Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/050201

Detection of Upper Outliers in Gamma Sample

Alok Kumar Singh * *and S. Lalitha*

Department of Statistics, University of Allahabad, Allahabad-211002, India

Received: 12 Oct. 2017, Revised: 12 Jan. 2018, Accepted: 24 Jan. 2018
Published online: 1 May 2018

Abstract: **Balasooriya and Gadag (1994)** proposed a location and scale invariant test based on the test statistic Zk for testing the k upper outliers in two-parameter exponential sample. Kumar et al have proposed test statistics for testing multiple upper outlier detection in gamma sample. In literature, various test statistics have been proposed to detect outliers in an exponential sample. **Likes (1966)** also proposed a new test statistics to detect outlier in the exponential case. In this paper, the test statistic proposed by Likes has been used to detect outliers in a two parameter gamma sample and the null distribution of the test statistics has been obtained. A simulation study is carried out to compare the theoretical developments.

Keywords: Gamma sample, Outlier detection, Performance criteria, Masking effect, Swamping effect

1 Introduction

The two parameters of a gamma distribution represent the scale and the shape parameters and because of this, it has flexibility in analysing any positive real data. It has increasing as well as decreasing failure rate depending on the shape parameter, which gives an extra edge over exponential distribution, which has only constant failure rate. Since sum of independent and identically distributed (i.i.d.) gamma random variables has a gamma distribution, it has more practical utility. For example, if a system is dependent on a particular component which requires n-spare parts for maintenance, and if the component and each spare parts has i.i.d. gamma lifetime distributions, then the lifetime distribution of the system also follows a gamma distribution. Another interesting property of the family of gamma distributions is that it has likelihood ratio ordering, with respect to shape parameter, when the scale parameter remains constant. It naturally implies the ordering in hazard rate as well as in distribution. Hence, if some outlying observations are present in a sample from Gamma distribution, the inferences made on the basis of this sample may not be dependable. Hence, a study of detection of outlying observations is needed. For gamma samples, a number of discordancy tests for a single and multiple upper outliers have been proposed by various authors; for example, see Barnett and Lewis (1994), Kale (1976), Kimber (1979, 1983), Chikkagoudar and Kunchur (1983), Likes (1987), Lewis and Fieller (1979), Balasooriya and Gadag (1994) and Zhang (1998) etc. Jabbari Nooghabi et. al. (2010) extended the work of Zerbet and Nikulin (2003) for gamma distribution. In this paper a test statistic is developed for identification of multiple upper outliers. To start with, the null and the alternative hypotheses are formulated. Thus the null hypothesis H_0 says that there is no outlying observation in the sample and the alternative hypothesis H_k states that there are k upper outliers in the sample.

2 The Test Statistic

In case of a sample from an exponential distribution, Likes (1987) proposed a test statistic for testing upper outliers

$$D_k = \frac{X_{(n)} - X_{(n-k)}}{X_{(n)} - X_{(1)}} \tag{1}$$

The test statistic (1) is based on score function for testing H_0 against H_k. The statistic (1) will have small values if upper outliers are present in the data and declare them as discordant if they exceed by a specified value. Since, exponential

[*] Corresponding author e-mail: alok.rjnis@gmail.com

distribution has constant failure rate, so it may not be appropriate to assume that the outliers are from the exponential distribution. However, it would be more appropriate to consider more general lifetime model for detecting upper outliers in the data.

The probability density function (pdf) of a gamma distribution with shape parameter and scale parameter θ is given by

$$f\left(x;\lambda,\theta\right)=\frac{1}{\Gamma\lambda\theta^{\lambda}}x^{\lambda-1}\exp\left(-\frac{x}{\theta}\right),\ x>0,\lambda>0,\ \theta>0.$$

For identification of outliers, a statistical test procedure called a discordancy test to decide whether or not the contaminant observations are to be declared as discordant is performed. Thus, given a sample $X_1,X_2,\ldots,X_n$ and its corresponding order statistics $X_{(1)},X_{(2)},\ldots,X_{(n)}$, it may be desirable to test the null hypothesis H_0 such that all the observations are members of a $G(l,\theta)$ against the alternative H_k that $n-k$ observations are from this model but k values come from a $G(\lambda,b\theta)$, $b\geq 1$. Clearly, H_k is a scale slippage alternative.

Thus for a discordancy test of these k observations, the test statistic D_k proposed by Likes (1987) for testing k upper outliers in exponential sample, is applied for gamma sample and the performance of the test is computed.

3 Null Distribution of the Test Statistics

For developing a test procedure, the null distribution of the test statistic has to be obtained. The null distribution of the test statistic D_k is obtained in a similar manner as that was obtained by Kumar and Lalitha (2012) for Exponential sample. The distribution of D_k under the null hypothesis H_0 is given in the following theorem.

Theorem Under the null hypothesis, the statistic D_k defined in (1) follows a beta distribution of second kind with parameters λk and $\lambda(n-1)$ respectively.

Proof: This theorem can be proved using the characteristic function and the inversion theorem. Let

$$Y_j=X_{(n-j+1)}-X_{(n-j)},\ j=1,2,\ldots n-1.$$

Then the numerator and denominator of (1) may be written as-
$\sum\limits_{j=1}^{k}Y_j=X_{(n)}-X_{(n-k)}$ and $\sum\limits_{j=1}^{n-1}Y_j=X_{(n)}-X_{(1)}$ respectively.

