Arab J. Math. Sc.
Volume 7, Number 2, December 2001
pp. 43-63

A MULTISTAGE TEST FOR DETECTING MULTIPLE OUTLIERS IN THE NORMAL CASE

A.A. AL-SHIHA AND S.S. YANG

ABSTRACT. In this article, a multistage procedure for detecting multiple outliers is proposed. The proposed test statistic called $L_{n,k}$ used in this procedure is based on the generalized likelihood ratio principle. In each sequential step of the proposed procedure, $L_{n,k}$ is considered as a block test for testing multiple outliers in a random sample from a normal distribution with zero mean and unknown variance σ^2. The test statistic $L_{n,k}$ is scale-invariant, and hence, it is suitable for any value of σ^2. The test statistic $L_{n,k}$ is derived and compared with some well known single- and multiple outliers tests and found to be more powerful and less affected by the problems of swamping and masking. In addition, a real data example is worked out to demonstrate the use of the procedure.

1. INTRODUCTION

The problem of detecting outliers is old, yet it is of considerable importance in applied statistics. Many authors use "outlier" to indicate any observation that does not come from the target distribution. The literature on outliers is vast and the books by Barnett and Lewis (1978) and Hawkins (1980) and the paper by Beckman and Cook (1983) provide useful surveys of much of the relevant literature.

Outlier detection problems associated with the normal distributions have been generally treated as problems in hypotheses testing. The null hypothesis is generally that the set of all observations is a random sample from some

1991 *Mathematics Subject Classification.* 62F03, 62F99, 62A10, 62E25.

normal distribution with unknown mean and unknown variance, while the alternative hypothesis is that some of the observations came from different (in mean or variance or both) normal distributions. Most outlier detection procedures depend on location-scale invariant test statistics to accommodate the general normality case. Many formal techniques for the identification of outliers have been proposed in the literature, some of which are single-outlier procedures that are designed to test for only one outlier. Other methods are multiple outliers procedures that can detect more than one outlier in the sample. Multiple outliers procedures generally include the sequential application of a single outlier test to test for more than one outlier. Barnett and Lewis (1978) call the repeated application of single outlier tests for rejection of multiple outliers a consecutive test. Another type of multiple outliers tests, called block tests by Barnett and Lewis (1978), are used to test sets of data simultaneously in one step. The masking effect and the swamping effect are two problems associated with multiple outliers tests. The masking effect is the inability of a multiple outlier's procedure to identify even a single outlier in the presence of several potential outliers. This problem usually happens when outliers of similar magnitude form subgroups in the sample. This phenomenon makes procedures based on the sequential application of one-outlier tests insensitive. This problem also occurs with block tests when the true number of outliers is underestimated. On the other hand, the swamping effect is the tendency of a multiple outlier's procedure to declare more outliers than there are in the data. Many procedures for detecting multiple outliers suffer from the swamping effect when the maximum number of potential outliers assumed is much larger than the true number of outliers.

We propose a new statistic called $L_{n,k}$ for detecting multiple outliers. In each step of the proposed procedure, $L_{n,k}$ is considered as a block test for testing multiple outliers in a random sample from a normal distribution with zero mean and unknown variance σ^2. The test statistic $L_{n,k}$ is scale-invariant, and hence, the procedure is suitable for any value of σ^2.

In Section 2, we present a brief review for some early multiple outliers tests. In Section 3, we derive the test statistic $L_{n,k}$, describe the proposed multistage procedure for detecting multiple outliers, and compare it with some well known single-and multiple outliers tests. In Section 4, we present an example to demonstrate the use of the proposed multistage procedure.

2. REVIEW OF SOME EARLY WORK ON OUTLIERS DETECTION

Several procedures for detecting outlying observations in the normal case can be found in literature. Since we are concerned about procedures detecting multiple outliers, our discussion will be mainly about these types of procedures.

Dixon (1950) proposed criteria for detecting one outlier. His criteria are ratios of the distance between a suspected observation and its nearest neighbor to the range of the sample (with zero, one, or two observations omitted). Dixon (1953) has suggested a sequential application of his criteria for detecting multiple outliers.

