Pakistan Journal of Applied Sciences 2 (4):488-490, 2002

Kolmogorov Smirnov Test for Generalized Pareto Distribution

[1]M. Arshad, [2]M. T. Rasool and [3]M. I. Ahmad
[1]University College of Administrative Sciences, Kotli, Azad Kashmir
[2]Government College Samundri, District Faisalabad, Pakistan
[3]Department of Mathematics and Statistics, College of Sciences
Sultan Qaboos University Musqat, Oman

Abstract: The Kolmogorov- Smirnov Statistic is considered for testing the goodness of fit of the three parameter Generalized Pareto distribution. The statistic for testing the goodness of fit of the completely specified distribution are modified by replacing the Generalized Pareto distribution parameters by their probability weighted moments estimates. The Table of critical values is derived for this empirical distribution function test for various sample sizes.

Keywords: Kolmogorov Smirnov, Pareto Distribution, Empirical Distribution

Introduction

The Generalized Pareto (GP) distribution was introduced by Pickands (1975) and has since been further studied by Davison (1984); Smith (1984; 1985) and Van Montfort and Witter (1985). Van Mortfort and Witter (1986) have demonstrated its application to the distribution of peaks over threshold of rainfall series , using the maximum likelihood method for estimation. Hosking and Wallis (1987) used the method of moments and the method of probability weighted moments for the estimation of the GP distribution. Using Monte carlo simulation, they concluded that the estimates obtained by either Method of Moments (MOM), Probability Weighted Moments (PWM) were more reliable than the Maximum Likelihood Estimates (MLE). They also Observed that the GP distribution gave a better fit to large peaks, thus suggesting its use for modeling Peaks Over Threshold (POT). The GP distribution was also used by Wang (1991) for the comparison of POT and Annual Maximum (AM) models by PWM method. A Comparative study for the estimates of the GP distribution was made by Moharram, Gosain and Kapoor (1993) and concluded that PWM method is best when the value of the shape parameter is less than zero and particularly if shape parameter might be less than -0.2, then PWM estimates will probably be preferred because of their low bias.

Ever since the GP distribution has applications in a number of fields, including reliablity studies, in the modeling of large insurance claims, as a failure-time distribution. It is frequently used a model in the study of income distribution (Aigner and Goldberger 1970). Its applications include use in the analysis of extreme events e. g. for the analysis of the precipitation data, In the flood Frequency analysis, In the analysis of the data of greatest wave heights or sea levels, maximum winds loads on building, in the maximum rain fall analysis, in the analysis of greatest values of Yearly floods, breaking strength of materials, air craft loads etc. In other words it is used in any situation in which the Exponential distribution might be used but in which some robustness is required against heavier tailed or lighter tailed alternatives. The GP distribution has been quite popular not only for flood frequency analysis but for fitting the distribution of extreme natural events in general.

Once a distribution function is assumed or selected for study at hand ,it remains to estimate its parameters and when the parameters of the model are estimated, it is then desirable to access how well the distribution fits the observed data. Goodness of fit tests are often essential to reveal departures from the assumed model. In this study the parameters are estimated by PWM method and the critical points are derived for the Kolmogorov-Smirnov test for the Generalized Pareto distribution.

Materials and Methods

The Generalized Pareto Distribution: A random variable X is said to be distributed as Generalized Pareto (GP) distribution, if its distribution function is of the following form:

$$F(X) = 1=(1-cz)^{1/c} \quad c \neq 0$$
$$= 1- \exp(z) \quad c=0$$

where

$$z = \frac{x-b}{a}$$

Where a=scale parameter, b=threshold parameter, c=shape parameter.

As a definition, the Gp distribution combines in to a single form the three types GP-1,GP-2 and GP-3 or LomaX distribution corresponding to c=0, c<0 and c>0 respectively, the Generalized Pareto distribution reduces to the two parameter exponential distrbution when c=0. When c=1 it reduces to the uniform distribution on the interval $0 \leq x \leq a$

There are so many methods of estimation, e. g. method of moments, method of least squares, method of maximum likelihood, ordinary least squares, generalized least squares, PWM etc. But here We have used the probability weighted moments method, because this method gives more accurate estimates of the parameters of a fitted distribution in small samples than the MLE estimates. Experience also shows that as compared with conventional moments L- moments are less subject to bias in estimation.

