Journal of Sciences, Islamic Republic of Iran 28(3): 267 - 272 (2017) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Detecting Outliers in Exponentiated Pareto Distribution

M. Jabbari Nooghabi*

Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Islamic Republic of Iran

Received: 30 April 2016 / Revised: 26 October 2016 / Accepted: 3 January 2017

Abstract

In this paper, we use two statistics for detecting outliers in exponentiated Pareto distribution. These statistics are the extension of the statistics for detecting outliers in exponential and gamma distributions. In fact, we compare the power of our test statistics based on the simulation study and identify the better test statistic for detecting outliers in exponentiated Pareto distribution. At the end, we describe an example from insurance company.

Keywords: Exponentiated Pareto sample; Z statistic, Dixon's statistic; Outliers; Upper outlier.

Introduction

The Pareto distribution was originally used to describe the allocation of wealth among individuals, since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. It can be shown that from a probability density function, 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 decreases steadily as wealth increases. Also, the Pareto distribution is useful for finding the average of annuity and benefit for an insurance problem. In economics, where this distribution is used as an income distribution, the threshold parameter is some minimum income with a known value. Dixit and Jabbari Nooghabi [4] compared the uniformly minimum variance unbiased estimator (UMVUE) of the probability density function (pdf), the distribution function (CDF) and the r^{th} moment for the Pareto distribution.

Now, if we assume that Y is a Pareto distributed random variable, then we take $X = \ln(Y)$ to have the corresponding exponentiated Pareto distribution as defined by Nadarajah [14]. Usually, Y is defined on the positive side of the real line and so one would hope that models on the basis of the distribution of X would have greater applicability. Nadarajah [14] introduced five exponentiated Pareto distributions and derived several of their properties including the moment generating function, expectation, variance, skewness, kurtosis, Shannon entropy, and the Rényi entropy. Note that another type of exponentiated Pareto distribution was considered by Shawky and Abu-Zinadah [15] and characterized using record values. Shawky and Abu-Zinadah [16] derived the maximum likelihood estimation of the different parameters of an exponentiated Pareto distribution. Also, they considered five other estimation procedures and compared them. Afify [2] obtained Bayes and classical estimators for a two parameter exponentiated Pareto distribution for when samples are available from complete, type I and type II censoring schemes. He proposed Bayes estimators under a squared error loss function as well as under a LINEX loss function using priors of non-informative type for the parameters.

Mahmoud [13] proposed the best linear unbiased estimates and the maximum likelihood estimates of the

* Corresponding author: Tel: +985138805694; Fax: +985138807155; Email: jabbarinm@um.ac.ir; jabbarinm@yahoo.com

location and scale parameters from the Exponentiated Pareto distribution based on progressively Type-II right censored order statistics.

However, in this paper we will restrict to the form defined by Nadarajah [14].

The generalized Chauvenet's test for rejecting outlier observations is suitable for detecting k outliers in a univariate data set. This test can be used for exponential case. Several authors considered the problem for testing one outlier in exponential distribution. Only two types of statistics for testing multiple outliers exist. First is Dixon's while the second is based on the ratio of some observations suspected to be outliers with respect to the sum of all observations in the sample. In fact, most of these authors have used a general case of gamma model and then the results for exponential model are given as a special case. This approach is focused on alternative models, namely slippage alternatives in exponential samples (see Barnett and Lewis [3]). Barnett and Lewis [3] gave a survey of literature in the connection. Kale [8] investigated the problem of identifying the outliers for one parameter exponential family. Zerbet and Nikulin [17] proposed a different statistic from the well-known *Dixon's statistic*, D_k, to test multiple outliers. Hadi et al. [6] presented an overview of the major developments in the area of detection of outliers. These include projection pursuit approaches as well as Mahalanobis distance-based procedures. Also, they discussed other methods, corresponding to the large datasets. Jabbari Nooghabi *et. al.* [7] and Kumar and Lalitha [11] extended the Zerbet and Nikulin [17] statistic for gamma distribution and showed that Z_k statistic is more powerful than Dixon's. Kumar [10] discussed an approach for testing multiple upper outliers with slippage alternative in an exponential sample irrespective of origin. The test statistic is based on a ratio of two estimates, obtained by the maximization of the two log-likelihood functions. He derived the exact null distribution of the test statistic. Kornacki [9] proposed an alternative method of outlier detection based on the Akaike information criterion. Lin and Balakrishnan [12] proposed an algorithm for evaluating the null joint distribution of Dixon-type test statistics for testing discordancy of k upper outliers in exponential samples. Gogoi and Das [5] compared the empirical powers of some statistics for detecting multiple upper outliers in exponential samples under slippage alternative. The results show that the maximum likelihood ratio test statistic is better than the other statistics followed by Dixon type test statistics to deal with upper outliers in exponential samples. Adil and Irshad [1] modified the Tukey's boxplot for detection of outliers when the data are skewed and proposed approach to detect outliers properly.