The test statistic can be rewritten as

$$D_k=\frac{\sum_{j=1}^{k}\left(X_{(n-j+1)}-X_{(n-j)}\right)}{\sum_{j=1}^{n-1}\left(X_{(n-j+1)}-X_{(n-j)}\right)}.$$

Or $D_k=\frac{\sum_{j=1}^{k}Y_j}{\sum_{j=1}^{n-1}Y_j}=\frac{V}{W}$, where, $V=\sum\limits_{j=1}^{k}Y_j$ and $W=\sum\limits_{j=1}^{n-1}Y_j$.

The joint characteristic function of and is-

$$\varphi_{v,w}\left(t,z\right)=E\left(e^{i(vt+wz)}\right)$$

$$=E\left(e^{i\sum\limits_{j=1}^{k}y_jt+i\sum\limits_{j=1}^{n-1}y_jz}\right)$$

$$=\int e^{i\sum\limits_{j=1}^{k}y_jt+i\sum\limits_{j=1}^{n-1}y_jz}f\left(y_1,y_2,\ldots,y_{n-1}\right)dy_1dy_2\ldots dy_{n-1}\ .$$

Under the null hypothesis, i.e. when no outlying observations are present, Y_j follows a Gamma distribution.Then V and W being functions of Y_j, the joint characteristic function of V and W is given by

$$\varphi_{v,w}\left(t,z\right)=\prod_{j=1}^{k}\int_{0}^{\infty}\frac{1}{\Gamma\lambda\theta^{\lambda}}y_j^{\lambda-1}e^{-ity_j/\theta}dy_j\prod_{j=1}^{n-1}\int_{0}^{\infty}\frac{1}{\Gamma\lambda\theta^{\lambda}}y_j^{\lambda-1}e^{-izy_j/\theta}dy_j.$$

$= \prod_{j=1}^{k} \frac{1}{\theta^{\lambda}} \left(\frac{it}{\theta}\right)^{-1} \prod_{j=1}^{n-1} \frac{1}{\theta^{\lambda}} \left(\frac{iz}{\theta}\right)^{-1}.$

Or

$\varphi(v,w) = \left(\frac{1}{\theta^{\lambda}}\right)^{k} \prod_{j=1}^{k} \left(\frac{it}{\theta}\right)^{-1} \left(\frac{1}{\theta^{\lambda}}\right)^{n-1} \prod_{j=1}^{k} \left(\frac{iz}{\theta}\right)^{-1}.$

Now using inversion theorem, the joint distribution of V and W can be obtained as follows-

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

$= \frac{1}{(2\pi)^2} \frac{1}{\theta^{\lambda k}} \int_0^\infty \prod_{j=1}^{k} \left(\frac{it}{\theta}\right)^{-1} e^{-ivt} dt \frac{1}{\theta^{\lambda(n-1)}} \int_0^\infty \prod_{j=1}^{n-1} \left(\frac{iz}{\theta}\right)^{-1} e^{-iwz} dz.$

Since $\int_0^\infty \frac{e^{-ivt}}{\left(\frac{it}{\theta}\right)^{\lambda}} dt = \frac{2\pi v^{\lambda-1} e^{-v/\theta}}{\Gamma\lambda}$, $v > 0$ and

$\int_0^\infty \frac{e^{-iwz}}{\left(\frac{iz}{\theta}\right)^{\lambda}} dt = \frac{2\pi w^{\lambda-1} e^{-w/\theta}}{\Gamma\lambda}$, $w > 0$,

$f_{v,w}(v,w) = \frac{1}{\theta^{\lambda k}} \frac{v^{\lambda k-1}}{\Gamma\lambda k} e^{-v/\theta} \frac{1}{\theta^{\lambda(n-1)}} \frac{v^{\lambda(n-1)-1}}{\Gamma\lambda(n-1)} e^{-w/\theta}.$

Hence, the joint distribution of v and w is

$f_{v,w}(v,w) = \frac{1}{\left(\theta^{\lambda}\right)^{n+k-1}} \frac{v^{\lambda k-1}}{\Gamma\lambda(n-1)} \frac{w^{\lambda(n-1)-1}}{\Gamma\lambda k} e^{-\frac{v}{\theta}} e^{-\frac{w}{\theta}}.$

Or

$f_{v,w}(v,w) = \frac{w^{\lambda(n-1)-1}}{\theta^{\lambda(n-1)}} \frac{e^{-\frac{w}{\theta}}}{\Gamma\lambda(n-1)} \frac{v^{\lambda k-1}}{\theta^{\lambda k}} \frac{e^{-\frac{v}{\theta}}}{\Gamma\lambda k}; v,w > 0; \lambda,\theta > 0.$

$= f(v)\, f(w),$

where $f(v) = \frac{v^{\lambda k-1}}{\theta^{\lambda k}} \frac{e^{-\frac{v}{\theta}}}{\Gamma\lambda k}$; $v > 0; \lambda,\theta > 0,$

and $f(w) = \frac{w^{\lambda(n-1)-1}}{\theta^{\lambda(n-1)}} \frac{e^{-\frac{w}{\theta}}}{\Gamma\lambda(n-1)}$; $w > 0; \lambda,\theta > 0.$

From this, the probability density function (pdf) of D_k is obtained as

$$f_{d_k}(D) = \frac{1}{B(\lambda k, \lambda(n-1))} \frac{d_k^{\lambda k-1}}{(1+d_k)^{\lambda(n+k-1)}}\ , \ 0 < d_k < \infty. \tag{2}$$

This is a $\beta_2(\lambda k,\ \lambda(n-1))$ density.

