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 modification of this test. Several authors considered the problem for testing one outlier in exponential distribution (Chikkagoudar and Kunchur, 1983; Kabe, 1970; Lewis and Fieller, 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 an exponential distribution,

$$\Pr \{X \leq x | H_0\} = F(x, \theta) = 1 - \exp(-x/\theta), \theta > 0, \theta \text{ is unknown}$$

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

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

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

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

where \( 0 < \beta < 1 \), \( \beta \) is unknown and \( X_{(n-k+1)}, X_{(n-k+2)}, \ldots, X_{(n)} \) 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 \( \alpha \).

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.

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

**Proof:**

To prove 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

\[
\sum_{j=n-k+1}^{n} (n - j + 1) Y_j = \sum_{j=n-k+1}^{n} (X_{(j)} - X_{(n-k)}), \text{ then,}
\]

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

The characteristic function of \((V, W)\) is
\( \varphi_{(v,w)}(t,z) = E(e^{i(tv+wz)}) \)
\[ = E(e^{i(\sum_{j=n-k+1}^{n} Y_j + \sum_{j=n-k+1}^{n} (n-j+1)Y_j z)}) \]
\[ = \int_{\mathbb{R}^{n-k}} e^{i(\sum_{j=n-k+1}^{n} Y_j + \sum_{j=n-k+1}^{n} (n-j+1)Y_j z)} dy_{n-k+1} \cdots dy_n. \]

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

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

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

\[ f_{(V,W)}(v,w) = \frac{1}{(2\pi)^2} \int_0^{\infty} \int_0^{\infty} \varphi_{(v,w)}(t,z)e^{-i(v+zw)} dt \, dz \]
\[ = \frac{1}{(2\pi)^2} \int_0^{\infty} \left[ \prod_{j=1}^{k} \frac{1}{b_j^2} \left( \frac{2}{b_j} - it - iz(k-j+1) \right)^{-1} e^{-iv} dt \right] e^{-iw} dz \]
\[ = \frac{1}{(2\pi)^2} \int_0^{\infty} \left[ \prod_{j=1}^{k} \frac{1}{b_j^2} \left( \frac{2}{b_j} - it - iz(k-j+1) \right)^{-1} e^{-it} dt \right] e^{-iw} dz \] (1)
Before the finding the joint probability density function (pdf) of variables \((V, W)\), we calculate these products:

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

\[
= (k!)^2 \left( \frac{\beta}{\theta} \right)^{2k}
\]

(2)

and

\[
\prod_{j=1}^{k} \left( 2 - i t - i z (k - j + 1) \right)^{-1} = -\prod_{j=1}^{k} \left( 2 \frac{\beta}{\theta} - i z \right) (k - j + 1 - i t)
\]

\[
= (-1)^{k} \prod_{j=1}^{k} \frac{1}{i t - \Delta_j}
\]

\[
= (-1)^k \sum_{j=1}^{k} \frac{1}{(i t - \Delta_j) \prod_{i=j+1}^{k} (\Delta_j - \Delta_i)}
\]

\[
= (-1)^k \sum_{j=1}^{k} \frac{1}{(i t - \Delta_j) ((-1)^{j+1}(j-1)! (k-j)!) \left( \frac{2 \beta}{\theta} - i z \right)^{k-1}}
\]

\[
= \sum_{j=1}^{k} \frac{(-1)^{k-j-1}}{(i t - \Delta_j) \left( \frac{2 \beta}{\theta} - i z \right)^{k-1} (j-1)! (k-j)!}
\]

(3)

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

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

\[
f_{(V,W)}(v,w) = \frac{(k!)^2}{(2\pi)^2} \left( \frac{\beta}{\theta} \right)^{2k} \sum_{j=1}^{k} \frac{(-1)^{k-j-1}}{(j-1)! (k-j)!} \int_{0}^{\infty} \int_{0}^{\infty} \frac{e^{-ivt} e^{-izw}}{it - \Delta_j} \left( \frac{2 \beta}{\theta} - i z \right)^{k-1} \, dt \, dz
\]

