Eur. J. Stat. 1 (2021) 58-73
doi: 10.28924/ada/stat.1.58

Multiple Upper Outlier Detection Procedure in Generalized Exponential Sample

Alok Kumar Singh[1,*], Abhinav Singh[2], Rohit Patawa[3]

[1]Department of Statistics, Raja Balwant Singh Degree College, Agra, India
alok.austats@gmail.com
[2]Department of Statistics, Rajiv Gandhi South Campus, Banaras Hindu University, Mirzapur, India
abhinavsinghstat@gmail.com
[3]Department of Community Medicine, Autonomous State Medical College, Firozabad, India
rohitpatawa@gmail.com
[*]Correspondence: alok.austats@gmail.com

Abstract. Hawkins [6] defined an outlier as an observation that is significantly different from the remaining observations in a dataset so as to arouse suspicion that it was generated by different mechanism. Barnett and Lewis [2] defined an outlier as an observation that deviates significantly in the sample in which it occurs. Spatial outliers are different from outliers and many authors like Singh and Lalitha [9]. Outlier detection procedures for two parameter gamma distribution have been discussed by many authors. But one major disadvantage of the gamma distribution is that the distribution (or survival) function cannot be expressed in a closed form if the shape parameter is not an integer. Since it is in terms of an incomplete gamma function, one needs to obtain the distribution/survival function or the failure rate by numerical integration. This is a limitation in the usage of gamma distribution. It is observed that the generalized exponential distribution can be used as an alternative to the gamma distribution in many situations. Different properties like monotonicity of the hazard functions and tail behaviours of the gamma distribution and that of the generalized exponential distribution are quite similar in nature. But the latter one has a nice compact distribution (or survival) function. It is observed that for a given gamma distribution there exists a generalized exponential distribution so that the two distribution functions are almost identical. Since the gamma distribution function does not have a compact form, efficiently generating gamma random numbers is known to be problematic. It was observed that for all practical purposes it is possible to generate

Received: 18 Sep 2021.

Key words and phrases. generalized exponential sample; outlier detection; performance criteria; masking effect; swamping effect; statistical test; scale invariant.

approximate gamma random numbers using generalized exponential distribution and the random samples thus obtained cannot be differentiated using any statistical tests. Many authors proposed a location and scale invariant test based on the test statistic Z_k for testing the upper outliers in two-parameter exponential sample. Kumar *et. al.* [7] and Singh and Lalitha [10] have proposed test statistics for testing multiple upper outlier detection in gamma sample. Various test statistics have been proposed to detect outliers in an exponential sample. Likes [8] 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 generalized exponential sample and the critical value of the test statistics has been obtained. A simulation study is carried out to compare the theoretical developments.

1. Introduction

Hawkins [6] defined an outlier as an observation that is significantly different from the remaining observations in a dataset so as to arouse suspicion that it was generated by different mechanism. Barnett and Lewis [2] defined an outlier as an observation that deviates significantly in the sample in which it occurs. Spatial outliers are different from outliers and many authors like Singh and Lalitha [9] and Singh [12]. Outlier detection procedures for two parameter gamma distribution have been discussed by many authors. But one major disadvantage of the gamma distribution is that the distribution (or survival) function cannot be expressed in a closed form if the shape parameter is not an integer. Since it is in terms of an incomplete gamma function, one needs to obtain the distribution/survival function or the failure rate by numerical integration. This is a limitation in the usage of gamma distribution.

It is observed that the generalized exponential distribution can be used as an alternative to the gamma distribution in many situations. Different properties like monotonicity of the hazard functions and tail behaviours of the gamma distribution and that of the generalized exponential distribution are quite similar in nature. But the latter one has a nice compact distribution (or survival) function. It is observed that for a given gamma distribution there exists a generalized exponential distribution so that the two distribution functions are almost identical. Since the gamma distribution function does not have a compact form, efficiently generating gamma random numbers is known to be problematic. It was observed that for all practical purposes it is possible to generate approximate gamma random numbers using generalized exponential distribution and the random samples thus obtained cannot be differentiated using any statistical tests. Moreover, if there is a skewed data set where gamma distribution is the best fitting distribution, then in such situations, the generalized exponential distribution also can be used. Gupta and Kundu [5] fitted distributions to two real data sets and observed that the fitted distribution functions were almost identical in many respects in both the cases. It was observed by Gupta and Kundu [4] that the two-parameter generalized exponential distribution with parameters ξ and σ which was denoted by GE(ξ,σ) can be used quite effectively in analysing many lifetime data, particularly in place of two-parameter gamma distribution.

The two parameters of an exponentiated exponential distribution (Generalized exponential) represent the shape and the scale parameter like a gamma distribution. It also has the increasing or decreasing failure rate depending on the shape parameter. The density function varies significantly depending on the shape parameter. It was observed that it has lots of properties which are quite similar to those of a gamma distribution but it has an explicit expression of the distribution function or the survival function. It has also likelihood ratio ordering with respect to the shape parameter, when the scale parameter was kept constant. It was also observed that for a fixed scale and shape parameters there is a stochastic ordering between both the distributions.

2. Materials and Methods

2.1 The Test Statistic

Let $X_1, X_2 \dots, X_n$ be the n observations of a sample from a Generalized exponential distribution with parameters ξ and σ and let $X_{(1)}, X_{(2)} \dots, X_{(n)}$ be the corresponding order statistics. Then to test the null hypothesis H_0, the following test statistic D_k is used, where D_k is defined by

$$D_k = \frac{X_{(n)}-X_{(n-k)}}{X_{(n)}-X_{(1)}}, \quad 0< D_k< 1. \tag{1}$$

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 declares them as discordant if they exceed by a specified value.

