A new statistic for detecting outliers in Rayligh distribution

Einolah Deiri

Received: 18 January 2019 / Accepted: 8 September 2020 / Published online: 23 September 2020
© The Author(s) 2020

Abstract

Zerbet and Nikulin (Commun Statist Theor Meth 32(3): 573–583, 2003) presented the new statistic $Z_k$ for detecting outliers in exponential distribution. They also compared this statistic with Dixon’s statistic $D_k$. Jabbari et al. (Commun Statist Theor Meth 39(4): 698–706, 2010) expend this statistic ($Z_k$) for Gamma distribution. In this paper, we generalize statistics $Z_k - Z^*$, for detecting outliers in Rayligh distribution and compare the results with the generalized Dixon’s statistic. Distribution of the test based on the statistic $Z^*_k$ under slippage alternatives is obtained. The criterion value and power of the new test are also calculated and compared with the criterion value of the Dixon’s statistic. The results show that the test based on statistic $Z^*_k$ is more powerful than the test based on the statistic $D_k$.

Keywords Dixon’s statistic · Rayligh sample · Outliers · Power of the test · Slippage hypothesis · Test of chauvenet · Upper outlier · Z statistic

1 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 for univariate data set. The Chauvenet’s test can be used for exponential case. Ibragimov and Khalna (1978) considered various modification of this test. Several authors considered the problem of 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 are 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. 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 which is different to the well-known Dixon’s statistic $D_k$ to detect multiple outliers. In this paper, we generalize the statistics $Z_k - Z^*_k$ for detecting outliers in Rayligh distribution. Distribution of the test based on these statistics under slippage alternatives is obtained and the tables of critical values are given for various sample size $n$ and number of outliers $k$. The power of these tests are also calculated and compared. The results show that the test based on statistic $Z^*_k$ is more powerful than the test based on statistic $D_k$.

2 Statistical inference

Let $X_1, X_2, \ldots, X_n$ are arbitrary independent random variables. In this paper, we want to test the hypothesis: $H_0: X_1, X_2, \ldots, X_n$ derive from a Rayligh distribution as.

\[ \Pr \{X \leq x | H_0\} = F(x;0) = 1 - \exp \left(-\frac{x^2}{\theta}\right) \cdot \theta > 0, \theta \text{ is unknown} \]

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

\[ f(x;\theta) = \frac{2}{\theta} x \exp \left(-\frac{x^2}{\theta}\right) \cdot x > 0 \]

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

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

where $\beta \geq 1$, $\beta$ is unknown and $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics corresponding to the observations $X_1, X_2, \ldots, X_n$. This hypothesis could be considered as 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). The hypothesis $H_0$ corresponds to the $\beta = 1$ To test $H_0$, we propose the following statistics:

$$Z_k^* = \frac{x^2_{(n-k)} - x^2_{(1)}}{\sum_{j=n-k+1}^n(x^2_{(j)} - x^2_{(1)})}$$

For $m = 1$, the above statistics ($Z_k^*$) proposed by Zerbet and Nikulin (2003).

Following the idea of the Chauvenet’s test, we assume that the decision criterion is: the hypothesis $H_0$ is rejected when $Z_k^* > z_e$ with $Z_e = Z_e(\alpha)$ being the critical value corresponding to the significance level $\alpha$.

### 3 The distribution of the statistics $Z_k^*$ under alternatives

In this section, we find the distribution of the statistics $Z_k^*$, according to Zerbet and Nikulin (2003) method. Then the distributions of these statistics under the slippage alternative hypothesis $H_k$ are obtained by the following theorem.

**Theorem 3.1.** The distribution of the statistic $Z_k^*$, under $H_k$ is given by:

$$p_r(Z_k^* < z|H_k) = \frac{(-1)^{k-r} \Gamma(k \beta + n - k) \Gamma(k + 2)}{2 \beta^2 \Gamma(k) \Gamma(k \beta + 1)} \times \left(\frac{z}{1 - k z}\right)^2 \sum_{j=2}^{n-k} (-1)^j \frac{(k \beta + n - k - j + 1)}{(j - 1) \Gamma(n - j + 1)} \frac{z^{(k \beta + n - k - j + 1)}}{(1 - k z) \beta}, 0 < \beta < 1$$

where

$$(a_0, a_1; b; z) = \sum_{j=0}^{\infty} \frac{\Gamma(a_0 + j) \Gamma(a_1 + j) \Gamma(b)}{\Gamma(a_0) \Gamma(a_1) \Gamma(b + j + 1)} z^j f!$$

**Proof.** To prove this theorem, we must obtain the distribution of the statistic $Z_k^*$ under the alternative hypothesis $H_k$.