In this paper, we use two statistics Z_k and D_k for detecting outliers in exponentiated Pareto distribution. The distribution of the test based on these statistics under slippage alternatives is obtained and the tables of critical values are given for various n (the sample size) and k (the number of outliers). The power of these tests are also calculated and compared. In the next section, we introduce the test statistics. In Sections 3 and 4, we obtain the distribution of the statistics. Section 5 used to compare the critical values and the powers. In the last section, we describe an example from an insurance company.

Statistical Inference

Let $X_1, X_2, \ldots, X_n$ be arbitrary independent random variables. In this paper, we test the following hypothesis:

H_0: $X_1, X_2, \ldots, X_n$ are iid random variables from exponentiated Pareto distribution with parameters α and θ (α is unknown and θ is known).

Therefore, the probability density function of these samples under the null hypothesis is:

$f_X(x;\alpha,\theta) = \alpha\theta^{\alpha}e^{-\alpha x}. \quad x \geq \ln(\theta) > 0,\ \alpha > 0$

But under the *slippage alternative*, H_k, we have

$X_{(1)}, X_{(2)}, \ldots, X_{(n-k)}$ derive from $f_X(x;\alpha,\theta)$,

$X_{(n-k+1)}, X_{(n-k+2)}, \ldots, X_{(n)}$ derive from $f_X(x;\alpha\beta,\theta)$,

where $\beta > 1$, β is unknown and $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics corresponding to the observations $X_1, X_2, \ldots X_n$. We suppose that the hypothesis be an important sub-hypothesis of the one saying that k of n observations are suspected to be outliers (for $\beta > 1$, these k observations are called *upper outliers*). So, H_0 is correspond to $\beta = 1$.

To test H_0, we use these statistics

$$Z_k = \frac{X_{(n-k)} - X_{(1)}}{\sum_{j=n-k+1}^{n}(X_{(j)} - X_{(1)})}, \qquad (1)$$

and

$$D_k = 1 - \frac{X_{(n-k)} - \ln(\theta)}{X_{(n)} - \ln(\theta)}. \qquad (2)$$

Following the idea of the Chauvenet's test, we assume that the decision criterion is:

$$H_0 \ is\ rejected\ when\ Z_k > c_1\ or\ D_k > c_2,$$

where $c_1 = c_1(\alpha_1)$ and $c_2 = c_2(\alpha_1)$ are the critical value corresponding to the significance level α_1 for Z_k and D_k statistics, respectively.

The Distribution of Z_k Under Alternatives

In this section, we find the distribution of the statistic Z_k according to Zerbet and Nikulin [17] method. Then,

the distribution of this statistic under the slippage alternative hypothesis H_k is obtained by the following Theorem.

Theorem 1. The distribution of the statistic Z_k under H_k is

$$P_k\{Z_k < z|H_k\} = \frac{\beta^k \Gamma(k\beta+n-k)}{\Gamma(k\beta+1)} \sum_{j=2}^{n-k} \frac{(-1)^{n+j-k}}{\Gamma(j-1)\Gamma(n-j-k+1)(k\beta+n-k-j+1)} \times \left\{\beta^{-k} - \left[\beta + (k\beta+n-k-j+1)\frac{z}{1-kz}\right]^{-k}\right\}, \quad 0 < z < \frac{1}{k}. \tag{3}$$

Proof. To proof see Zerbet and Nikulin [17] or Jabbari Nooghabi *et. al.* [7] and and Kumar and Lalitha [11].

