A new statistic for detecting outliers in Rayleigh distribution

Einolah Deiri¹

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_k^* , for detecting outliers in Rayleigh 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 · Rayleigh 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 Rayleigh 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 $X_1, X_2, \dots, X_n$ are arbitrary independent random variables. In this paper, we want to test the hypothesis: $H_0 : X_1, X_2, \dots, X_n$ derive from a Rayleigh distribution as.

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

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

$$f_x(x; \theta) = \frac{2}{\theta} x \exp\left(-\frac{x^2}{\theta}\right) \quad \theta > 0 \quad x > 0$$

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

$$X_{(1)}, X_{(2)}, \dots, X_{(n-k)} \text{ derive from } F(x; \theta)$$

✉ Einolah Deiri
e.deiri53@qaemiau.ac.ir

¹ Department of Statistics, Islamic Azad University, Qaemshahr Branch, Tehran, Mazandaran, Iran

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

where $\beta \geq 1$, β is unknown and $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are the order statistics corresponding to the observations $X_1, X_2, \dots, X_n$. 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:

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

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 $Z_k^* > z_c$ with $Z_c = Z_c(\alpha)$ being the critical value corresponding to the significance level α .

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:*

$$pr\{Z_k^* < z | H_k\} = \frac{(-1)^{n-k} \Gamma(k\beta + n - k) \Gamma(k + 2)}{2\beta^2 \Gamma(k) \Gamma(k\beta + 1)}$$

$$\times \left(\frac{z}{1 - kz} \right)^2 \sum_{j=2}^{n-k} \frac{(-1)^j (k\beta + n - k - j + 1)}{\Gamma(j - 1) \Gamma(n - j - k + 1)}$$

$$X_2 F_1(2, k + 1; 3 - \frac{z(k\beta + n - k - j + 1)}{(1 - kz)\beta}, 0 < z$$

$$< \frac{1}{k}$$

where

$$(a_0, a_1; b_1; z) = \sum_{j=0}^{\infty} \frac{\Gamma(a_0 + j) \Gamma(a_1 + j) \Gamma(b_1)}{\Gamma(a_0) \Gamma(a_1) \Gamma(b_1 + j)} \frac{z^j}{j!}$$

Proof. To proof 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:

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

where $V = X_{(n-k)}^2 - X_{(1)}^2$ and $W = \sum_{j=n-k+1}^n (X_{(j)}^2 - X_{(n-k)}^2)$.

let $Y_j = X_{(j)}^m - X_{(j-1)}^m$ we obviously obtain that:

$$\sum_{j=2}^{n-k} Y_j = X_{(n-k)}^2 - X_{(1)}^2 \quad \text{and} \quad \sum_{j=n-k+1}^n (n-j+1) Y_j = \sum_{j=n-k+1}^n (X_{(j)}^2 - X_{(n-k)}^2) \quad \text{Then,}$$

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

The characteristic function of (v, w) is

$$\varphi_{(v,w)}(t, z) = E\left(e^{i(v_t - w_z)}\right)$$

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

$$= e^{it \sum_{j=2}^{n-k} Y_j t} + \sum_{j=n-k+1}^n (n-j+1) Y_j z$$

$$\times f_{(Y_2, \dots, Y_n)}(Y_2, \dots, Y_n) dy_2 \dots dy_n$$

Knowing that $Y_i, j = 1, 2, \dots, n - k$ follows the Rayleigh distribution of parameters 1 and $\theta(k\beta + n - k - j + 1)^{-1}$, and $Y_{n-k+j}, j = 1, 2, \dots, k$ have the same distribution but with parameters 1 and $\left(\frac{\theta}{\beta}\right)(k - j + 1)^{-1}$ (see Chikkagoudar and Kunchur (1983)), then the characteristic function $\phi_{(v,w)}$ is

$$\varphi_{(v,w)}(t, z) = \int_0^{+\infty} e^{it \sum_{j=2}^{n-k} Y_j} \left[\prod_{r=2}^{n-k} \frac{1}{a_r} e^{-\frac{y_r}{a_r}} \right]$$

$$\times e^{iz \sum_{j=n-k+1}^n (n-j+1) Y_j} \left[\prod_{r=1}^k \frac{1}{b_r} e^{-\frac{y_{n-k+r}}{b_r}} \right] dy_2 \dots dy_n$$

$$\prod_{r=2}^{n-k} \int_0^{+\infty} \frac{1}{a_j} e^{-\frac{y_j}{a_j} + it y_j} dy_j$$

$$\times \prod_{j=1}^k \int_0^{+\infty} \frac{1}{b_j} e^{iz(k-j+1)y_{n-k+j} \frac{y_{n-k+j}}{b_j}} dy_{n-k+j}$$

$$= \prod_{j=2}^{n-k} \left[\int_0^{+\infty} \frac{1}{a_j} e^{-y_j \left(\frac{1}{a_j} - it \right)} dy_j \right]$$

$$\times \prod_{j=1}^k \left[\int_0^{+\infty} \frac{1}{b_j} e^{-y_{n-k+j} \left(\frac{1}{b_j} - iz(k-j+1) \right)} dy_{n-k+j} \right]$$