Now, let $\frac{1}{1+d_k} = d_k{}'$. then equation (2) transform to

$$f_{d_k{}'}(D) = \frac{1}{B(\lambda(n-1),\ \lambda k)} \left(d_k{}'\right)^{\lambda(n-1)-1} \left(1-d_k{}'\right)^{\lambda k-1}\ , \ 0 < d_k{}' < 1. \tag{3}$$

where, $B(a,b) = \frac{\Gamma a \Gamma b}{\Gamma(a+b)}$ is a complete beta function.

4 Critical Values

It can be seen that if k largest observations are outlying, then the statistic D_k will assume small value. Hence, the critical values d_α for level of significance α may be obtained by using from the following equation.

$$P[D_k < d_\alpha | H_0] = \alpha \tag{4}$$

Consequently, $\int_0^{d_\alpha} f_{d_k{}'}(D)\, dD = \alpha$ has to be solved for d_α for obtaining critical values.

Thus on using (3),

$d_\alpha = I_{\lambda(n-1),\ \lambda k}(\alpha)$, where, $I_{a,b}(x) = \frac{1}{B(a,b)} \int_0^x y^{a-1}(1-y)^{b-1}$

is an incomplete beta function of first kind. This equation has to be solved for obtaining the values of d_α.

Hence, the test procedure is as follows: Reject H_0, when $D_k < d_\alpha$ otherwise it may be accepted. Here k denotes the number of largest observations that are declared as discordant at α level of significance.

An Example: Kimber and Steven(1981) in which the time intervals data are given by 25, 52, 7, 61, 446, 34, 87, 76,4, 17, 19, 240, 116, 45, 64, 141, 31, 503, 10, 181, 101. For testing k=2 upper outliers, the value of the test statistic for k=1 and 2 were found to be $D_1 = 0.1142285$ and $D_2 = 0.5270541$ respectively at 5 percent level of significance. The critical values for k=1 and 2 are $d_1 = 0.2641682$and $d_2 = 0.6231105$. Hence, in this case both upper extreme observations were declared as outliers.

5 Performance Criteria

To compute the performance of the discordancy test, David (1981) and Barnett and Lewis (1994) have described different performance criteria for single and multiple outliers in a sample. These are refined by Hayes and Kinsella (2003). Under H_k, the following probabilities were defined for different values of k.

$$p_{ij}^k = P(AcceptH_i|H_j); i, j = 1, 2, , k. \tag{5}$$

When $k = 2$, from (5), the probabilities p_{11}^2 and p_{22}^2 of correct decisions and p_{12}^2, p_{21}^2 of masking and swamping effects respectively were computed for the level of significance $\alpha = 0.05$ and for different choices of n & b.

When $k \geq 3$, from (5), similar probabilities can be defined for performance studies. For a good performance, the probabilities in (5) should be high for $i = j$, $i, j = 1, 2, , k$, while it should be low for $i < j$, $i, j = 1, 2, , k$, (case of masking) and $i > j$, $i, j = 1, 2, , k$ (case of swamping).

6 Simulation Study

Here, a simulation study is carried out to compute the performance of the test statistic (1) using the method given by Lin et al (2014). The powers were evaluated and also the probabilities of masking and swamping for the case when $k = 2$ & 3 were determined. For given n, k and b, the samples of size n under the hypothesis H_k, were first generated by choosing a sample of size n from $G(1, 1)$. After that, these samples were arranged in ascending order of magnitude to obtain the ordered samples. For $k = 2, N = 10000$, replications of size $n = 10$, the samples were generated from $G(1, 1)$ distribution and $(n-1)^{th}$ and n^{th} observations were replaced by bx_{n-1} and bx_n, where $b > 1$. The test statistic D_k was computed and compared with the respective critical values. The different performance probabilities given in (5) were obtained for $b = 10(5)50, 50(10), 100(50)150$. Graphs of these probabilities were plotted which are shown in figure 1 to 9 for different value of significance.

It can be seen from 1 and 5 that has low power at initial value of b but as b increases, the power of the test increases very rapidly and become steady for both the cases. The probability p_{22}^2 , shown in figure 2, also increases as b increases. The probability of swamping and masking effects $p_{12}^2 \& p_{21}^2$, respectively, shown in figures 3, and 4, are very low for all values of b.

From fig. 1, it can be seen that the probability p_{11}^2 increases moderately till $b = 20$, but beyond that it increases very rapidly.

From fig. 2, it can be seen that the probability p_{22}^2 increases moderately till $b = 22$, but beyond that it increases very rapidly.

From fig.3, it can be seen that the probability p_{12}^2 is same at all value for b and very low. From fig.4, it can be seen that the probability p_{21}^2 high at initial value of b and moderately drop at $b = 20$ beyond that it decreases very rapidly.