Corollary 1. The distribution of Z_k under H_0 is obtained from Theorem 1 by taking $\beta = 1$.

The Distribution of D_k Under Alternatives

In this section, the following Theorem is used to find the distribution of D_k under alternatives.

Theorem 2. The distribution of D_k under H_k is as follows:

$$P_k\{D_k < x|H_k\} = \frac{\beta\Gamma(k\beta+n-k+1)}{\Gamma(k\beta+1)} \sum_{j=1}^{n-k}\sum_{i=1}^{k} \frac{(-1)^{n+i+j}}{(n-k-j)!\,(j-1)!\,(k-i)!\,(i-1)!} \times \frac{1}{k\beta+n-k-j+1}\left[(\beta(k-i+1))^{-1} - (\beta(k-i+1) + \left(\frac{1}{x}-1\right)(k\beta+n-k-j+1))^{-1}\right], \quad 0 < x < 1. \tag{4}$$

Proof. Same as the Theorem 1, we set

$$R_k = \frac{\sum_{j=1}^{n-k} Y_j}{\sum_{j=n-k+1}^{n} Y_j} = \frac{P}{Q}. \tag{5}$$

The characteristic function of (P, Q) is

$$\phi_{P,Q}(t,s) = \boldsymbol{E}\left(e^{i(Pt+Qs)}\right) = \int_{\Re^n} e^{i(\sum_{j=1}^{n-k} y_j t + \sum_{j=n-k+1}^{n} y_j s)} \times f_{(Y_1,Y_2,\dots,Y_n)}(y_1, y_2, \dots, y_n)\,dy_1 dy_2 \dots dy_n. \tag{6}$$

Then according to distribution of Y_j, $j = 1,2,\dots,n-k$ and Y_{n-k+j}, $j = 1,2,\dots,k$, we have

$$\phi_{P,Q}(t,s) = \prod_{j=1}^{n-k}\left[\int_0^{\infty} \frac{1}{a_j} e^{-y_j\left(\frac{1}{a_j}-it\right)} dy_j\right] \times \prod_{j=1}^{k}\left[\int_0^{\infty} \frac{1}{b_j} e^{-y_{n-k+j}\left(\frac{1}{b_j}-is\right)} dy_{n-k+j}\right] \prod_{j=1}^{n-k}\frac{1}{a_j}\left(\frac{1}{a_j}-it\right)^{-1} \times \prod_{j=1}^{k}\frac{1}{b_j}\left(\frac{1}{b_j}-is\right)^{-1}.$$

So, the joint density function of (P, Q) is

$$f_{(P,Q)}(p,q) = \frac{1}{(2\pi)^2}\int_0^{+\infty}\left[\prod_{j=1}^{n-k}\frac{1}{a_j}\left(\frac{1}{a_j}-it\right)^{-1} e^{-itp}\right]dt \times \int_0^{+\infty}\left[\prod_{j=1}^{k}\frac{1}{b_j}\left(\frac{1}{b_j}-is\right)^{-1} e^{-isq}\right]ds. \tag{7}$$

Same as Zerbet and Nikulin [17], we obtain these products

$$\prod_{j=1}^{n-k}\frac{1}{\frac{1}{a_j}-it} = \sum_{j=1}^{n-k}\frac{(-1)^{n-k+j}}{\left(it-\frac{1}{a_j}\right)(j-2)!\,(n-k-j)!\,\alpha^{n-k}}, \tag{8}$$

$$\prod_{j=1}^{n-k}\frac{1}{a_j} = \frac{\Gamma(k\beta+n-k+1)\alpha^{n-k}}{\Gamma(k\beta+1)}, \tag{9}$$

and

$$\prod_{j=1}^{k}\frac{1}{\frac{1}{b_j}-is} = \sum_{j=1}^{k}\frac{(-1)^{k+j}}{\left(is-\frac{1}{b_j}\right)(j-2)!\,(k-j)!\,(\alpha\beta)^{k-1}}. \tag{10}$$