Grubbs (1950) proposed some statistics to test the significance of the largest observation, the smallest observation, the two largest observations, the two smallest observations, and the largest and smallest observations. In the case of one outlier, Grubbs's statistic is statistically equivalent to the test statistic called Extreme Studentized Deviate (ESD) proposed by Pearson and Chandra Sekar (1936). Grubbs derived for the first time the exact distribution of ESD. ESD has been shown that it maximizes the probability of making the correct decision under the mean-shift model in the case of one outlier by Paulson (1952), and Kudo (1956), and hence, it is optimal in this sense. Ferguson (1961) showed the same property of ESD when one outlier is generated under the variance-inflation model. Therefore, we shall choose the ESD test in our comparison study that will be presented in this paper.

Ferguson (1961) has shown that tests based on the sample kurtosis are locally best invariant test for testing small shifts in the mean in either direction. He also showed that tests based on the sample kurtosis are locally best invariant test for testing small positive shifts in the variance. Ferguson (1961) suggested the sequential application of the kurtosis test to test for multiple outliers. Since tests based on the sample kurtosis were shown to be optimal, we shall include the test based on the sequential application of the kurtosis in our comparison study.

Tietjen and Moore (1972) generalized Grubbs's statistics to a class of statistics that can detect exactly k outliers. They proposed two Grubbs-type statistics for the detection of multiple outliers. For testing the k largest observations of the sample $x_1, x_2, ..., x_n$, they proposed the test statistic $L_k = \sum_{i=1}^{n-k}(x_{(i)} - \overline{x}_k)^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2$ where $x_{(i)}$ is the i-th smallest order statistic, $\overline{x}_k = \sum_{i=1}^{n-k} x_{(i)}/(n-k)$, and $\overline{x}$ is the mean of the full sample. For testing the k smallest observations they proposed the test statistic $L_k^* = \sum_{i=k+1}^{n} (x_{(i)} - \overline{x}_k^*)^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2$ where $\overline{x}_k^* = \sum_{i=k+1}^{n} x_{(i)}/(n-k)$. For testing exactly k outliers in the sample, they proposed the test statistic $E_k = \sum_{i=1}^{n-k}(z_{(i)} - \overline{z}_k)^2 / \sum_{i=1}^{n}(z_i - \overline{z})^2$ where $\overline{z}_k = \sum_{i=1}^{n-k} z_i/(n-k)$, and z_i is the observation whose absolute residual, $|z_i - \overline{z}|$, is the i-th largest. Tietjen and Moore suggested estimating k by using the largest gap in the order sample when using L_k and L_k^*. They did not suggest any method to estimate k when using E_k. Hawkins (1979) showed that even if k is correctly identified, E_k might detect wrong observations as outliers in the case where the outliers occur on one side of the data. For this reason, we choose not to include it in our comparison study.

Rosner (1975) developed multiple outliers test statistics which require only the knowledge of the maximum number of possible outliers, k. Hawkins (1978) and Cook and Beckman (1980) showed that Rosner's procedure suffers from the swamping effect, i.e., it tends to declare more outliers than there are in the sample. This particular problem is also confirmed by our simulation study when we compared Rosner's procedure with the $L_{n,k}$ procedure (not shown here). For this reason, we choose not to include it in our comparison study.

The proposed $L_{n,k}$ procedure will be compared with the sequential procedures based on ESD and KUR test statistics for detecting nonzero means in a random sample from a normal distribution with unknown variance.