Greenwood (1979) defined the probability weighted moments of X as

$$M_{p,r,s}=E[X^p \{F(X)\}^p \{1-F(X)\}^s]$$
$$= \int X^p \{F(X)\}^r \{1-F(X)\}^s dF(X)$$

Where p, r, s are the integers.

The parameters of the GP distribution are estimated as

$$\hat{c} = \frac{5Y_1 - 3Y_2}{Y_2 - Y_1}$$

Where $Y_1 = 2\beta_1 - \beta_0$, $Y_2 = 3\beta_2 - \beta_1$

$$\hat{a} = Y_1 (2 + c) (1 + c)$$

and

$$\hat{b} = \beta_0 - \frac{\hat{a}}{1+\hat{c}}$$

Goodness of fit techniques: Goodness of fit techniques means the methods of examining how well a sample of data agrees with a given distribution as its population.

- Test of chi-square types
- Moment ratio techniques
- Tests based on correlation
- Tests based on empirical distribution function

Most of these test statistics suffer from serious limitations. In general test of chi-square type have less power due to loss of information caused by grouping. The distribution theory of chi-squared statistics is a large sample theory. The higher order moments are usually under estimated and this fact prevents the use of moment ratio techniques and so would be the case with correlation types tests.

Several power studies have revealed Empirical Distribution Function (EDF) tests to be more powerful than other tests of fit for a wide range of sample sizes (Stephen, 1974; 1977). Recently satisfactory use of EDF tests has been difficult due to lack of readily available Tables of significance points for the case where the parameters of the assumed distribution have to be estimated from the sample data. The significance points that have been available are appropriate to the case where the parameters of the distribution are known. This case is referred to by stephen (1974; 1977) as case zero. Such Tables are of limited value in practice because the parameters of the distribution are seldom known. When the parameters are estimated the critical values are considerably smaller than for the specified parameters case. Thus the use of these critical values which are for specified parameters case to assess the agreement of a theoretical distribution when parameters are estimated from the data may result in accepting fitted distribution that ought to be rejected.

Important EDF Tests: Kolmogorov-Smirnov (KS) test, Anderson-Darling (AD) test and Cramer-Von-Mises (CVM) test . These tests can be used for smaller sample sizes. The use of these tests has been restricted until recently due to lack of percentage points of these testing when the parameters of the model are to be estimated from sample data.

We have derived the critical points for kolmogorov-Smirnov test statistic for the Generalized Pareto distribution for various sample sizes when the parameters are estimated by the probability weighted moments method.

Empirical distribution function: Suppose a given random sample of size n is $X_1, X_2, \ldots, X_n$ and let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistic and also suppose that the cumulative distribution function of X is F(x). The EDF is defined as:

$$F_n(X) = \frac{\text{Number of observations} \le x}{n}$$

More precisely, the definition is

$$F_n(x) = 0 \qquad X < X_{(i)}$$

$$F_n(x) = \frac{i}{n} \qquad X_{(i)} \le x \le X_{(i+1)} \quad i = 1, 2, \ldots, n-1$$

$$F_n(x) = 1 \qquad X_{(n)} \le x$$

Thus $F_n(x)$ is a step function which shows the proportion of observations less than or equal to x for any x. While F (x) is the probability of an observation less than or equal to x ($F(x) = P(X \le x)$).

EDF Statistic: A statistic measuring the difference between $F_n(x)$ and F (x) will be called an EDF statistic. The most well known EDF statistic is D Introduced by Kolmogorov-Smirnov.

$$D^+ = \text{Sup} \{F_n(x) - F(x)\}$$

$$D^- = \text{Sup} \{F(x) - F_n(x)\}$$

and

$$D = \text{Sup} \mid F_n(x) - F(X) \mid = \max(D^+, D^-)$$

For computational purposes the Kolmogorov-Smirnov test statistic D is:

Let

$$D^+ = \max \left(\frac{i}{n} - F(X_i)\right)$$

and

$$D^- = \max \left(F(X_i) - \frac{i-1}{n}\right)$$

then

$$D = \max (D^+, D^-)$$

The PWM estimates are denoted by D, F(x) is the cdf and x_i is the order statistic.

Simulation Procedures and Results: Our object was to workout the tall area probabilities for the distribution of Kolmogorov -Smirnov test for Generalized Pareto distribution when the parameters are estimated by probability weighted moments. For this purpose simulation procedure was used to approximate the distribution of KS tests. It is very difficult to derive the KS distribution mathematically.