From fig. 6, it can be seen that probability p_{22}^3 increases very rapidly till $b = 90$ after that it become steady. From fig.7, the probability p_{33}^3 is very low till $b = 100$ and after that it increases very rapidly and from $b = 300$, it becomes steady. From fig.8 & 9, masking and swamping effects for $k = 3$ are also very low.

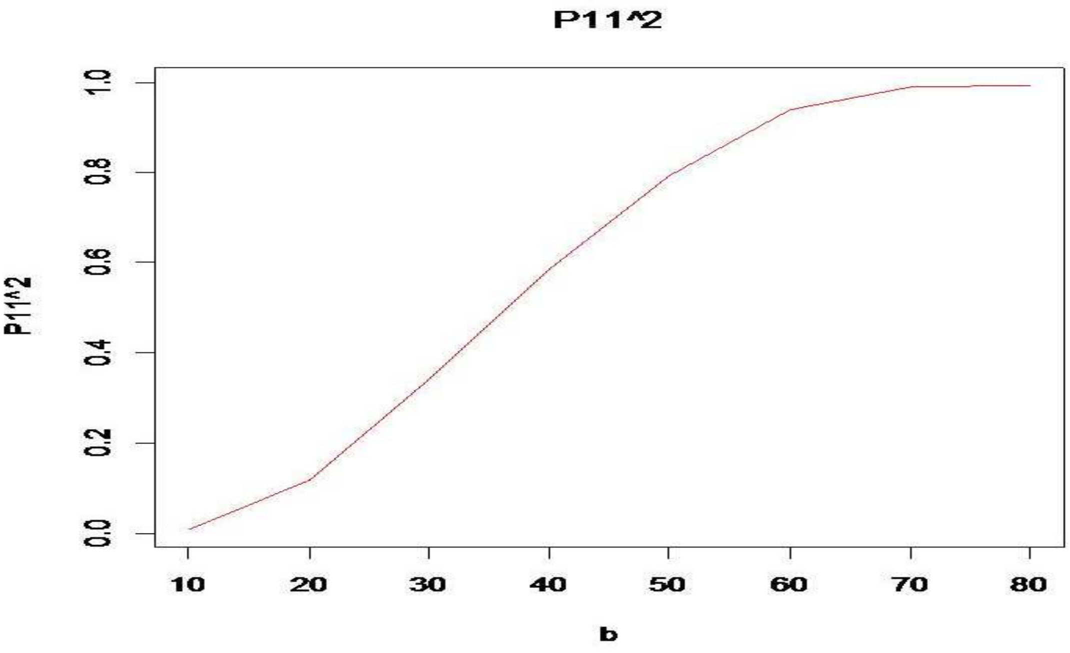

Fig. 1: Over all Power of the statistic D_k for $n = 10$ and $k = 2$ and $\alpha = 0.05$.

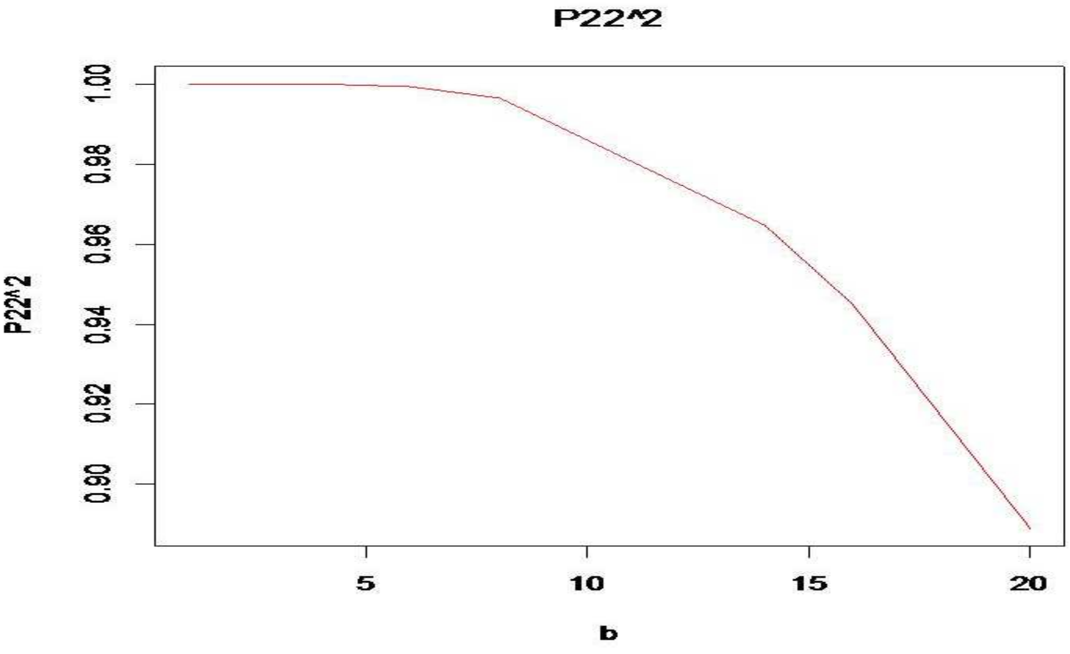

Fig. 2: Performance criterion p_{22}^2 of testing procedure for $n = 10$ and $k = 2$ and $\alpha = 0.05$.