In this article, the test statistic in equation (1) for detection k upper outliers in a generalized exponential sample whose *pdf* is given by

$$g(x;\ \xi,\sigma) = \xi\sigma(1-e^{-\sigma x})^{\xi-1}e^{-\sigma x}, \qquad x,\xi,\sigma > 0$$

and *cdf* by

$$G(x) = (1-e^{-\sigma x})^{\xi}, \qquad \xi,\sigma,x > 0.$$

Here ξ is the shape parameter and σ is the scale parameter.

2.2 Test Statistic and its Null distribution

Let $X_1, X_2 \dots, X_n$ be the given sample of size *n* from a generalized exponential distribution with parameters ξ and σ and $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be its corresponding order statistics.

For a discordancy test of k observations, the test statistic D_k, where

$$D_k = \frac{X_{(n)}-X_{(n-k)}}{X_{(n)}-X_{(1)}}, \ 0<D_k< 1.$$

is used. As pointed earlier, a small value of $D_k$would indicate the presence of outliers in the sample.

The null hypothesis H_0 to be tested is that all the observations are from a $GE(\xi,\sigma)$ with unknown σ against the alternative H_k that $n-k$ observations are from this model but the largest k observations are from a $GE(\xi,b\sigma)$, where $b \ge 1$. Clearly, $H_k$is a scale slippage alternative.

The test statistic D_k is based on score function for testing H_0 against H_k. The statistic would have small value if upper outliers are present in the data. Hence, the test procedure declares k observations as discordant if the calculated value of $D_k$turns out to be smaller than the critical value.

2.3 Critical Values

To obtain the critical values of D_k, its distribution should be known. The critical values were obtained using simulation technique. For this a sample of size n with certain values of the parameters $\xi$and σ was generated using R software and the statistic D_k was calculated for certain number k of suspected outliers. This was then repeated 10,000 times and the percentile points were obtained. Thus, the critical values d_α for level of significance α are the 100 α percentile points.The critical values are tabulated in table 1for $n = 10(10)20(10)100$ and $k = 1(1)4$.

Table 1. Critical values of D_k for k=1,2,3,4 and for 5% and 1% significance levels respectively.

n	α	k=1	k=2	k=3	k=4
10	1%	0.006042	0.067464	0.185511	0.293034
	5%	0.029176	0.151137	0.300307	0.429253
	10%	0.056591	0.214216	0.372921	0.500508
11	1%	0.004805	0.062619	0.163167	0.278066
	5%	0.026322	0.14387	0.282663	0.397856
	10%	0.052705	0.20415	0.350842	0.472195
12	1%	0.004497	0.059241	0.161822	0.262801
	5%	0.024257	0.137	0.272635	0.38014
	10%	0.049472	0.192076	0.337611	0.450177
13	1%	0.005223	0.056194	0.150323	0.255565
	5%	0.02472	0.134153	0.25282	0.358246
	10%	0.049321	0.190579	0.318158	0.427167
14	1%	0.004581	0.058862	0.14096	0.241623
	5%	0.022506	0.132777	0.248728	0.352901
	10%	0.044536	0.182692	0.310938	0.41849
15	1%	0.005203	0.053461	0.135822	0.230424
	5%	0.021298	0.122147	0.233576	0.342719
	10%	0.043453	0.173112	0.292832	0.405803
16	1%	0.00466	0.055098	0.137079	0.218975
	5%	0.021243	0.118557	0.233099	0.327414
	10%	0.042047	0.167129	0.292668	0.386553

17	1%	0.004294	0.046762	0.125175	0.203538
	5%	0.020812	0.110805	0.222959	0.310444
	10%	0.040943	0.162648	0.279762	0.376369
18	1%	0.004411	0.048616	0.118354	0.202017
	5%	0.020936	0.111603	0.212512	0.308207
	10%	0.041196	0.162127	0.272565	0.369029
19	1%	0.004401	0.047399	0.119898	0.200467
	5%	0.020382	0.105712	0.205328	0.30329
	10%	0.041403	0.155321	0.263921	0.3623
20	1%	0.003441	0.048893	0.120339	0.194023
	5%	0.021173	0.109005	0.209382	0.298411
	10%	0.041354	0.152012	0.266518	0.359262
30	1%	0.003512	0.042054	0.097164	0.164616
	5%	0.01788	0.092843	0.172887	0.242659
	10%	0.036171	0.131588	0.22379	0.299844
40	1%	0.003026	0.038536	0.087897	0.144893
	5%	0.015359	0.083886	0.156341	0.223244
	10%	0.031671	0.120529	0.199067	0.271653
50	1%	0.002944	0.033747	0.076896	0.133318
	5%	0.015072	0.077107	0.143757	0.209698
	10%	0.028739	0.113386	0.187139	0.255565
60	1%	0.002608	0.032246	0.078724	0.125443
	5%	0.013639	0.076726	0.134553	0.196133
	10%	0.028342	0.110145	0.173315	0.241797
70	1%	0.002812	0.032768	0.072084	0.118248
	5%	0.014676	0.072091	0.129618	0.187277
	10%	0.028169	0.10446	0.16893	0.229832
80	1%	0.002347	0.0294	0.073016	0.112608
	5%	0.013099	0.069653	0.126299	0.178585
	10%	0.026422	0.100813	0.164354	0.221678
90	1%	0.002221	0.026848	0.068724	0.109606
	5%	0.012303	0.064777	0.125187	0.175619
	10%	0.02592	0.094795	0.160938	0.216777
100	1%	0.002366	0.024634	0.068387	0.109742
	5%	0.01166	0.064519	0.121957	0.166099
	10%	0.023832	0.093919	0.155518	0.205803

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

2.4 Performance study

For performance study, the outlying observations were planted in the original sample by generating another sample from a $GE(\xi, b\sigma)$ distribution. To compute the performance criteria for single and multiple outliers in a sample, the following probabilities were defined for different value of k-

$p_{ij}^k = P\left(AcceptH_i \middle| H_j\right), i, j = 1,2, \dots, k.$

Then 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 $n\ and\ b$ were obtained.

