Australian Journal of Basic and Applied Sciences, 4(11): 5614-5620, 2010
ISSN 1991-8178

A New Statistic for Detecting Outliers in Exponential Case

Ali Shadrokh and Hassan Pazira

Department of Statistics, Payame Noor University of Tehran, Iran

Abstract: Zerbet and Nikulin presented the new statistic Z_k for detecting outliers in exponential distribution. In this article, we propose a statistic (T_k) different from the well known Dixon's statistic D_k and similar to the statistic Z_k to test for multiple outliers. Distribution of the test based on this new statistic under slippage alternatives is obtained. The power of the new test is also calculated, it is compared to the power of the statistic D_k. The results show that the test based on statistic T_k is more powerful than the test based on the statistic D_k.

Key words: Dixon's statistic; Exponential sample; Outlier; Power of the test; Slippage hypothesis; Test of Chauvenet; Upper outlier; Z statistic.

INTRODUCTION

Bol'shev (1969) generalized the Chauvenet's test for rejecting outlier observations (see Bol'shev, 1969; Voinov and Nikulin, 1993, 1996). This method is suitable for detecting *k* outliers in an univariate data set. The Chauvenet's test can be used for exponential case. Also, Ibragimov and Khalna (1978) considered various modication of this test. Several authors considered the problem for testing one outlier in exponential distribution (Chikkagoudar and Kunchur, 1983; Kabe, 1970; Lewis and Fiellerm, 1979; Likes, 1966). Only two types of statistics for testing multiple outliers exist. First is Dixon's while the second is based on the ratio of the some of the observations suspected to be outliers to the sum of all observations of the sample. In fact, most of these authors have considered a general case of gamma model and 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, 1978). Zerbet and Nikulin (2003) proposed a statistic different from the well-known Dixon's statistic D_k to test for multiple outliers. In this article, we propose a statistic (T_k) similar to the statistic Z_k to test for multiple outliers. Distribution of the test based on this statistic under slippage alternatives is obtained and the tables of critical values are given for various *n* (size of the sample) and *k* (the number of outliers). The power of the new test is also calculated, it is compared to the power of the statistic D_k. The results show that the test based on statistic T_k is more powerful than the test based on the statistic D_k.

2 Statistical Inference:

Let X_1, X_2, ..., X_n be arbitrary independent random variables. In this article, we want to test the hypothesis: H_0: X_1, X_2, ..., X_n derive from a exponential distribution,

$$.\Pr\left\{X \le x \middle| H_0\right\} = F\left(x,\theta\right) = 1-\exp(-x/\theta), \theta > 0, \theta \textit{ is unknown}$$

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

$$f_x\left(x,\theta\right) = \frac{1}{\theta}\exp(-\frac{x}{\theta}), \qquad \theta > 0, x > 0$$

but, under the *slippage alternative* H_k , we have:

$$X_{(1)}, X_{(2)}, ..., X_{(n-k)} \sim f_x(x,\theta)$$

Corresponding Author: Hassan Pazira, Department of Statistics, Payame Noor University of Tehran, Iran
E-mail: Pazira.b@Gmail.com

$$X_{(n-k+1)}, X_{(n-k+2)}, \ldots, X_{(n)} \sim f_x(x, \theta/\beta)$$

where $0 < \beta \leq 1$, 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 $0 < \beta < 1$, these k observations are called *upper outliers*). H_0 corresponds to the $\beta = 1$. To test H_0, we propose the following statistic:

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

Following the idea of the Chauvenet's test, we assume that the decision criterion is the hypothesis H_0 is rejected when $T_k > t_c$, with $t_c = t_c(\alpha)$ being the critical value corresponding to the significance level α.

3 The Distribution of the Statistic T_k Under Alternatives:

In this section, we find the distribution of the statistic T_k, according to Zerbet and k Nikulin (2003) method. Then the distribution of this statistic under the slippage alternative hypothesis H_k is obtained by the following theorem.

Theorem 3.1:

The distribution of the statistic T_k under H_k is given by this formula.

$$\Pr\left\{T_k < \middle| H_k\right\} = (k!)^2 \left(\frac{\beta}{2\theta}\right)^k \sum_{j=1}^{K} \frac{(-1)^{k-j} \left(1 - \left[1 - t(k-j+1)/(kt-1)\right]^{k-1}\right)}{(j-1)!(k-j+1)!}, 0 < t < 1/k.$$

Proof:

To proof this theorem, we must give the distribution of the statistic T_k under the alternative hypothesis H_k.