Therefore

$$f_{(P,Q)}(p,q) = \frac{\alpha\Gamma(k\beta+n-k+1)}{\Gamma(k\beta+1)}\sum_{j=1}^{n-k}\frac{(-1)^{n-k+j}}{(j-1)!\,(n-k-j)!}e^{-\alpha(k\beta+n-k-j+1)p} \times \alpha\beta k!\sum_{i=1}^{k}\frac{(-1)^{k+i}}{(k-i)!\,(i-1)!}e^{-\alpha\beta(k-i+1)q}, \quad p > 0,\ q > 0. \tag{11}$$

So, the density and distribution function of R_k are

$$f_{R_k}(r) = \frac{\beta\Gamma(k\beta+n-k+1)}{\Gamma(k\beta+1)}\sum_{j=1}^{n-k}\sum_{i=1}^{k}\frac{(-1)^{n+j+i}[\beta(k-i+1)+r(k\beta+n-k-j+1)]^{-2}}{(n-k-j)!\,(j-1)!\,(k-i)!\,(i-1)!}, \tag{12}$$

and

$$P_k\{R_k < r\} = \frac{\beta\Gamma(k\beta+n-k+1)}{\Gamma(k\beta+1)}\sum_{j=1}^{n-k}\sum_{i=1}^{k}\frac{(-1)^{n+i+j}}{(n-k-j)!\,(j-1)!\,(k-i)!\,(i-1)!} \times\frac{1}{k\beta+n-k-j+1}[(\beta(k-i+1))^{-1}-(\beta(k-i+1)+r(k\beta+n-k-j+1))^{-1}],\quad r>0. \qquad (13)$$

With substituting $R=\frac{1}{D}-1$, the proof is complete.

Corollary 2. The distribution of D_k under H_0 is obtained from Theorem 2 by taking $\beta=1$.

Results

A Simulation Example

In this section, we give the critical values of statistics Z_k and D_k for the levels of significance $\alpha_1=0.05$ and

Table 1. Critical values of Z_k for $\alpha_1=0.05$ and $\alpha_1=0.1$

	k									
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	0.95438	0.39434	0.18751	--	--	--	--	--	--	--
	0.90981	0.35284	0.15470	--	--	--	--	--	--	--
6	0.96130	0.41440	0.22380	--	--	--	--	--	--	--
	0.92308	0.37940	0.19632	--	--	--	--	--	--	--
7	0.96562	0.42604	0.24199	0.14622	--	--	--	--	--	--
	0.93148	0.39508	0.21788	0.12683	--	--	--	--	--	--
8	0.96866	0.43375	0.25326	0.16164	--	--	--	--	--	--
	0.93735	0.40558	0.23145	0.14416	--	--	--	--	--	--
9	0.97086	0.43928	0.26101	0.17163	0.11669	--	--	--	--	--
	0.94170	0.41318	0.24089	0.15555	0.10343	--	--	--	--	--
10	0.97262	0.44348	0.26672	0.17872	0.12540	--	--	--	--	--
	0.94509	0.41897	0.24792	0.16376	0.11305	--	--	--	--	--
15	0.97766	0.45527	0.28212	0.19689	0.14632	0.11281	0.08880	0.07062	--	--
	0.95503	0.43539	0.26708	0.18502	0.13660	0.10457	0.08174	0.06438	--	--
20	0.98022	0.46098	0.28933	0.20493	0.15513	0.12227	0.09899	0.08161	0.06808	0.05719
	0.96009	0.44343	0.27608	0.19458	0.14666	0.11515	0.09289	0.07626	0.06331	0.05291
25	0.98182	0.46449	0.29359	0.20966	0.16013	0.12759	0.10456	0.08742	0.07415	0.06361
	0.96326	0.44839	0.28151	0.20023	0.15245	0.12116	0.09903	0.08260	0.06987	0.05978
30	0.98294	0.46691	0.29654	0.21283	0.16347	0.13105	0.10812	0.09112	0.07796	0.06754
	0.96555	0.45183	0.26400	0.20403	0.15633	0.12509	0.10303	0.08667	0.07402	0.06404

Upper and lower values in each cell refer to $\alpha_1=0.05$ and $\alpha_1=0.1$, respectively.