\[
= \frac{(k!)^2}{2\pi} \left( \frac{\beta}{\theta} \right)^{2k} \sum_{j=1}^{k} \frac{(-1)^{k-j}}{(j-1)! (k-j)!} \int_{0}^{\infty} e^{-\Delta_j^{-1} i zw} \left( \frac{2 \beta}{\theta} - i z \right)^{k-1} \, dz
\]
in the process to find joint pdf of variables (V,W), we knowing that
\[
\int_{\Delta_j} e^{i\omega t} \, dt = -2\pi e^{-\Delta_j}
\]
and also,
\[
\int_{\Delta_j} e^{i(w-v(k+1-j))z} \left(\frac{2\beta}{\theta} - iz\right)^{-k} \, dz = \frac{2\pi(w-v(k+1-j))^k}{(k-2)!} \left(\frac{2\beta}{\theta} - i\omega\right)^{-k}
\]
As a conclusion, the pdf of \(U_k\) is
\[
f_{U_k}(u) = (k-1)(k!)^2 \left(\frac{\beta}{2\theta}\right)^k \sum_{j=1}^{k} (-1)^{k-j} \left[1-u(k-j+1)\right]^{k-2} \frac{(j-1)(k-j)!}{(j-1)!(k-j)!} \]
then,
\[
Pr\{U_k < u\} = \int_{0}^{u} f_{U_k}(x) \, dx = (k!)^2 \left(\frac{\beta}{2\theta}\right)^k \sum_{j=1}^{k} (-1)^{k-j} \left[1-u(k-j+1)\right]^{k-1} \frac{(j-1)(k-j)!}{(j-1)!(k-j+1)!}
\]
The distribution function of \(T_k\) can be found from (1) by using the relation
\[
Pr\{T_k < t|H_k\} = Pr\left\{U_k < \frac{t}{kt-1}|H_k\right\}, \quad 0 < t < 1/k
\]
finally, the proof is complete.

**Corollary:**
Under \(H_0\) the distribution of statistic \(T_k\) is obtained from the theorem taking \(\beta = 1\).

4 Power Comparison of the Tests Based on \(T_k\) and \(D_k\)
In this section, we give the critical values of the statistics \(T_k\),
\[
D_k = \frac{X_{(n-k)} - X_{(n-1)}}{X_{(n)}}
\]
for the levels of significance \(\alpha = 0.05\) and \(\alpha = 0.1\), for \(k = 2, 3, \ldots\) such that \(k < n\) and \(n = 6(l+1)\) in Tables 1 and 2 respectively.

**Tables:** In the following Tables, Upper value in each cell refers to \(\alpha = 0.05\) and lower value to \(\alpha = 0.1\).
Table 1: Critical values of $T_k$ for $\alpha = 0.05$ and $\alpha = 0.1$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k$</th>
<th>$T_k$</th>
<th>$T_k$</th>
<th>$T_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>0.2967</td>
<td>0.1509</td>
<td>0.0893</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.3081</td>
<td>0.1302</td>
<td>0.0746</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.2871</td>
<td>0.1533</td>
<td>0.0857</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.3198</td>
<td>0.1795</td>
<td>0.1042</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.3212</td>
<td>0.1741</td>
<td>0.1188</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.3019</td>
<td>0.1721</td>
<td>0.1029</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>0.3472</td>
<td>0.1925</td>
<td>0.1290</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.3128</td>
<td>0.1899</td>
<td>0.1121</td>
</tr>
</tbody>
</table>

Table 2: Critical values of $D_k$ for $\alpha = 0.05$ and $\alpha = 0.1$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k$</th>
<th>$D_k$</th>
<th>$D_k$</th>
<th>$D_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>1.6141</td>
<td>1.7085</td>
<td>1.8515</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1.6152</td>
<td>1.7094</td>
<td>1.8516</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1.5810</td>
<td>1.6471</td>
<td>1.8372</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1.5167</td>
<td>1.5587</td>
<td>1.7712</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1.5189</td>
<td>1.5809</td>
<td>1.7868</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1.4931</td>
<td>1.5202</td>
<td>1.7819</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1.4819</td>
<td>1.4819</td>
<td>1.7489</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1.4635</td>
<td>1.4819</td>
<td>1.7566</td>
</tr>
</tbody>
</table>

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