At first, we compute the corresponding alternative distribution of the statistic:

$$U_{(k)} = \frac{X_{(n)} - X_{(n-k)}}{\sum_{j=n-k+1}^{n} (X_{(j)} - X_{(n-k)})} = \frac{V}{W}, k > 1$$

where $V = X_{(n)} - X_{(n-k)}$ *and* $W = \sum_{j=n-k+1}^{n} (X_{(j)} - X_{(n-k)})$.

Let $Y_j = X_{(j)} - X_{(j-1)}$, we obviously obtain that: $\sum_{j=n-k+1}^{n} Y_j = X_{(n)} - X_{(n-k)}$ and

$$\sum_{j=n-k+1}^{n} (n-j+1) \quad Y_j = \sum_{j=n-k+1}^{n} (X_{(j)} - X_{(n-k)}),$$ then,

$$U = \frac{\sum_{j=n-k+1}^{n} \quad Y_j}{\sum_{j=n-k+1}^{n} (n-j+1) \quad Y_j} = \frac{V}{W}.$$

The characteristic function of (V, W) is

$$\varphi_{(v,w)}(t,z) = E(e^{i(vt+wz)})$$
$$= E(e^{i(\sum_{j=n-k+1}^{n} Y_j t + \sum_{j=n-k+1}^{n} (n-j+1) Y_j z)})$$
$$= \int_{\Re^{n-1}} e^{i(\sum_{j=n-k+1}^{n} Y_j t + \sum_{j=n-k+1}^{n} (n-j+1) Y_j z)}$$
$$\times \quad f_{(Y_{n-k+1},\dots,Y_n)}(y_{n-k+1},\dots,y_n) dy_{n-k+1} \dots dy_n .$$

Knowing that $Y_j,\ j = 1,2,\dots,n-k$ follows the exponential distribution of parameter $\theta(k\beta + n - k - j + 1)^{-1}$ as well as $Y_{n-k+j},\ j = 1,2,\dots,k$ but of parameter $(\theta/\beta)(k-j+1)^{-1}$ (see Chikkagoudar and Kunchur (1983)), we can show the characteristic function $\varphi_{(V,W)}$ is

$$\varphi_{(v,w)}(t,z) = \int_0^{+\infty} e^{it\sum_{j=n-k+1}^{n} y_i} \left[\prod_{r=1}^{k} \frac{1}{b_r} e^{-y_{n-k+r}/b_r} \right]$$
$$\times e^{iz\sum_{j=n-k+1}^{n} (n-j+1) y_i} \left[\prod_{r=1}^{k} \frac{1}{b_r} e^{-y_{n-k+r}/b_r} \right] dy_{n-k+1} \dots dy_n$$
$$= \prod_{j=1}^{k} \int_0^{+\infty} \frac{1}{b_j^2} e^{\frac{-2y_{n-k+j}}{b_j} + ity_{n-k+j} + iz(k-j+1) y_{n-k+j}} dy_{n-k+j}$$
$$= \prod_{j=1}^{k} \left[\frac{1}{b_j^2} \int_0^{+\infty} e^{-y_{n-k+j}(\frac{2}{b_j} - it - iz(k-j+1))} dy_{n-k+j} \right]$$

Then we have,

$$\varphi_{(v,w)}(t,z) = \prod_{r=1}^{k} \frac{1}{b_j^2} \left(\frac{2}{b_j} - it - iz(k-j+1) \right)^{-1}$$

with $b_j = (\theta/\beta)(k-j+1)^{-1}$. From the formula of inversion we can find the joint density function of (*V, W*) as follows:

$$f_{(V,W)}(v,w) = \frac{1}{(2\pi)^2} \int_0^{+\infty} \int_0^{+\infty} \varphi_{(v,w)}(t,z) e^{-i(tv+zw)} dt dz$$

$$= \frac{1}{(2\pi)^2} \int_0^{+\infty} \left[\int_0^{+\infty} \left(\prod_{j=1}^{k} \frac{1}{b_j^2} \left(\frac{2}{b_j} - it - iz(k-j+1) \right)^{-1} e^{-itv} dt \right) e^{-izw} dz \right] \quad (1)$$