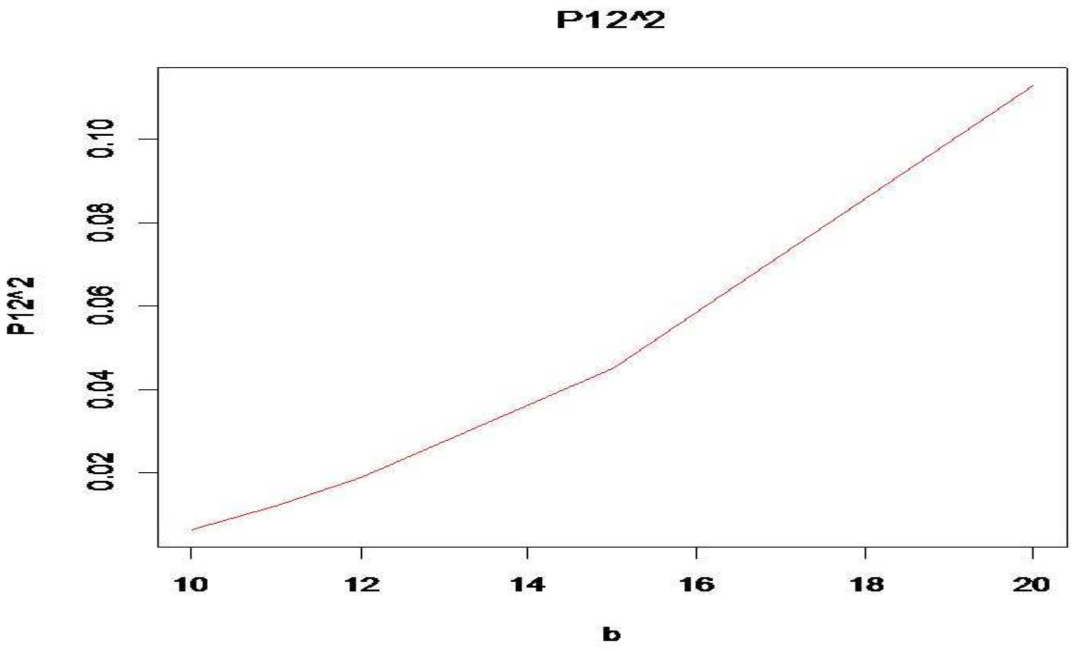

Fig. 3: Masking effect probability of D_k for $n = 10$ and $k = 2$ and $\alpha = 0.05$.

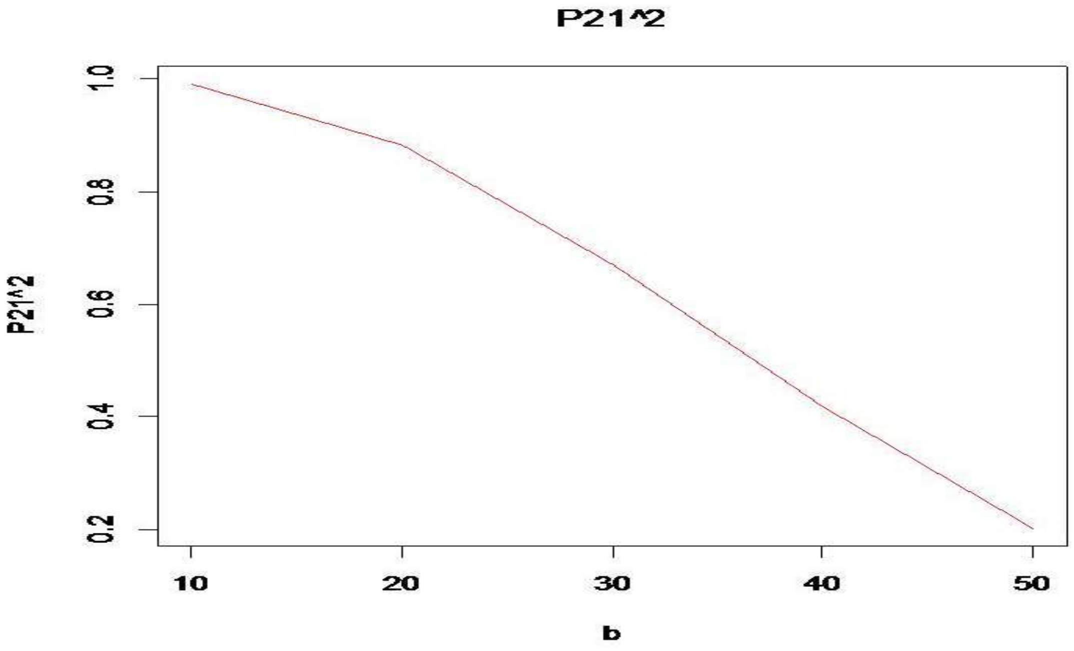

Fig. 4: Swamping effect probability of D_k for $n = 10$ and $k = 2$ and $\alpha = 0.05$.