It can be seen that the probabilities p_{11}^2, p_{22}^2, p_{12}^2 and $p_{21}^2$are equivalent to

$p_{11}^2 = P(D_1 < d_1, D_2 \geq d_2 | H_1)$

$p_{22}^2 = P(D_2 < d_2 | H_2)$

$p_{12}^2 = P(D_1 < d_1, D_2 \geq d_2 | H_2)$

$p_{21}^2 = P(D_2 < d_2 | H_1)$,

where d_i, are the critical values obtained in table 1 for k=i, i =1,2,3.

3. Results

A simulation study was carried out to compute the performance of the test statistic D_k using the method given by Lin et al [10]. The powers were evaluated and also the probabilities of masking and swamping for the case when $k = 2\ and\ 3$were determined. For given n, k and, b the samples of size n under the hypothesis H_k, are first generated by choosing a sample of size $n - k$ from $GE(1,1)$ and a sample of size k from $GE(1, b), b \geq 1.$ After that, these samples were arranged in ascending order to obtain the ordered samples. For $k = 2, N = 10000$, replications of size $n = 10$, the samples were generated from GE(1,1) distribution and (n-1)th and nth observations were generated from GE(1, b) distribution, where $b \geq 1$. The test D_k was applied and results were noted. The probabilities of the outlying observations were contaminants were obtained by dividing the number of incidents where the statistic fell in the critical region by the total number of repetition, *i.e.* 10,000. The different type of powers and swamping and masking effect were obtained for $b = 0(1)10(10)50$. Graphs of these probabilities were plotted which are shown in figure 1 to 12 and performance values depicted in the table 2.

Table.2. Performance values of test procedure

	Performance probabilities					
	p_{11}^2		p_{11}^3		p_{22}^3	
b	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05	α=0.01
10	0.0074	0.004	0.0094	0.0036	0.0083	0.0036
20	0.1073	0.0654	0.1148	0.0649	0.1075	0.063
30	0.333	0.2205	0.3302	0.2263	0.3285	0.2181
40	0.5759	0.4269	0.5887	0.4311	0.5819	0.4279
50	0.8015	0.6463	0.8	0.6406	0.7974	0.6423

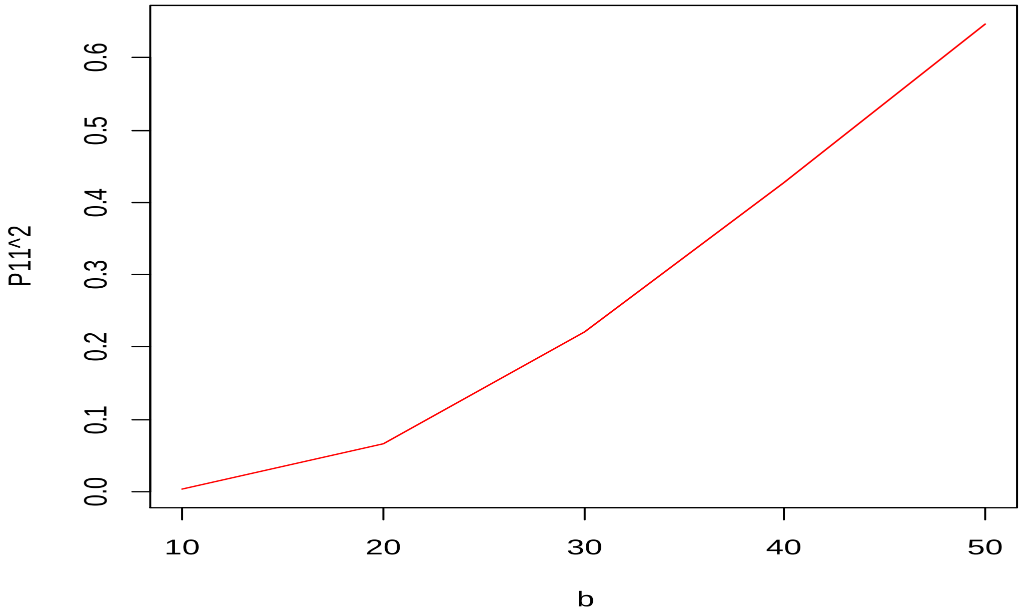

Figure.1 Power probability p_{11}^2 of for n=10 when k=2, α=0.05 and for different b.