Before the finding the joint probability density function (pdf) of variables (V, W), we calculate these products:

$$\prod_{j=1}^{k}\frac{1}{b_j^2}=\prod_{j=1}^{k}\left(\frac{\beta}{\theta}\right)^2(k-j+1)^2$$

$$=(k!)^2\left(\frac{\beta}{\theta}\right)^{2k} \qquad (2)$$

and

$$\prod_{j=1}^{k}\left(\frac{2}{b_j}-it-iz(k-j+1)\right)^{-1}=\prod_{j=1}^{k}\frac{1}{\left(\frac{2\beta}{\theta}-iz\right)(k-j+1)-it}$$

$$=(-1)^k\prod_{j=1}^{k}\frac{1}{it-\Delta_j}$$

$$=(-1)^k\sum_{j=1}^{k}\frac{1}{(it-\Delta_j)\prod_{i=1,i\neq j}^{k}(\Delta_j-\Delta_i)}$$

$$=(-1)^k\sum_{j=1}^{k}\frac{1}{(it-\Delta_j)\prod_{i=1,i\neq j}^{k}(i-j)\left(\frac{2\beta}{\theta}-iz\right)}$$

$$=(-1)^k\sum_{j=1}^{k}\frac{1}{(it-\Delta_j)((-1)^{j+1}(j-1)!(k-j)!\left(\frac{2\beta}{\theta}-iz\right)^{k-1})}$$

$$=\sum_{j=1}^{k}\frac{(-1)^{k-j-1}}{(it-\Delta_j)\left(\frac{2\beta}{\theta}-iz\right)^{k-1}(j-1)!(k-j)!} \qquad (3)$$

where $\Delta_j=(2\beta/\theta-iz)(k-j+1)$.

Therefore, with replacement of Eqs. (2) and (3) into Eq. (1), we can calculate the joint pdf of variables (V,W):

$$f_{(V,W)}(v,w)=\frac{(k!)^2}{(2\pi)^2}\left(\frac{\beta}{\theta}\right)^{2k}\sum_{j=1}^{k}\frac{(-1)^{k-j-1}}{(j-1)!(k-j)!}\int_0^{+\infty}\left[\int_0^{+\infty}\left(\frac{e^{-itv}}{it-\Delta_j}dt\right)\frac{e^{-izw}}{\left(\frac{2\beta}{\theta}-iz\right)^{k-1}}dz\right]$$

$$=\frac{(k!)^2}{2\pi}\left(\frac{\beta}{\theta}\right)^{2k}\sum_{j=1}^{k}\frac{(-1)^{k-j}}{(j-1)!(k-j)!}\int_0^{+\infty}\frac{e^{-v\Delta_j-izw}}{\left(\frac{2\beta}{\theta}-iz\right)^{k-1}}dz$$

$$= \frac{(k!)^2}{2\pi}\left(\frac{\beta}{\theta}\right)^{2k} \sum_{j=1}^{k} \frac{(-1)^{k-j} e^{-2v\beta(k-j+1)/\theta}}{(j-1)!(k-j)!} \int_0^{+\infty} \frac{e^{-i(w-v(k-j+1))z}}{\left(\frac{2\beta}{\theta}-iz\right)^{k-1}} dz$$

$$= \quad k(k-1)k!\left(\frac{\beta}{\theta}\right)^{2k} e^{-2w\beta/\theta} \sum_{j=1}^{k} \frac{(-1)^{k-j}\left[w-v(k-j+1)\right]^{k-2}}{(j-1)!(k-j)!} \tag{4}$$

in the process to find joint pdf of variables (V,W), we knowing that

$$\int_0^{+\infty} \frac{e^{-itv}}{it-\Delta_j} dt = -2\pi e^{-v\Delta_j}$$

and also,

$$\int_0^{+\infty} \frac{e^{-i(w-v(k-j+1))z}}{\left(\frac{2\beta}{\theta}-iz\right)^{k-1}} dz = \frac{2\pi(w-v(k-j+1))^{k-2}}{(k-2)!} e^{-2(w-v(k-j+1))\beta/\theta}$$

As a conclusion, the pdf of U_k is

$$f_{U_k}(u) = (k-1)(k!)^2\left(\frac{\beta}{2\theta}\right)^k \sum_{j=1}^{k} \frac{(-1)^{k-j}\left[1-u(k-j+1)\right]^{k-2}}{(j-1)!(k-j)!} \tag{5}$$

then,

$$\Pr\left\{U_k < u\right\} = \int_0^u f_U(x)dx = (k!)^2\left(\frac{\beta}{2\theta}\right)^k \sum_{j=1}^{k} \frac{(-1)^{k-j}\left(1-\left[1-u(k-j+1)\right]^{k-1}\right)}{(j-1)!(k-j+1)!} \tag{6}$$

The distribution function of T_k can be found from (1) by using the relation

$$\Pr\left\{T_k < t \middle| H_k\right\} = \Pr\left\{U_k < \frac{t}{kt-1}\middle| H_k\right\}, \quad 0 < t < 1/k$$

finally, the proof is complete.