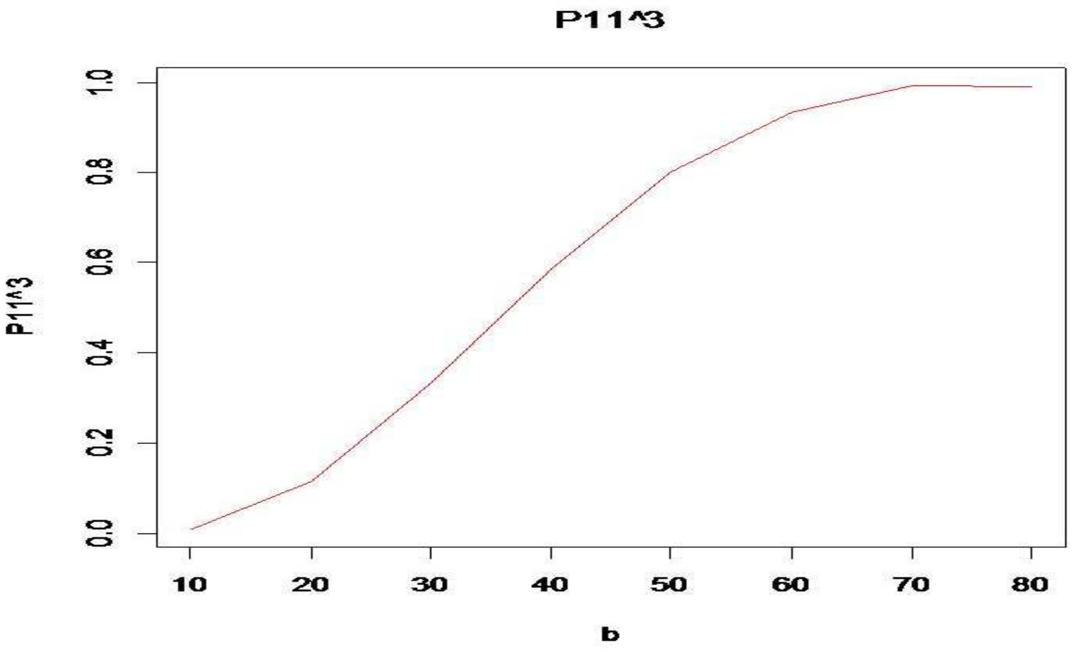

Fig. 5: Power of testing procedure for $n = 15$ and $k = 3$ and $\alpha = 0.05$.

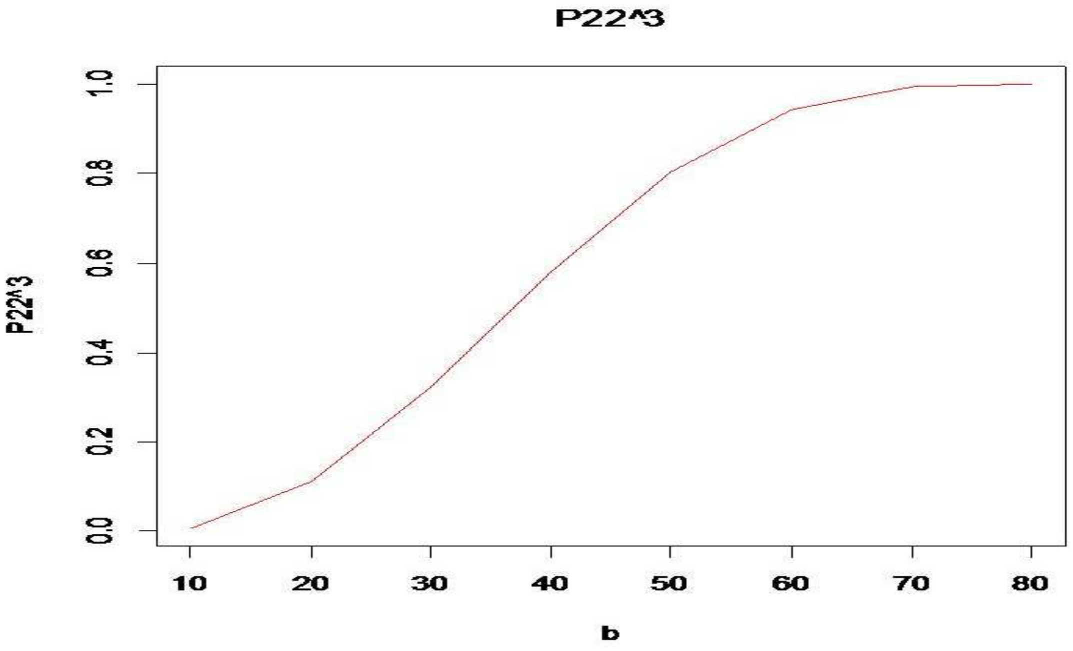

Fig. 6: Performance criterion p_{22}^{3} of testing procedure for $n = 15$ when $k = 3$ and $\alpha = 0.05$.