Table 2. Critical values of D_k for $\alpha_1=0.05$ and $\alpha_1=0.1$

	k									
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	0.78208	0.86148	0.87817	--	--	--	--	--	--	--
	0.71366	0.80173	0.76693	--	--	--	--	--	--	--
6	0.74585	0.81986	0.82261	--	--	--	--	--	--	--
	0.67517	0.75459	0.69865	--	--	--	--	--	--	--
7	0.71730	0.78668	0.77956	0.85710	--	--	--	--	--	--
	0.64537	0.71814	0.64927	0.85709	--	--	--	--	--	--
8	0.69403	0.75961	0.74518	0.83328	--	--	--	--	--	--
	0.62143	0.68892	0.61147	0.80006	--	--	--	--	--	--
9	0.67467	0.73700	0.71701	0.79997	0.87497	--	--	--	--	--
	0.60165	0.66489	0.58145	0.77781	0.85711	--	--	--	--	--
10	0.65818	0.71778	0.69337	0.77774	0.85707	--	--	--	--	--
	0.58495	0.64467	0.55689	0.75003	0.83328	--	--	--	--	--
15	0.60137	0.65176	0.61466	0.76831	0.81815	0.81824	0.83327	0.84612	--	--
	0.52828	0.57662	0.47831	0.73966	0.74998	0.75004	0.76928	0.78577	--	--
20	0.56667	0.61169	0.56847	0.62497	0.63641	0.64284	0.68747	0.71433	0.73338	0.76475
	0.49425	0.53625	0.43439	0.57599	0.59998	0.60489	0.61536	0.64288	0.66663	0.68748
25	0.54244	0.58384	0.53706	0.59997	0.61537	0.61536	0.64704	0.68746	0.68923	0.70591
	0.47076	0.50861	0.40532	0.55553	0.56070	0.57145	0.60002	0.60960	0.61113	0.65003
30	0.52417	0.56289	0.51382	0.59997	0.60302	0.61141	0.61723	0.62497	0.64998	0.68419
	0.45318	0.48798	0.38431	0.51235	0.54544	0.55178	0.56252	0.57145	0.58821	0.64998

Upper and lower values in each cell are correspond to $\alpha_1=0.05$ and $\alpha_1=0.1$, respectively.

$\alpha_1 = 0.1$, for $k = 1,2,3,\ldots$ such that $k \leq n/2$, and $n = 5(1)10(5)30$ in Tables 1 and 2, respectively.

The powers of Z_k and D_k statistics are compared in Figure 1 for $\alpha_1 = 0.05$ and 0.1, $k = 1,2,3$ and sample size $n = 10,20$ respect to β.

Computing the power of the tests and critical values are achieved using R software.

Figure 1 shows that for $\alpha_1 = 0.05$ and $\alpha_1 = 0.1$, the test based on Z_k is more powerful than the test based on D_k in all values of n, k and β.

An Actual Example

In an insurance company one of services is motor insurance. A claim can be made of at least 500,000 Rials as compensation for the motor insurance. So, the threshold of the distribution of claim is 500,000 Rials. The vehicles involved are of different cost of which some of them may be very expensive. Claim amounts varies according to the damage occurred to the vehicles. It has been observed that claims of these vehicles (expensive/severe damaged vehicle) are several times higher than normal vehicles. In this paper, we have drawn a random sample of size 20 of the claim amounts. It is observed that such claims follow a Pareto distribution in the presence of outliers. Here, the number of outliers (k) is unknown.

The data of claims from the insurance company of Iran records for the year 2008 is given bellow:

750000, 780000, 630000, 1750000, 1450000
3000000, 8650000, 4210000, 890000, 950000
1240000, 1800000, 1630000, 9010000, 4750000
3250000, 1135000, 1326000, 1280000, 760000.