We first compute the corresponding alternative distribution of the statistic:

$$U(k) = \frac{X_k^2 - X_{(1)}^2}{\sum_{j=n-k+1}^k(X_{(j)}^2 - X_{(1)}^2)} = V(k = \frac{X_k^2 - X_{(1)}^2}{W})$$

where $V = X_k^2 - X_{(1)}^2$ and $W = \sum_{j=n-k+1}^k(X_{(j)}^2 - X_{(1)}^2)$. Let $Y_i = X_{(i)} - X_{(1)}$, we obviously obtain that:

$$\sum_{j=2}^{n-k} Y_j = X_{(n-k)} - X_{(1)}$$

$$\sum_{j=n-k+1}^n (n - j + 1)Y_j = \sum_{j=n-k+1}^n (X_{(j)}^2 - X_{(1)}^2).$$

Thus, we obtain that:

$$U = \frac{\sum_{j=2}^{n-k} Y_j}{\sum_{j=n-k+1}^n (n - j + 1)Y_j} = V$$

The characteristic function of $(v, w)$ is

$$\phi_{(v, w)}(t, z) = E\left(e^{i(v \cdot w)}\right) = E\left(e^{i \sum_{j=2}^{n-k} Y_j + \sum_{j=n-k+1}^n (n - j + 1)Y_j z})\right) = e^{i \sum_{j=2}^{n-k} Y_j} + \sum_{j=n-k+1}^n (n - j + 1)Y_j z$$

$X_f(Y_{(1)}, \ldots, Y_{(n)})dY_2 \ldots dY_n$

Knowing that $Y_{(1)}, \ldots, Y_{(n-k)}$ follow the Rayleigh distribution of parameters $1$ and $\theta(k \beta + n - k - j + 1)$, and $Y_{n-k+1}, \ldots, k$ have the same distribution but with parameters $1$ and $(\frac{\beta}{\theta})(k - j + 1)^{-1}$ (see Chikkagoudar and Kunchur (1983)), then the characteristic function $\phi_{(V, W)}$ is

$$\phi_{(V, W)}(t, z) = \int_0^{+\infty} e^{i \sum_{j=2}^{n-k} Y_j} \prod_{r=1}^{n-k} \frac{1}{a_r} e^{-\frac{\omega}{a_r}} \prod_{r=2}^{n-k} \frac{1}{b_r} e^{-\frac{\omega}{b_r}} dY_2 \ldots dY_n$$

$$\prod_{r=1}^{n-k} \int_0^{+\infty} \frac{1}{a_j} e^{-\frac{\omega}{a_j}} dY_j$$

$$\prod_{j=1}^k \int_0^{+\infty} \frac{1}{b_j} e^{-\frac{\omega}{b_j} Y_{n-k+j}} dy_{n-k+j}$$

$$\prod_{j=2}^{n-k} \int_0^{+\infty} \frac{1}{a_j} e^{-\frac{\omega}{a_j} (\frac{1}{\gamma_j} - w)} dy_j$$

$$\prod_{j=2}^{n-k} \int_0^{+\infty} \frac{1}{a_j} e^{-\frac{\omega}{a_j} (\frac{1}{\gamma_j} - w)} dy_j$$
Therefore we have,

\[
\varphi_{(v,w)}(t, z) = \prod_{j=1}^{k} \left[ \int_0^{+\infty} \frac{1}{b_j} \left( 1 - \frac{1}{a_j} - it \right)^{-1} e^{-iz(k-j+1)} \, dt \right]
\]

with \( a_j = \theta(k\beta + n - j + 1) \) and \( b_j = (\theta \beta)(k\beta + n - j + 1) \). Therefore, the joint density function of \((v, w)\) can be obtained as follows:

\[
f_{(v,w)}(v, w) = \prod_{j=2}^{n-k} \left[ \frac{1}{b_j} \Gamma \left( \frac{k}{\theta} \right) \right] e^{-iz(k-j+1)}
\]

To find the joint probability density function of variables \((v, w)\), we first calculate the following products:

\[
\prod_{j=2}^{n-k} \left( \frac{1}{a_j} - it \right) = \sum_{j=2}^{n-k} \left( -1 \right)^{n+k-j-1} \frac{\theta^{n-k-2}}{(it - \frac{1}{a_j})(j-2)!} (n-j-k)!
\]

\[
\prod_{j=2}^{n-k} \frac{1}{a_j} = \frac{\Gamma(-k\beta - n + k + 1 + it\theta)\theta^{n-k-1}}{(-1)^{n+k-j-1}\Gamma(-k\beta)}
\]

\[
\prod_{j=2}^{n-k} \frac{1}{a_j} = \frac{\Gamma(k\beta + n - k)\theta^{n-k-1}}{\theta^{n-k-1}\Gamma(k\beta + 1)}
\]

\[
\prod_{j=1}^{k} \frac{1}{b_j} = \frac{1}{(\theta \beta)(k\beta + n - k + 1)!}
\]

Therefore, the joint density function 

\[
\int_0^{+\infty} \frac{1}{u} \left( 1 - \frac{1}{u} - it \right)^{-1} e^{-iz(k-j+1)} \, dt = -2\pi e^{-\frac{z^2}{2}}, \quad v > 0, \quad \theta > 0, \quad \beta \geq 1
\]

As a conclusion, the pdf of \( U_k \) is

\[
f_{u_k}(u) = \frac{(-1)^n u^2(2 + 2\beta)(\beta + n - k)}{2^\beta 2!(k\beta + 1)^2}
\]

Then the distribution function of \( Z^*_k \) can be found from (1) using the relation

\[
pr\{U_k < u\} = \frac{(-1)^n u^2(2 + 2\beta)(\beta + n - k)}{2^\beta 2!(k\beta + 1)^2}, \quad 0 < u
\]

\[
pr\{Z^*_k < z|H_k\} = pr\left\{ u_k \leq \frac{z}{1 - k\beta}, H_k \right\}, \quad 0 < z < 1/k
\]

and the proof is complete.