3. DEVELOPMENT OF THE $L_{n,k}$ PROCEDURE

Suppose $x_1, x_2, ..., x_n$ are n observations which, under a null hypothesis, constitute a random sample from $N(0, \sigma^2)$. The alternative hypothesis is that m of these observations have nonzero means and they are considered as outliers. That is, $x_1, x_2, ..., x_n$ is a sample composed of two subsets

$$J_1 = \{x_{i_1}, x_{i_2}, ..., x_{i_{n-m}}\} \text{ and } J_2 = \{x_{i_{n-m+1}}, x_{i_{n-m+2}}, ..., x_{i_n}\},$$

where $x_{i_j} \sim N(0, \sigma^2), j = 1, 2, ..., n-m, x_{i_j} \sim N(\theta_{i_j}, \sigma^2), j = n-m+1, n-m+2, ..., n$, and m is some integer such that $0 \leq m < n$. The subset J_1 is called the set of "inliers" and the subset J_2 is called the set of "outliers". Our purpose is to identify the subsets J_1 and J_2.

Derivation of the Test Statistic $L_{n,k}$

The inference problem of detecting multiple outliers is formalized as follows. Suppose $x_1, x_2, ..., x_n$ are independent random variables such that $x_i \sim N(\theta_i, \sigma^2), i = 1, 2, ..., n$, where all parameters are unknown. Let k be the maximum number of possible outliers such that $0 < k < n$. Let $I = \{1, 2, ..., \binom{n}{k}\}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Let $F_k = \left\{A_{i(k)} : i \in I\right\}$ be the collection of all different subsets of size k of the distinct elements of $\{1, 2, ..., n\}$. Testing that the x_j's with $j \in A_{i(k)}$ for some $i \in I$ are outliers in the sample $x_1, x_2, ..., x_n$ is equivalent to testing:

$$\begin{aligned} &H_0: \quad \theta_1 = \theta_2 = ... = \theta_n = 0, \text{ against} \\ &H_1^{(i)}: \quad \theta_j \neq 0 \text{ for } j \in A_{i(k)}, \text{ and } \theta_j = 0 \text{ for } j \notin A_{i(k)}. \end{aligned} \tag{3.1}$$

The generalized likelihood ratio test (G.L.R.T.) statistic for testing (3.1) is:

$$\lambda_i(x, k) = \frac{\underset{H_0}{Sup}\, L(\theta_1, \theta_2, ..., \theta_n, \sigma^2 | x)}{\underset{H_1^{(i)}}{Sup}\, L(\theta_1, \theta_2, ..., \theta_n, \sigma^2 | x)},$$

where

$$L(\theta_1, \theta_2, ..., \theta_n, \sigma^2 | x) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\sum_{i=1}^{n} (x_i - \theta_i)^2 / (2\sigma^2)\right\}.$$

It can be easily shown that

$$\lambda_i(x,k) = \left(1 + \frac{\sum\limits_{j\in A_{i(k)}} x_j^2}{\sum\limits_{j\notin A_{i(k)}} x_j^2}\right)^{-n/2}$$

The null hypothesis in (3.1) is rejected for small values of $\lambda_i(x,k)$ or equivalently it is rejected for large values of the following statistic:

$$T_{i(k)} = \frac{\sum\limits_{j\in A_{i(k)}} x_j^2/k}{\sum\limits_{j\notin A_{i(k)}} x_j^2/(n-k)}.$$

Let $S_{i(k)} = \sum\limits_{j\in A_{i(k)}} x_j^2/\sigma^2$ and $U_{i(k)} = \sum\limits_{j\notin A_{i(k)}} x_j^2/\sigma^2$. Under $H_0, S_{i(k)}$ and $U_{i(k)}$ are independent Chi-square random variables with k and $n-k$ degrees of freedom, respectively. Therefore, $T_{i(k)} = \frac{S_{i(k)}/k}{U_{i(k)}/(n-k)}$ has F-distribution with k and $n-k$ degrees of freedom for all $i \in I$.

Now, to test for presence of exactly k outliers in the sample $x_1, x_2, ..., x_n$, we test:

(3.2) $\quad H_0: \ \theta_1 = \theta_2 = ... = \theta_n = 0$, against
$\quad H_1: \$ exactly k of the θ_i's are not equal to 0.