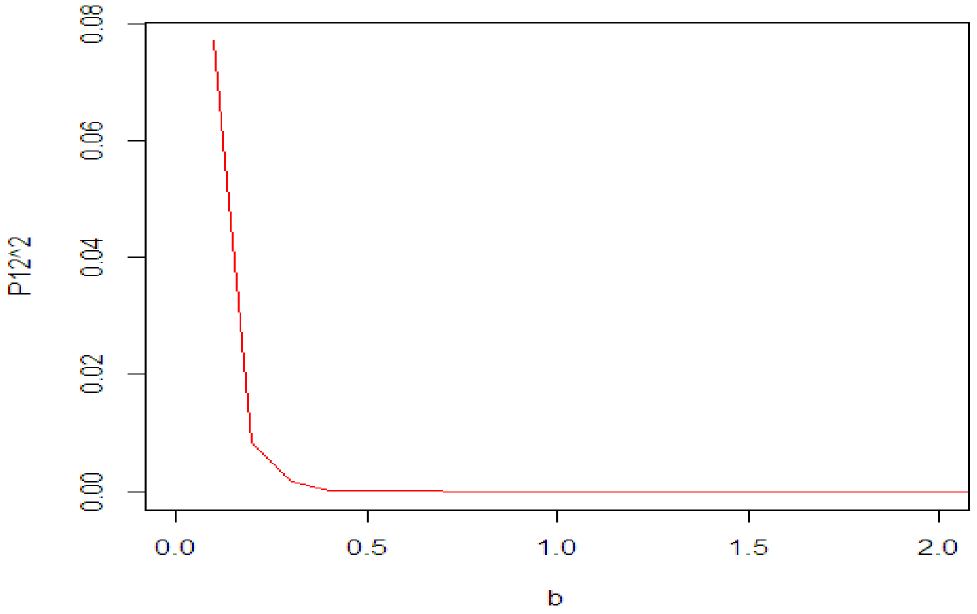

Figure.2 Masking effect of test procedure for *n*=10 when *k*=2 and α=0.05.

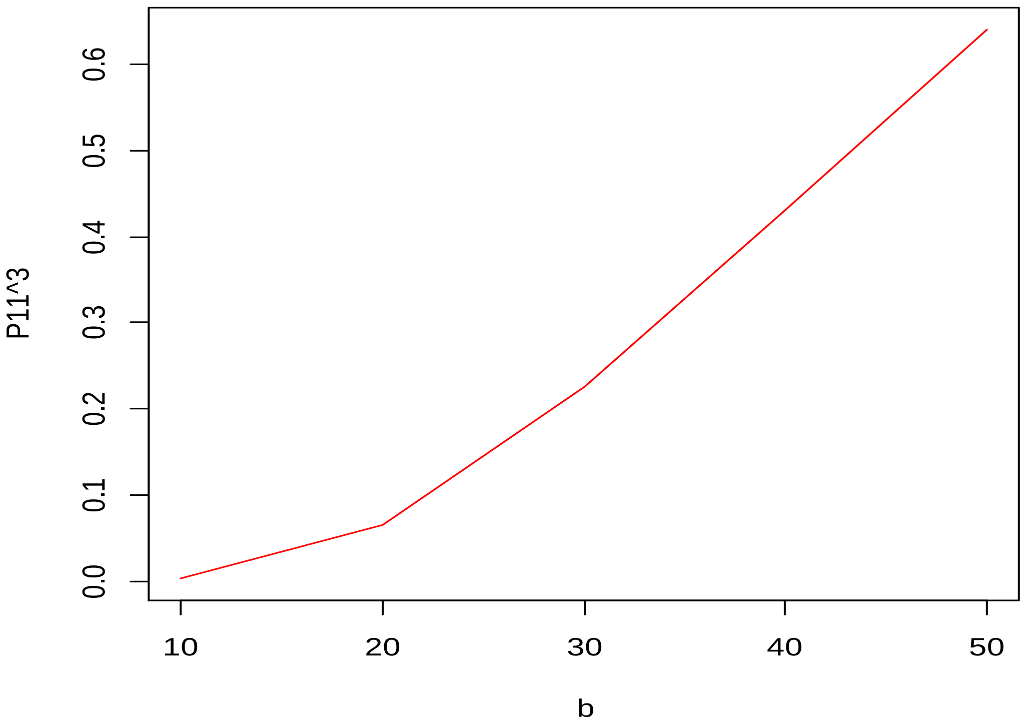

Figure.3 Power of test procedure for *n*=15 when *k*=3 and α=0.05.

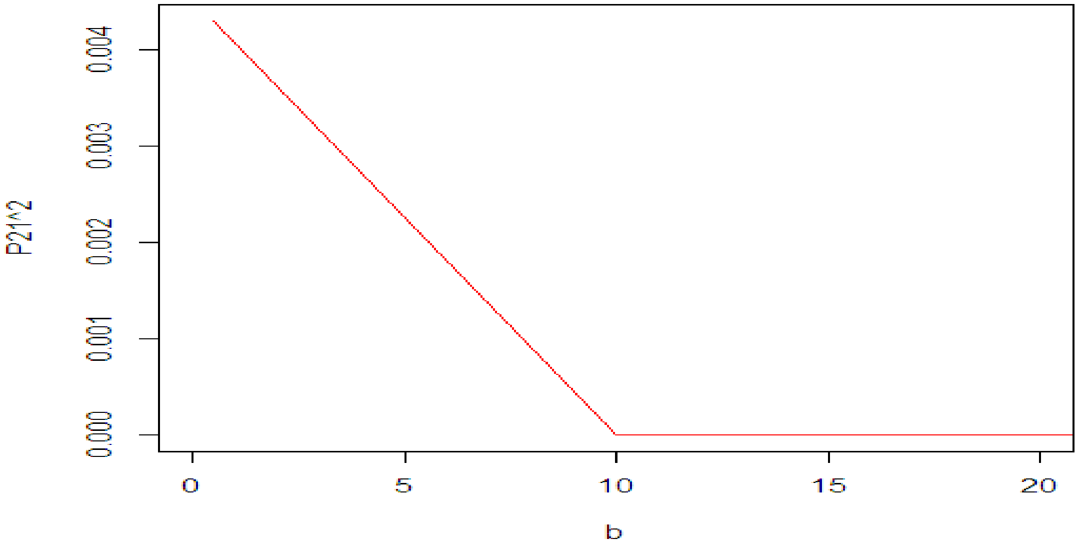

Figure.4. Swamping effect of test procedure for *n*=10 when *k*=2 and α=0.05.