Corollary:

Under H_0 the distribution of statistic T_k is obtained from the theorem taking β = 1.

4 Power Comparison of the Tests Based on T_k and D_k

In this section, we give the critical values of the statistics T_k,

$$D_k = \frac{X_{(n)} - X_{(n-k)}}{X_{(n)}}$$

for the levels of significance $\alpha = 0.05$ and $\alpha = 0.1$, for $k = 2, 3, \ldots$ such that $k < n$ and $n = 6(1)11$ in Tables 1 and 2 respectively.

Tables: In the following Tables, Upper value in each cell refers to α= 0.05 and lower value to α = 0.1.

Table 1: Critical values of T_k for α = 0.05 and α = 0.1

	k		
n	2	3	4
6	0.2967	0.1509	0.0893
6	0.2709	0.1362	0.0716
7	0.3035	0.1605	0.0906
7	0.2871	0.1533	0.0857
8	0.3198	0.1795	0.1042
8	0.2908	0.1692	0.0901
9	0.3212	0.1741	0.1188
9	0.3019	0.1721	0.1029
10	0.3390	0.1881	0.1128
10	0.3161	0.1710	0.1038
11	0.3478	0.1925	0.1290
11	0.3128	0.1899	0.1121

Table 2: Critical values of D_k for α = 0.05 and α = 0.1

	k		
n	2	3	4
6	1.6141	1.7085	1.8515
6	1.6152	1.7094	1.8511
7	1.5810	1.6471	1.8372
7	1.5812	1.6140	1.8372
8	1.5167	1.5587	1.7712
8	1.5189	1.5809	1.7868
9	1.5056	1.5198	1.7819
9	1.4931	1.5202	1.7819
10	1.4759	1.5056	1.7566
10	1.4819	1.0949	1.7636
11	1.4571	1.4819	1.7489
11	1.4635	1.4819	1.7446

5 Conclusions:

According to Tables 1, 2 and 3, we can see the critical value of T_k increases when n is increasing. But, the critical value of D_k decreases when n is increasing.

Also, the critical value of T_k decreases when k is increasing. But, the critical value of D_k increases when k is increasing.

These two properties are similar to the Zerbet and Nikulin (2003) approach.

REFERENCES

Barnett, Vic., Toby Lewis, 1978. Outlier in Statistical Data. New York: John Wiley and Sons, Inc.

Bol'shev, L.N., 1969. On tests for rejecting outlying observations. Trudy Inta prikladnoi Mat. Tblissi Gosudart. univ., 2: 159-177 (In Russian).

Bol'shev, L.N. and M. Ubaidullaeva, 1974. Chauvenet± test in theclassical theory of errors. Theory Prob. Applications, 19: 683-692.

Chikkagoudar, M.S. and S.H. Kunchur, 1983. Distribution of test statistics for multiple outliers in exponential samples. Comm. Stat. Theory. and Meth., 12: 2127-2142.

Greenwood and P.E. Nikulin, 1996. A Guide to Chi-Squared Testing. New York: John Wiley and Sons, Inc.

Ibrakimov, I.A. and Khalna, 1978. Some asymptotic results concerning the Chauvenet test. Ter. Veroyatnost. i Primenen, 23(3): 593-597.

Kabe, D.G., 1970. Testing outliers from an exponential population. Metrika, 15: 15-18.

Laurent and G. Andre, 1963. Conditional distribution of order statistics and distribution of the reduced ith order statistic of the exponential model. Ann. Math. Statist., 34: 652-657.

Lewis, T. and N.R.J. Fiellerm, 1979. A recursive algorithm for null distribution for outliers: I. Gamma samples. Technometrics, 21: 371-376.

Likes, J., 1966. Distribution of Dixon's statistics in the case of an exponential population. Metrika, 11: 46-54 (91, 96, 136, 198-200, 204, 209, 210).

Voinov, V.G. and Nikulin, M.N., 1993. Unbaised Estimators and Their Applications. 1. Kluwer Academic Publishers.

Voinov, V.G. and M.N. Nikulin, 1996. Unbaised Estimators and Their Applications, 2. Kluwer Academic Publishers.

Zerbet, A. and M.N. Nikulin, 2003. A new statistic for detecting outliers in exponential case. Commun. Statist. Theor. Meth., 32(3): 573-583.