The G.L.R.T. statistic for testing (3.2) is given by:

$$\begin{aligned} \lambda(x,k) &= \frac{\underset{H_0}{Sup}\, L(\theta_1,\theta_2,\ldots,\theta_n,\sigma^2|x)}{\max\limits_{i\in I} \underset{H_1^{(i)}}{Sup}\, L(\theta_1,\theta_2,\ldots,\theta_n,\sigma^2|x)} \\ &= \min_{i\in I} \lambda_i(x,k) \\ &= \left(1 + \tfrac{k}{n-k} \max_{i\in I} T_{i(k)}\right)^{-n/2}. \end{aligned}$$

H_0 is rejected for small values $\lambda(x,k)$, or equivalently it is rejected for large values of the following statistic:

$$L_{n,k} = \max_{i\in I} T_{i(k)}. \qquad (3.3)$$

It is difficult to determine the null distribution of $L_{n,k}$ analytically even though $T_{i(k)} \sim F_{(k,n-k)}$ for all $i \in I$ because the $T_{i(k)}$'s are not independent random variables. Therefore, the percentiles of the null distribution of $L_{n,k}$ were empirically simulated. Table 3.1 gives some simulated percentiles of $L_{n,k}$ for some values of n and k.

Table 3.1: Some simulated percentiles of $L_{n,k}$

		(1-α) percentile of $L_{n,k}$					(1-α) percentile of $L_{n,k}$		
		α					α		
n	k	0.01	0.05	0.10	n	k	0.01	0.05	0.10
25	5	16.067	12.101	10.487	24	5	16.438	12.439	10.698
	4	15.233	11.516	10.037		4	15.437	11.719	10.161
	3	14.868	11.172	9.772		3	14.930	11.289	9.818
	2	15.132	11.159	9.709		2	15.128	11.327	9.727
	1	16.860	11.994	10.173		1	17.298	11.988	10.159
23	5	17.202	12.759	10.880	22	5	17.959	13.092	11.159
	4	15.983	11.975	10.309		4	16.591	12.175	10.463
	3	15.259	11.412	9.954		3	15.799	11.537	9.997
	2	15.277	11.369	9.778		2	15.881	11.564	9.748
	1	17.120	12.034	10.112		1	17.044	12.024	10.020

Let $y_i = x_i^2 (i = 1, 2, ..., n)$ and $y_{(i)}$ be the i-th smallest order statistic of the random sample $y_1, y_2, ..., y_n$. It can be shown that the statistic $L_{n,k}$ can be written as:

$$L_{n,k} = \frac{\sum_{j=n-k+1}^{n} y_{(j)}/k}{\sum_{j=1}^{n-k} y_{(j)}/(n-k)}. \tag{3.4}$$

The formula in (3.4) is easier than the formula in (3.3) for calculating $L_{n,k}$ for a given sample. Since in the successive stages of the proposed procedure, the test statistic $L_{n,k}$ is calculated with values of n and k reduced by one at each stage, the following recursive formula is useful in applying this procedure:

$$L_{n-j,k-j} = \frac{1}{k-j}\left[(k-j+1)L_{n-j+1,k-j+1} - \frac{y_{(n-j+1)}}{\overline{D}_{n-k}}\right]; j = 1, 2, ..., k-1,$$

where $\overline{D}_{n-k} = \sum_{i=1}^{n-k} y_{(i)}/(n-k)$.

3.2 Selection of k

The true number of the outliers, r, is generally unknown and cannot be specified in advance. Therefore, to apply an appropriate outlier detection procedure, a prior estimate, k, of r is required. If k is selected in advance and is smaller than the true number of outliers present, masking effect may cause the risk of not detecting any outliers even if they are present. On the other hand, if k is too large, swamping effect may lead to declaring valid observations as outliers. In our procedure, the initial estimate for r is denoted by k, which will be carefully chosen, is the number used to calculate $L_{n,k}$.