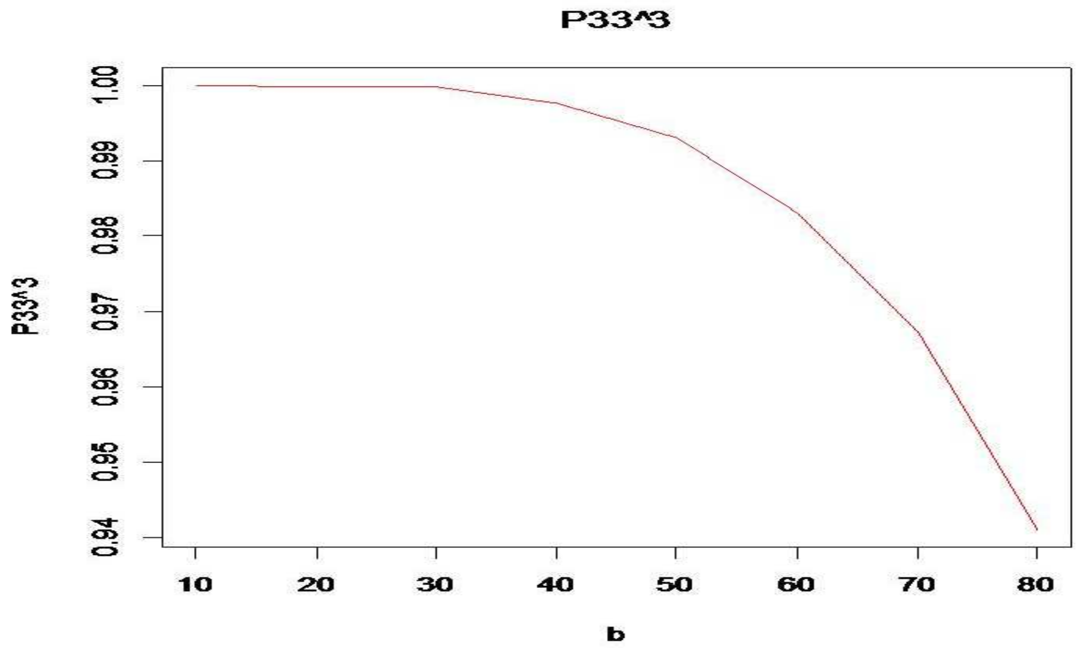

Fig. 7: Performance criterion p_{33}^{3} of testing procedure for $n = 15$ when $k = 3$ and $\alpha = 0.05$.

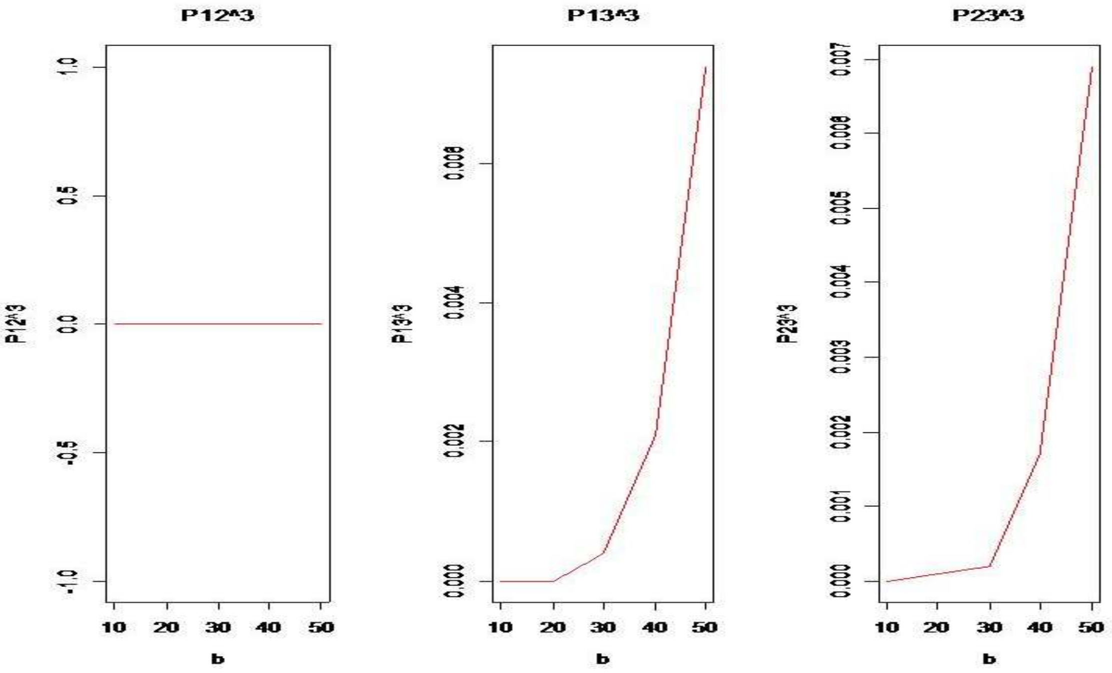

Fig. 8: Masking effect of testing procedure for $n = 15$ when $k = 3$ and $\alpha = 0.05$.