To convert the data to exponentiated Pareto distribution, at first we use the natural logarithm transformation. So, the converted data follow an exponentiated Pareto distribution. Therefore, the values of Z statistic are 0.98467, 0.38261, 0.26020, 0.17834, 0.14397, 0.08466, 0.07595, 0.06568, 0.05405 and

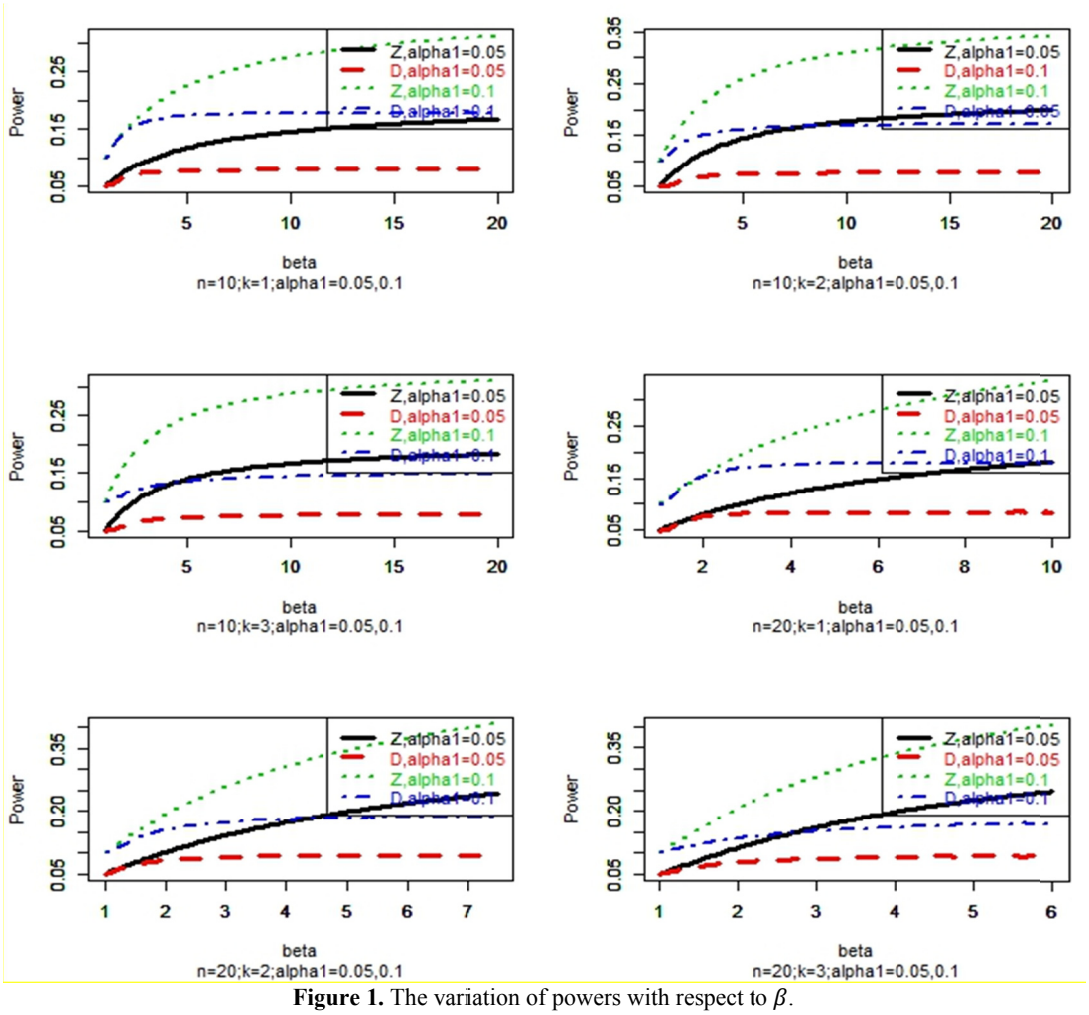

Figure 1. The variation of powers with respect to β.