The Generalized Pareto random variable X was generated by taking a random sample size n = 10, for each value of the shape parameter (-0.2, -0.1, 0.1, 0.2). Each sample was used to calculate the probability weighted moment estimates. By using these parameter estimates, the Kolmogorov-Smirnov test statistic values D were calculated. This process was repeated 1000 times for each value of the shape parameter.

Thus comprising a total 4000 valves for D and 50th, 75th, 85th, 90th, 95th, and 99th percentiles were found from this Kolmogorov-Smirnov distribution.

This whole procedure was repeated for the sample sizes 15, 20, 25, 30, 35, 40, 50, 100. A computer programme was developed in MINITAB, a statistical computer package, and the following Table 1 for critical values of Kolmogove-Smirnov test for Generalized Pareto distribution when the parameters are estimated by PWM was developed. The critical values of KS test for a specified case are given in Table 2.

Table 1: Critical Values of KS test for GP distribution when the parameters are estimated

n/P	0.50	0.25	0.15	0.10	0.05	0.01
10	0.162	0.189	0.206	0.219	0.239	0.281
15	0.137	0.161	0.176	0.185	0.202	0.231
20	0.121	0.141	0.154	0.164	0.180	0.208
25	0.109	0.128	0.141	0.148	0.162	0.195
30	0.103	0.121	0.132	0.140	0.153	0.174
35	0.094	0.111	0.120	0.127	0.138	0.165
40	0.087	0.104	0.113	0.120	0.129	0.149
50	0.081	0.094	0.103	0.111	0.119	0.142
100	0.057	0.067	0.073	0.077	0.084	0.094

Table 2: Critical Values of KS test for a Specified Case

n/P	0.20	0.10	0.05	0.02	0.01
10	0.323	0.369	0.409	0.457	0.489
15	0.226	0.304	0.338	0.337	0.404
20	0.232	0.265	0.294	0.329	0.352
25	0.208	0.238	0.264	0.295	0.317
30	0.190	0.218	0.242	0.270	0.290
35	0.177	0.202	0.224	0.251	0.269
40	0.165	0.189	0.210	0.235	0.252
50	0.139	0.173	0.192	0.215	0.230
100	0.107	0.122	0.136	0.152	0.163

Conclusion

By comparing these two Tables we have seen that the critical values for the specified case are larger than the case when the parameters of the distribution are estimated from the sample data. So by using these critical values which are for specified parameters case to access the agreement of a theoretical distribution when parameters are estimated from the data may result in accepting fitted distribution that ought to be rejected.

References

Aigner, D. J and Goldberger, A. S., 1970. Estimation of pareto's law from grouped observation J. the Amer. Stat. Assoc. 713-727.

Davidon,A. C., 1984. Modeling axcesess over high threshold with an application. In:J. Tiago de Oliveria (editor), Statistical extremes and applications. Riedl, Dordrecht, pp: 461-482.

Greenwood, J. A., Landwehr, J. M., 1979. Matalas, W. C. Probability weighted moments. Definition and relation to parameters of distribution expressible in inverse from. Wather Resour. Res. 15: 1049-1054.

Hosking, J. R. M and Walli,J. R., 1987. Parameter and quantile estimators for Generalized Pareto distribution. Technometrics, 29: 339-349.

Moharram, S. H., Gosain, A. K., Kapoor, P. N., 1993. A comparative study for the estimators of the Generalized Pareto distribution. J. of Hydrology, 150; 169-185.

Pickend, J., 1975. Statistical inference using extreme order statistics. Ann. Stat. 3: 119-131.

Stephens, M. A. , 1977. Goodness of fit for the extreme value distribution. Biomatrika, 64: 585-588.

Smith, R. L., 1984. Threshold methods for sample extreme and application. Reidl, Dordrecht, PP. 621-638.

Smith, R. L., L., 1985. Maximum likelihood estimation in a class of non reguler cases. Biometrika, 72: 67-90.

Stephen, M. A., 1974. EDF Statistic for goodness of fit and som comparison. J. Amer. Stat. Assoc. 69: 730-737.

Van Montfort, M. A. J. and Witter, J. V., 1986. The Generalized Pareto distribution applied to rainfall depth. Hydrol. Sci. J., 31: 151-162.

Van Montfort, M. A. J. and Witter, J. V., 1985. Testing exponentially aginst Generalized Pareto distribution. J. Hydrol., 78: 305-315.

Wang, Q. J., 1991. The pot model described by the Generalized Pareto distribution with poisson arrival rate J. Hydrol., 129: 263-280.