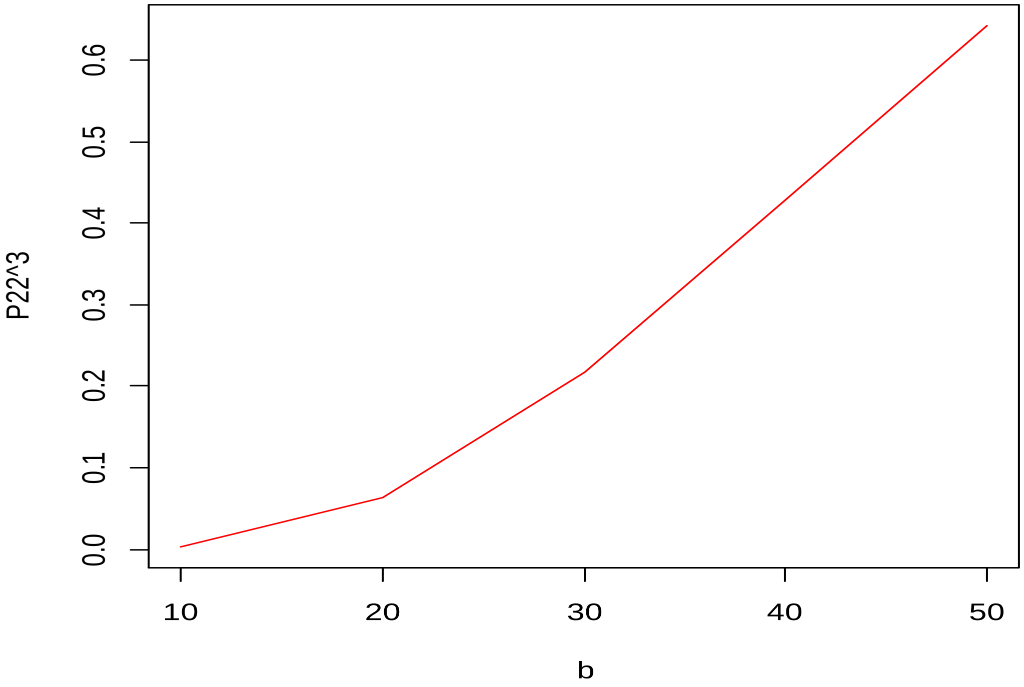

Figure.5 Performance criterion p_{22}^{3} of test procedure for *n*=15 when *k*=3 and α=0.05.

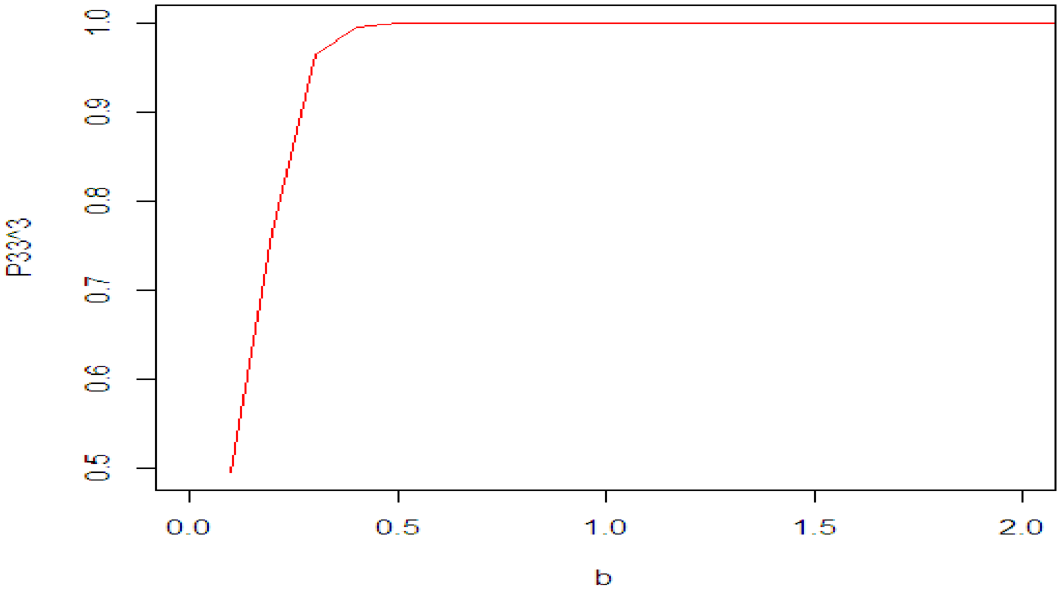

Figure.6 Performance criterion p_{33}^{3} of test procedure for n=15 when k=3 and α=0.05.