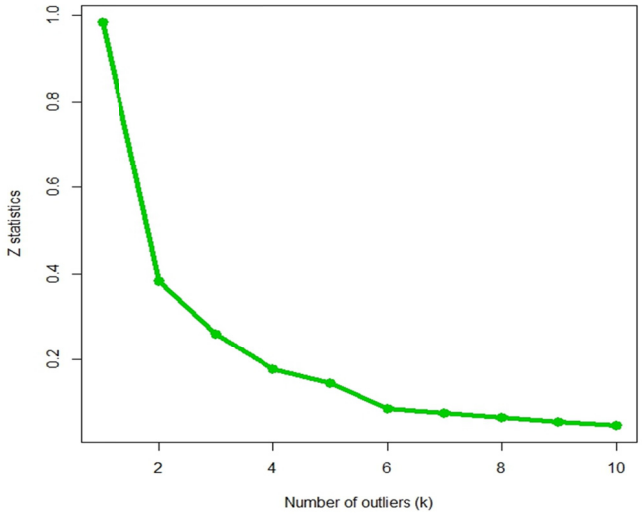

Figure 2. Values of Z statistic respect to k for actual example.

0.04578 for k=1:10, respectively. Comparing these values with critical values in Table 1. corresponding to $n=20$, shows that the number of outlier is $k=1$ and 9010000 is its value for α_1=0.05 or 0.1. Figure 2 shows values of Z statistic respect to k and it is decreased when k is increasing.

Acknowledgement

The author is grateful to the referees and the editor for making valuable suggestions which led to an improved version of the paper.

References

1. Adil I.H. and Irshad A.R. A Modified Approach for Detection of Outliers. *Pak.j.stat.oper.res.* **11** (1): 91-102 (2015).
2. Afify W.M. On estimation of the exponentiated Pareto distribution under different sample schemes. *Stat. Methodol.* **7** (2): 77-83 (2010).
3. Barnett V. and Lewis T. Outliers in Statistical Data. Second edn. *John Wiley and Sons, New York* (1984).
4. Dixit U.J. and Jabbari Nooghabi M. Efficient estimation in the Pareto distribution. *Stat. Methodol.* **7** (6): 687-691 (2010).
5. Gogoi B. and Das M.Kr. Detection of Multiple Upper Outliers in Exponential Sample under Slippage Alternative. *IARJSET* **2** (8): 63-69 (2015).
6. Hadi A.S., Imon A.H.M.R and Werner M. Detection of outliers. *WIREs. Comp. Stat.* **1** (1): 57–70 (2009).
7. Jabbari Nooghabi M., Jabbari Nooghabi H. and Nasiri P. Detecting Outliers in Gamma Distribution. *Comm. Statist. Theory Methods* **39** (4): 698–706 (2010).
8. Kale B.K. Detection of Outliers. *Sankhya B* **38:** 356–363 (1976).
9. Kornacki A. Detection of outlying observations using the Akaike information criterion. *Biometrical Letters* **50** (2): 117-126 (2013).
10. Kumar N. Test for multiple upper outliers in an exponential sample irrespective of origin. *Statistics* **47** (1): 184-190 (2013).
11. Kumar N. and Lalitha S. Testing for upper outliers in Gamma sample. *Comm. Statist. Theory Methods* **41** (5): 820–828 (2012).
12. Lin Chien-tai and Balakrishnan N. Tests for Multiple Outliers in an Exponential Sample. *Comm. Statist. Simulation Comput.* **43** (4): 706-722 (2014).
13. Mahmoud M.A.E., Yhiea N.M. and El-Said S.M. Estimation of parameters for the exponentiated Pareto distribution based on progressively type-II right censored data. *J. Egyptian Math. Soc.* **24** (3): 431-436 (2016).
14. Nadarajah S. Exponentiated Pareto Distributions. *Statistics* **39** (3): 255-260 (2005).
15. Shawky A.I. and Abu-Zinadah H.H. Characterizations of the Exponentiated Pareto Distribution Based on Record Values. *Applied Mathematical Sciences* **2** (26): 1283-1290 (2008).
16. Shawky A.I. and Abu-Zinadah H.H. Exponentiated Pareto Distribution: Different Method of Estimations. *Int. J. Contemp. Math. Sciences* **4** (14): 677-693 (2009).
17. Zerbet A. and Nikulin M. A New Statistic for Detecting Outliers in Exponential Case. *Comm. Statist. Theory Methods* **32** (3): 573–583 (2003).