Marasinghe (1985), in identification of outliers in regression problem, points out that tests based on block statistics have enough power to reject the no-outlier hypothesis even when k is overestimated. Also, our simulation results show that the reduction in the power of the multiple outliers test using $L_{n,k}$ statistic is small when k is chosen to be slightly larger than the actual number of outliers, r (see Table 3.4). To estimate k, we propose to use the half-normal plot. In the half-normal plot the "inlier" observations tend to fall on a straight line passing through the origin and with a slope approximately equal to $1/\sigma$, while the outlying observations tend to fall "far" away from this line. Usually, a number $k_0 (k_0 \geq 0)$ of the observations clearly appear far away from the line, while a number $k_1 (k_1 \geq 0)$ of the observations may or may not fall on that line. We suggest that one should overestimate r by $k = k_0 + k_1$. This procedure means that any suspected observation will also be treated as an outlier in the initial process of estimating r by k.

3.3 The Multistage Procedure for Detecting Multiple Outliers

The proposed multistage procedure takes advantage of the property that tests based on block statistics like $L_{n,k}$ have enough power to reject the no-outlier hypothesis even when k is overestimated. It is based on using $L_{n,k}$ repeatedly. At each stage, one observation from the sample is deleted. The proposed procedure allows the percentage points of $L_{n,k}$ to be used repeatedly.. Marasinghe (1985) has used a similar approach for detecting multiple outliers in linear regression. Marasinghe (1985) points out that John and Draper (1978) have used a similar approach to his for detecting multiple outliers in two-way tables.

The nested sequence of hypotheses tested by the proposed multistage procedure can be reformulated as testing the following sequence of hypotheses. First, the null $H_0^{(1)}$: no-outliers versus $H_1^{(1)}$: at most k outliers is tested using $L_{n,k}$. If $H_0^{(1)}$ is rejected, the observation that corresponds to the largest squared value is removed and $H_0^{(2)}$: no-outliers versus $H_1^{(2)}$: at most $(k-1)$ outliers is tested using $L_{n-1,k-1}$, computed from the remaining sample. This process is continued until a null hypothesis $H_0^{(m+1)}$: no-outliers versus $H_1^{(m+1)}$: at most $(k-m)$ outliers fails to be rejected $(m \leq k)$. This procedure assumes that the observation removed in each stage is a true outlier, in which case the Type I error rate at each stage is the level specified for the test. In this procedure, the percentage points of $L_{n,k}$ can be used in successive stages with values of n and k reduced by one at each stage. The final estimate of r obtained by this sequential procedure is actually m and the declared outliers are those observations that correspond to $y_{(n)}, y_{(n-1)}, \ldots, y_{(n-m+1)}$. In applying this procedure, and in order not to miss potential outliers, we suggest using a large significance level such as $\alpha = 0.1$.

3.4 Power Comparison Between $L_{n,1}$, ESD, and KUR for Detecting One Outlier

We have mentioned earlier that the procedures of detecting one outlier based on the statistics ESD and KUR have some optimal properties. In this section, we empirically compare powers of these tests to the power of $L_{n,1}$ for some chosen sample sizes. The power of a procedure for detecting outliers is defined as the ratio of the number of the samples detecting the correct number of outliers to the total number of samples. The statistics ESD, and KUR are defined as follows:

$$ESD = \frac{\max\limits_{i=1,\ldots,n} |x_i - \overline{x}|}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 / n}},$$

$$KUR = \frac{n \sum\limits_{i=1}^{n} (x_i - \overline{x})^4}{\left[\sum\limits_{i=1}^{n} (x_i - \overline{x})^2\right]^2},$$

where $\overline{x} = \sum_{i=1}^{n} x_i/n$.

First, we compare the empirical powers for detecting one outlier in the presence of exactly one outlier in the sample. This comparison was performed for $n = 10$, and 20, and for $\alpha = 0.01$, and 0.05. In this simulation study we consider $x_i \sim N(0,1), i = 1, 2, ..., n-1$, and $x_n \sim N(\lambda, 1)$, and $x_1, x_2, ..., x_n$ are independent, and $\lambda = 0, 1, ..., 10$. For each combination of n, α, and λ, we independently simulated 10,000 random samples from $N(0,1)$. For each random sample, the outlier was introduced by adding known quantity λ to the last observation and computed the values of $L_{n,1}$ ESD, and KUR. The empirical power of each test is then the total number of the random samples that rejected the no-outlier hypothesis divided by 10,000. The results of this study are presented in Table 3.2 and Figure 3.1. This simulation study shows that $L_{n,1}$ is more powerful than both ESD and KUR for almost all combinations of n, α, and λ.