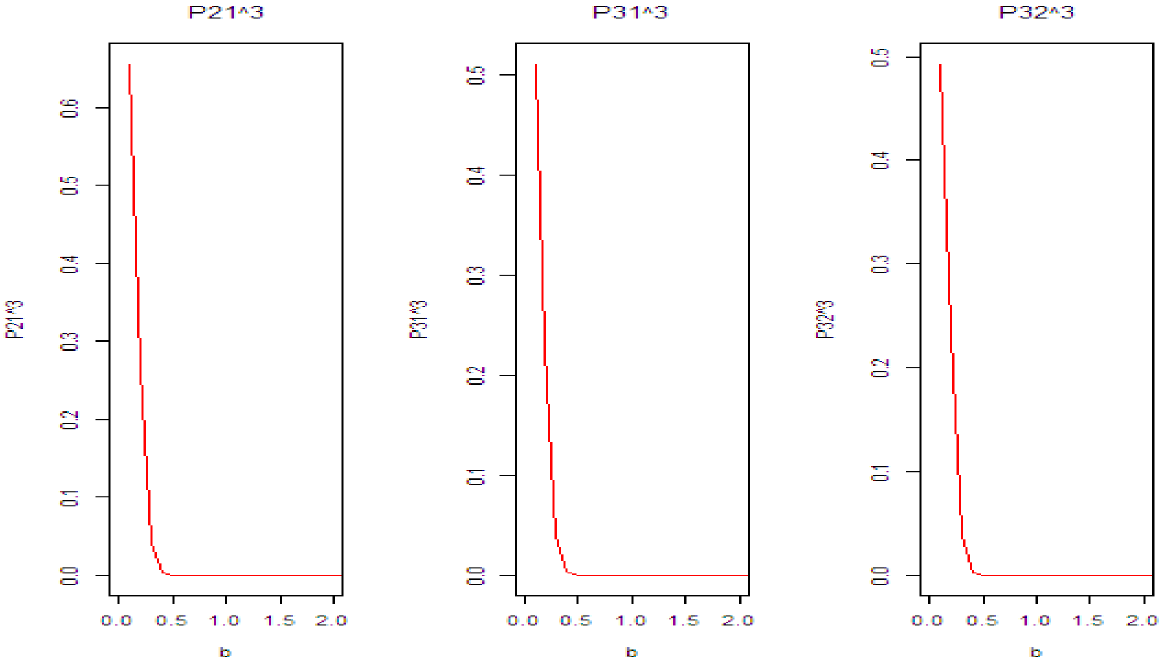

Figure.7 Swamping effect of test procedure for n=15 when k=3 and α=0.05.

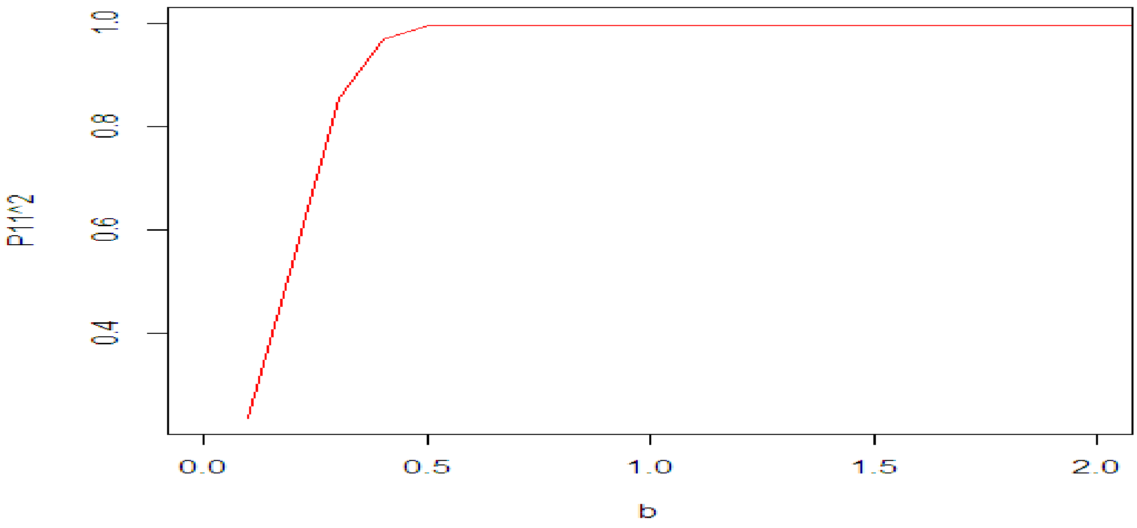

Figure.8 Power of test procedure for *n*=10 when *k*=2 and α=0.01.

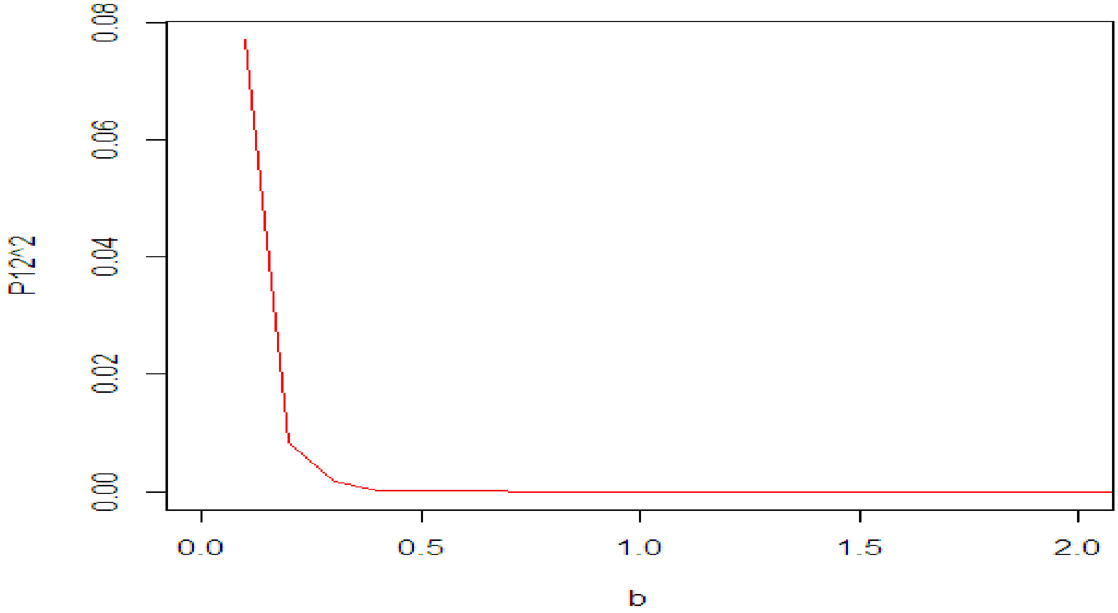

Figure.9 Masking effect of test procedure for *n*=10 when *k*=2 and α=0.01.

