), 289-300.

[4] Kang, S. B. and Cho, Y. S. (2001). A study on Distribution Based on the Normalized Sample Lorenz Curve. *The Korean Communications in Statistics,* Vol. 8(1), 185-192.

[5] Kulldorff, G. and Vannman, K. (1973). Estimation of the Location and Scale Parameters of a Pareto Distribution by Linear Functions of Order Statistics, *Journal of American Statistical Association*, Vol. 68, 218-227.

[6] Likes, J. (1969). Minimum Variance Unbiased Estimation of the Parameters of power-function and Pareto's Distribution, *Statistische Hefte*, 10, 104-110.

[7] Malik, H. J. (1970). Estimation of the Parameters of the Pareto Distribution, *Metrika*, Vol. 16, 126-132.

[8] Moothathu, T. S. K. (1985). Sampling Distribution of Lorenz Curve and Gini Index of the Pareto Distribution. *Sankhyā*, Vol. 47(B), 247-278.

[9] Moothathu, T. S. K. (1990). The Best Estimator of Lorenz Curve, Gini Index and Theil Entropy Index of Pareto Distribution. *Sankhyā*, Vol. 52(B), 115-127.

[10] Woo, J. and Kang, S. B. (1990). Estimation for Functions of Two Parameters in the Pareto Distribution, *Youngnam Statistical Letters*, Vol. 1, 67-76.