Second in order to investigate the masking effect on each of the tests $L_{n,1}$,ESD, and KUR, we compare the empirical powers for detecting one outlier in presence of exactly two outlier in the sample. This comparison was performed for $n = 20$ and $\alpha = 0.05$. In this simulation study we consider $x_i \sim N(0,1), i = 1, 2, ..., n-2, x_{n-1} \sim N(\lambda, 1)$, and $x_n \sim N(\lambda_2, 1)$, and $x_1, x_2, ..., x_n$ are independent. We chose $\lambda_2 = 5, 6, 7$, and 8, and for each value of λ_2, we used $\lambda_1 = 0, 2, 3, ..., \lambda_2$. For each combination of n, λ_1, and λ_2, we independently simulated 10,000 random samples from $N(0,1)$. For each random sample, the outliers were introduced by adding known quantity λ_1 to the $(n-1)$-th observation and by adding known quantity λ_2 to the n-th observation and computed the values of $L_{n,1}$, ESD, and KUR. The empirical power of each test is then the total number of the random samples that rejected the no-outlier hypothesis divided by 10,000. The results of this study are presented in Table 3.3 and Figure 3.2. This simulation study shows that $L_{n,1}$ is less effected by the masking problem and more powerful than ESD for almost all combinations of n, λ_1, and λ_2. It is also true that $L_{n,1}$ is less effected by the masking problem and slightly more powerful than KUR for most of the combinations of n, λ_1, and λ_2.

Table 3.2: Empirical Powers of $L_{n,1}$, ESD, and KUR for Detecting One Outlier in Presence of One Outlier (This Table Is Illustrated By Figure 3.1)

		$\alpha = 0.01$			$\alpha = 0.05$		
n	λ	$L_{n,1}$	ESD	KUR	$L_{n,1}$	ESD	KUR
10	0.0	0.0118	0.0117	0.0111	0.0534	0.0526	0.0536
	1.0	0.0131	0.0134	0.0106	0.0613	0.0533	0.0533
	2.0	0.0406	0.0302	0.0277	0.1300	0.1031	0.1028
	3.0	0.1515	0.1062	0.0955	0.3433	0.2679	0.2616
	3.5	0.2483	0.1779	0.1649	0.4865	0.3867	0.3787
	4.0	0.3705	0.2739	0.2581	0.6463	0.5300	0.5187
	4.5	0.4925	0.3698	0.3512	0.7713	0.6655	0.6501
	5.0	0.6251	0.4997	0.4761	0.8633	0.7716	0.7544
	5.5	0.7403	0.6051	0.5810	0.9286	0.8584	0.8463
	6.0	0.8365	0.7153	0.6887	0.9686	0.9245	0.9157
	6.5	0.9092	0.8106	0.7890	0.9866	0.9589	0.9536
	7.0	0.9496	0.8786	0.8632	0.9956	0.9829	0.9789
	8.0	0.9886	0.9568	0.9472	0.9993	0.9972	0.9964
	9.0	0.9985	0.9889	0.9864	1.0000	0.9999	0.9998
	10.0	0.9998	0.9979	0.9969	1.0000	0.9999	0.9998
20	0.0	0.0112	0.0105	0.0080	0.0505	0.0508	0.0477
	1.0	0.0141	0.0136	0.0119	0.0567	0.0540	0.0523
	2.0	0.0448	0.0418	0.0375	0.1263	0.1173	0.1144
	3.0	0.1715	0.1528	0.1370	0.3632	0.3256	0.3166
	3.5	0.3105	0.2790	0.2497	0.5383	0.4914	0.4764
	4.0	0.4713	0.4307	0.3977	0.7073	0.6641	0.6429
	4.5	0.6322	0.5837	0.5446	0.8333	0.7929	0.7752
	5.0	0.7704	0.7248	0.6942	0.9153	0.8846	0.8707
	5.5	0.8730	0.8364	0.8105	0.9647	0.9524	0.9430
	6.0	0.9428	0.9224	0.9045	0.9874	0.9809	0.9756
	6.5	0.9756	0.9630	0.9519	0.9956	0.9927	0.9917
	7.0	0.9915	0.9851	0.9802	0.9994	0.9986	0.9975
	8.0	0.9993	0.9985	0.9975	1.0000	1.0000	1.0000
	9.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	10.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 3.1: Empirical Power of $L_{n,1}$, *ESD*, and *KUR* for Detecting One Outlier in Presence of One Outlier (n = 10, 20, and α = 0.01, 0.05)