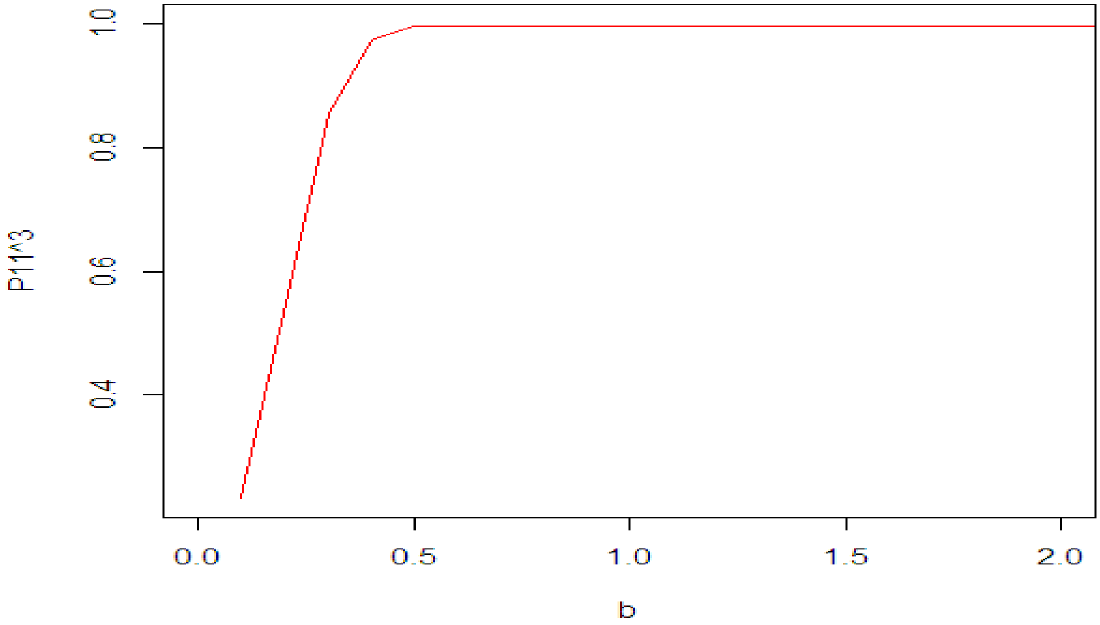

Figure.10 Power of test procedure for *n*=15 when *k*=3 and α=0.01.

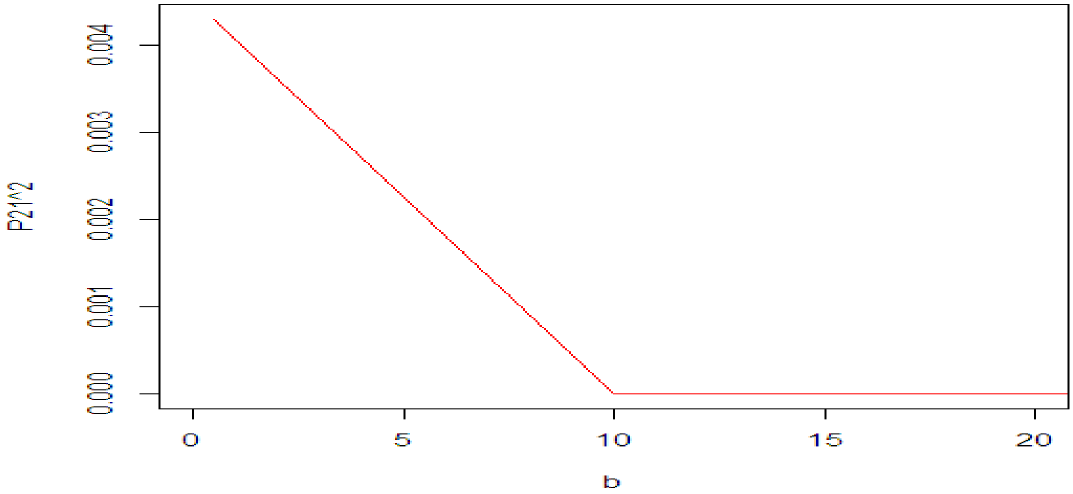

Figure.11 Swamping effect of test procedure for *n*=10 when *k*=2 and α=0.01.