**Corollary:** Under \( H_k \) the distribution of statistic \( Z^*_k \) is obtained from the Theorem 3.1 using \( \beta = 1 \).
Table 1 Critical values of $Z^*_k$ for $\alpha = 0.05$ and $\alpha = 0.1$

| $n$ | $\alpha$ | 1 | 2 | 3 |
|-----|-----------|---|---|---|
| 10  | 0.05      | 1.1084 | 0.7513 | 0.4861 |
|     | 0.10      | 0.9852 | 0.7439 | 0.4827 |
| 12  | 0.05      | 1.6437 | 0.9072 | 0.6392 |
|     | 0.10      | 1.6411 | 0.9031 | 0.6327 |
| 15  | 0.05      | 1.8939 | 1.2476 | 0.9721 |
|     | 0.10      | 1.8892 | 1.2361 | 0.9605 |
| 20  | 0.05      | 2.0431 | 1.7646 | 1.2484 |
|     | 0.10      | 2.0411 | 1.7593 | 1.2405 |
| 25  | 0.05      | 2.5328 | 1.8227 | 1.5758 |
|     | 0.10      | 2.5279 | 1.8124 | 1.5721 |

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

| $n$ | $\alpha$ | 1 | 2 | 3 |
|-----|-----------|---|---|---|
| 10  | 0.05      | 0.6529 | 0.7382 | 0.8601 |
|     | 0.10      | 0.6761 | 0.7643 | 0.8896 |
| 12  | 0.05      | 0.4308 | 0.5149 | 0.6367 |
|     | 0.10      | 0.4417 | 0.5243 | 0.6526 |
| 15  | 0.05      | 0.3130 | 0.4256 | 0.5636 |
|     | 0.10      | 0.3295 | 0.4442 | 0.5826 |
| 20  | 0.05      | 0.2165 | 0.3346 | 0.4603 |
|     | 0.10      | 0.2292 | 0.3540 | 0.4914 |
| 25  | 0.05      | 0.1745 | 0.2089 | 0.3082 |
|     | 0.10      | 0.1962 | 0.2474 | 0.3376 |

4 Power comparison of the tests and conclusions

The critical values of the statistics $Z^*_k$ and $D_k$, for the significance levels of $\alpha = 0.05$ and $\alpha = 0.1$, for $k = 1, 2, \ldots$ such that $k < n$, $n = 8(1)12$ is given in Tables 1 and 2, respectively. Meantime, the Dixon’s statistics $D_k$ is given by

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

for more details about the distribution of the Dixon’s statistic, see Likes (1966) and Chikkagoudar and Kunchur (1983).

According to Tables 1 and 2, we can see the critical value of $Z^*_k$ increases when $n$ is increase. But, the critical value of $D_k$ decreases when $n$ is increase.

Also, the critical value of $Z^*_k$ decreases when $k$ is increase. But, the critical value of $D_k$ increases when $k$ is increase.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Barnett, V. and Lewis, T. (1978) Outlier in statistical data. John Wiley and Sons Inc, New York

Bol’shev LN (1969) On tests for rejecting outlying observations. Trudy Inta prikladnoi Mat. Tblissi Gosudart univ 2:159–177 (In Russian)

Bol’shev LN, Ubadullaeva M (1974) Chauvinist ± test in the classical theory of errors. Theory Prob Appl 19:683–692

Chikkagoudar MS, Kunchur SH (1983) Distribution of test statistics for multiple outliers in exponential samples. Comm Stat Theory and Meth 12:2127–2142

Greenwood, and Nikulin PE (1996) A guide to chi-squared testing, John Wiley and Sons, Inc, New York

Ibrakimov IA, Kholna, (1978) Some asymptotic results concerning the Chauvenet test. Ter Veroyatnost i Primenen 23(3):593–597

Jabbari Nooghabi M, Jabbari Nooghabi H, Nasiri P (2010) Detecting outliers in gamma distribution. Commun Statist Theor Meth 39(4):698–706

Kabe DG (1970) Testing outliers from an exponential population. Metrika 15:15–18

Laurent, and Andre G (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, Fiellem NRJ (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 VG, Nikulin MN (1993) Unbiased estimators and their applications. 1. Kluwer Academic Publishers, Dordrecht

Voinov VG, Nikulin MN (1996) Unbiased estimators and their applications. 2. Kluwer Academic Publishers, Dordrecht

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

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.