Therefore we have,

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

$$\times \prod_{j=1}^k \frac{1}{b_j} \left(\frac{1}{b_j} - iz(k-j+1) \right)^{-1}$$

with $\alpha_j = \theta(k\beta + n - k - j + 1)^{-1}$ and $b_j = \left(\frac{\theta}{\beta}\right)(k-j+1)^{-1}$. Therefore, the joint density function of (v, w) can be obtained as follows:

$$\begin{aligned} f_{(v,w)}(v,w) &= \frac{1}{(2\pi)^2} \int_0^{+\infty} \int_0^{+\infty} \varphi_{(v,w)}(t,z) e^{-i(tv+zw)} dt dz \\ &= \frac{1}{(2\pi)^2} \int_0^{+\infty} \left[\prod_{j=2}^{n-k} \frac{1}{a_j} \left\{ \frac{1}{a_j} - it \right\}^{-1} e^{-itv} dt \right] \\ &\quad \times \int_0^{+\infty} \left[\prod_{j=1}^k \frac{1}{b_j} \left\{ \frac{1}{b_j} - iz(k-j+1) \right\}^{-1} e^{-izw} dz \right] \end{aligned} \quad (1)$$

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

$$\begin{aligned} \prod_{j=2}^{n-k} \frac{1}{\frac{1}{a_j} - it} &= \sum_{j=2}^{n-k} \frac{(-1)^{n+j-k-1} \theta^{n-k-2}}{\left(it - \frac{1}{a_j} \right) (j-2)! (n-j-k)!} \\ &= \frac{\Gamma(-k\beta - n + k + 1 + it) \theta^{n-k-1}}{(-1)^{n-k+1} \Gamma(-k\beta + it\theta)} \end{aligned} \quad (2)$$

$$\begin{aligned} \prod_{j=2}^{n-k} \frac{1}{a_j} &= \frac{(-1)^{n-k+1} \Gamma(-k\beta)}{\theta^{n-k-1} \Gamma(-k\beta - n + k + 1)} \\ &= \frac{\Gamma(k\beta + n - k)}{\theta^{n-k-1} \Gamma(k\beta + 1)} \end{aligned} \quad (3)$$

$$\begin{aligned} \prod_{j=1}^k \frac{1}{\frac{1}{b_j} - iz(k-j+1)} &= \frac{1}{\left(\frac{\beta}{\theta} - iz\right)^k k!} \\ &= \frac{\Gamma(-k\beta - n + k + 1 + it\theta) \theta^{n-k-1}}{(-1)^{n-k+1} \Gamma(-k\beta + it\theta)} \end{aligned} \quad (4)$$

And

$$\prod_{j=1}^k \frac{1}{b_j} = k! \left(\frac{\beta}{\theta} \right)^k \quad (5)$$

By replacement Eqs. 2–5 in Eq. (1), the joint pdf of variables (v, w) will be as follows:

$$f_{(v,w)}(v,w) = \sum_{j=2}^{n-k} \frac{(-1)^{n+j-k} \Gamma(k\beta + n - k) v e^{\frac{-v}{a_j}} \beta^{k-1} w^{k-1} e^{-w \frac{\beta}{\theta}}}{\Gamma(k\beta + 1) \theta^{k+2} (j-2)! (n-j-k)! (k-1)!} \quad (6)$$

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

$$\int_0^{+\infty} \frac{e^{-itv}}{it - \frac{1}{a_j}} dt = -2\pi e^{-v/a_j}, v, w > 0, \theta > 0, \beta \geq 1$$

As a conclusion, the pdf of U_k is

$$f_{u_k}(u) = \frac{(-1)^{n-k} u \Gamma(k+2) \Gamma(k\beta + n - k)}{\beta \theta \Gamma(k) \Gamma(k\beta + 1)} \quad (7)$$

$$\times \sum_{j=2}^{n-k} \frac{(-1)^j \beta^{k+1} (k\beta + n - k - j + 1)}{[\beta + u(k\beta + n - k - j + 1)]^{k+2} \Gamma(j-1) \Gamma(n-j-k+1)}$$

then,

$$pr\{U_k < u\} = \frac{(-1)^{n-k} u^2 (k+2) \Gamma(k\beta + n - k)}{2\beta^2 \Gamma(k) \Gamma(k\beta + 1)} \quad (8)$$

$$\begin{aligned} &\times \sum_{j=2}^{n-k} \frac{(-1)^j (k\beta + n - k - j + 1)}{\Gamma(j-1) \Gamma(n-j-k+1)} \\ &\times {}_2F_1(2, k+1; 3; -\frac{(k\beta + n - k - j + 1)u}{\beta}), 0 < u \end{aligned}$$

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

$$pr\{Z_k^* < z | H_k\} = pr\left\{u_k < \frac{z}{1 - kz} | H_k\right\}, 0 < z < 1/k$$

and the proof is complete.

Corollary: Under H_0 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	α	k		
		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	α	k		
		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	0.5862
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 $k = 1, 2, \dots$ 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, 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, Ubaidullaeva M (1974) Chauvinist $\pm$ 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, Khalna, (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, Fiellerm 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.