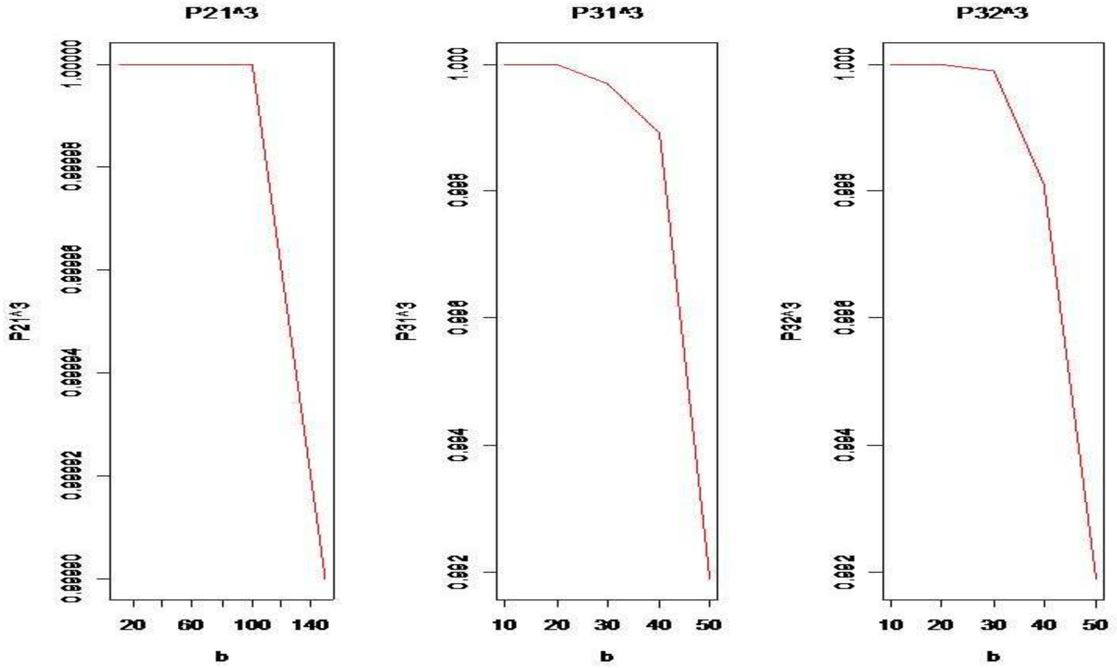

Fig. 9: swamping effect of testing procedure for $n = 15$ when $k = 3$ and $\alpha = 0.05$.

This implies that larger the deviation in scale parameter the lower the effect on the power of testing procedure beyond $b = 90$.

Hence, it shows that all powers and masking, swamping effects have similar pattern for different value of $n\&k$.Similar pattern can be seen for other values of level of significance, i.e for $\alpha = 0.01$ and 0.1.

7 Concluding Remarks

On the basis of performance, in terms of general power and probabilities of swamping and masking effects, it can be concluded that the test based on D_k has a good performance as it correctly identifies the contaminant observations as discordant. Also, D_k has very low probability of masking effect i.e. of not identifying the contaminant observations as outliers. Masking effect of D_k is not very good however, it is considerably low.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Balasooriya, U., Gadag, V. (1994). Tests for upper outliers in the two-parameter exponential distribution.J. Statist. Computat.Simul.50: 249-259.

[2] Barnett, V. A., Lewis, T. (1994).Outliers in Statistical Data. Chichester: John Wiley and Sons.

[3] Chikkagoudar, M. S., Kunchur, S. M. (1983). Distributions of test statistics for multiple outliers in exponential samples.Commun.Statist.Theor.Meth.12: 2127-2142.

[4] David, H. A. (1981). Order Statistics. New York: Wiley.

[5] Hayes, K., Kinsella, T.(2003).Spurious and non-spurious power in performance criteria for tests of discordancy. Statistician 52: 69-82.
[6] Jabbari Nooghabi, M., Jabbari Nooghabi, H., Nasiri, P. (2010).Detecting outliers in gamma distribution.Commun.Statist.Theor.Meth.39: 698-706.
[7] johnson, N. L., Kotz, S., Balakrishnan, N. (1994).Continuous Univariate Distributions. New York: John Wiley & sons.
[8] Kale, B. K. (1976).Detection of outliers.Sankhya B 38: 356-363.
[9] Kimber, A. C. (1979). Tests for a single outlier in a gamma sample with unknown shape and scale parameters. Appl. Statist. 28: 243-250.
[10] Kimber, A. C. (1983). Discordancy testing in gamma samples with both parameters unknown. Appl. Statist. 32: 304-310.
[11] Kumar N, Lalitha S (2012) Testing for upper outliers in gamma sample. Commun Stat Theory Methods 41: 820-828.
[12] Lewis, T., Fieller, N. R. J. (1979). A recursive algorithm for null distribution for outliers:I. gamma samples. Technometrics 21: 371-376.
[13] Likes, J. (1987). Some tests for $k = 2$ upper outliers in an exponential sample. Biometr. J. 29: 313-324.
[14] Lin, C. T., & Wang, S. C. (2015). Discordancy tests for two-parameter exponential samples. Statistical Papers, 56(2), 569-582.
[15] Zerbet, A., Nikulin, M. (2003).A new statistic for detecting outliers in exponential case. J. Statist. Computa.Simul.32: 573-583.
[16] Zhang, J. (1998). Tests for multiple upper or lower outliers in an exponential sample. J. Appl. Statist. 25: 245-255.

Alok Kumar Singh pursuing PhD in Statistics at Univeristy of Allahabad, India. His research interests are in the areas of outlier, spatial outlier, classical and bayesain inference, applied demography, applied mathematical modeling etc.

S. Lalitha is Professor of Statistics at University Allahabad, India. She was recieved her PhD from IIT Kanpur, India. She is referee and editor of several International journals in the frame of pure and applied statistics. She has expertise on outlier, spatial outlier and modeling area of research.