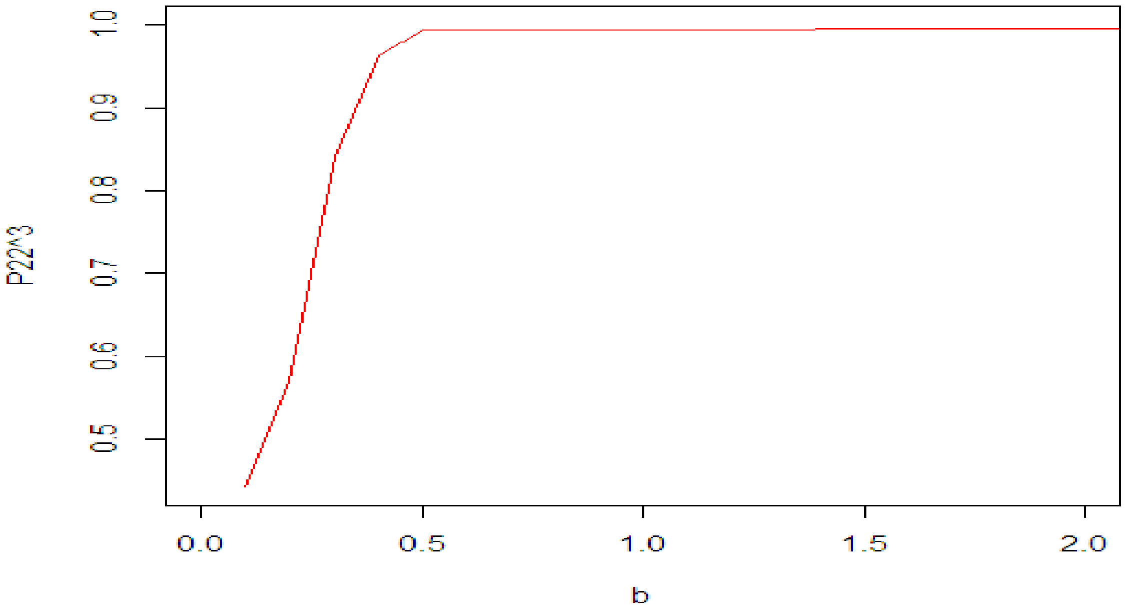

Figure.12 Performance criterion p_{22}^{3} of test procedure for n=15 when k=3 and α=0.01.

4. Discussion

From fig.1 and fig 12 it can be seen that D_k 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}^{3} , shown in figure 5, also increases as b increases. The probability of swamping and masking effects p_{12}^{2} *and* p_{21}^{2} , respectively, shown in figure 2 and 4, are very low for any value of b.

From fig. 1, it can be seen that the probability p_{11}^{2} increases very rapidly and become constant.

From fig.2, 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=5 beyond that it decreases very rapidly.

From these graphs, it can be seen that all powers and masking, swamping effects have similar pattern for different values of $n\ and\ k$. Similar pattern can be seen for other value of level of significance, *i.e.* for α = 0.01.

From fig. 6, it can be seen that probability p_{33}^{3} increases very rapidly till b=0.45and beyond this point, there is no variation. From fig. 7, it can be seen that the swamping effects for $k = 3$ are also very low.

From fig.9, it can be seen that the masking effect increases for a value of b=0.26 and beyond which it drastically drops. This implies that larger the deviation in scale parameter the lower the effect on the power of testing procedure as b increases.

5. Conclusions and suggestions

The performance of the test statistic D_k is reasonably good as it correctly identifies the contaminant observations as discordant. All the powers, masking and swamping effects have similar pattern for different value of $n\ and\ k$. Also, D_k has very low probability of masking effect *i.e.* of wrongly not identifying the contaminant observations as outliers. As the shape parameter increases, the swamping effect also becomes smaller. So test statistic D_k can be used for generalized exponential sample.

References

[1] U. Balasooriya, V. Gadag, Tests for upper outliers in the two-parameter exponential distribution, Journal of Statistical Computation and Simulation. 50 (1994) 249–259. https://doi.org/10.1080/00949659408811614.

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

[3] M.S. Chikkagoudar, S.H. Kunchur, Distributions of test statistics for multiple outliers in exponential samples, Commun. Stati. – Theory Meth. 12 (1983) 2127–2142. https://doi.org/10.1080/03610928308828596.

[4] R.D. Gupta, D. Kundu, Generalized exponential distributions, Aust. NZ J. Stat. 41 (1999) 173–188. https://doi.org/10.1111/1467-842X.00072.

[5] R.D. Gupta, D. Kundu, Discriminating between Weibull and generalized exponential distributions, Comput. Stat. Data Anal. 43 (2003) 179–196. https://doi.org/10.1016/S0167-9473(02)00206-2.

[6] D.M. Hawkins, Identification of outliers, Chapman and Hall, London, 1980.

[7] N. Kumar, S. Lalitha, Testing for upper outliers in gamma sample, Commun. Stat. – Theory Meth. 41 (2012) 820–828. https://doi.org/10.1080/03610926.2010.531366.

[8] J. Likeš, Some tests for k≥2 upper outliers in an exponential sample, Biom. J. 29 (1987) 313–321. https://doi.org/10.1002/bimj.4710290312.

[9] C.-T. Lin, S.-C. Wang, Discordancy tests for two-parameter exponential samples, Stat Papers. 56 (2015) 569–582. https://doi.org/10.1007/s00362-014-0597-3.

[10] A.K. Singh, S. Lalitha, A novel spatial outlier detection technique, Commun. Stat. – Theory Meth. 47 (2018) 247–257. https://doi.org/10.1080/03610926.2017.1301477.

[11] A.K. Singh, S. Lalitha, Detection of Upper Outliers in Gamma Sample, J. Stat. Appl. Probab. Lett. 5 (2018), 247-257. http://doi.org/10.18576/jsapl/050201.

[12] A.K. Singh, Multivariate Analysis of Crime Data using Spatial Outlier Detection Algorithm, J. Stat. Appl. Probab. 5 (2016), 433-438.