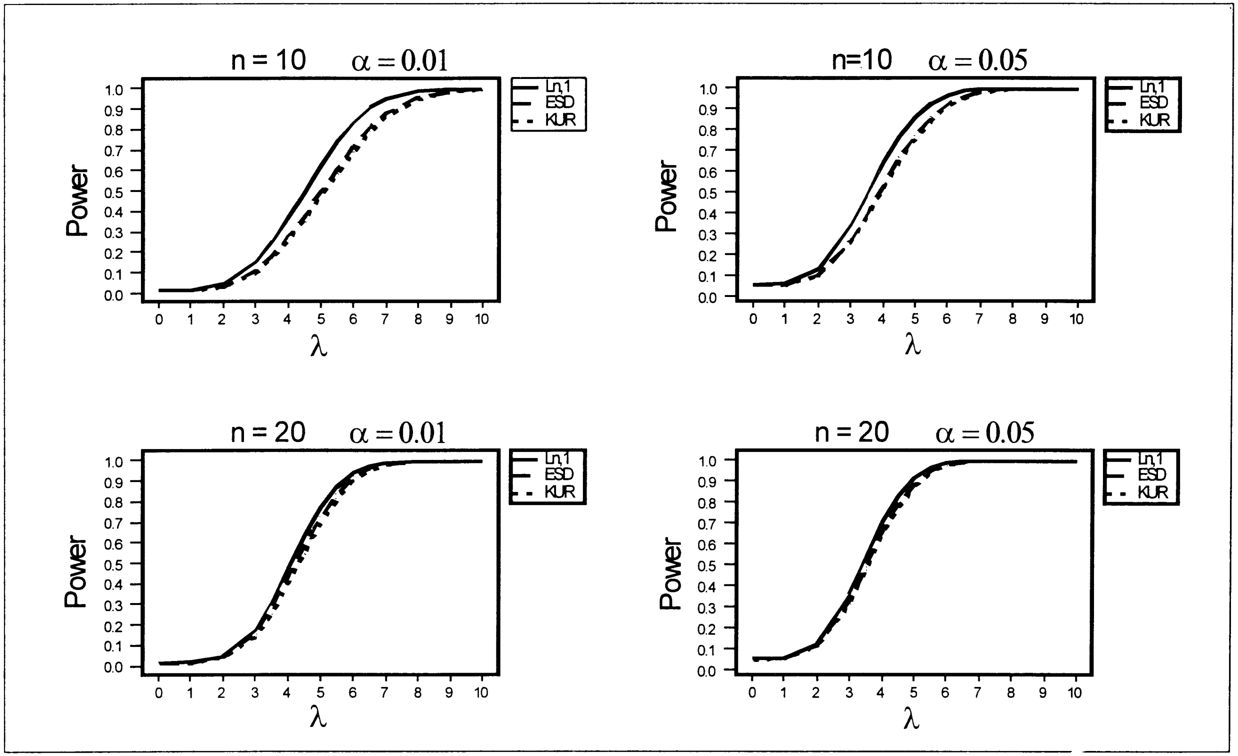
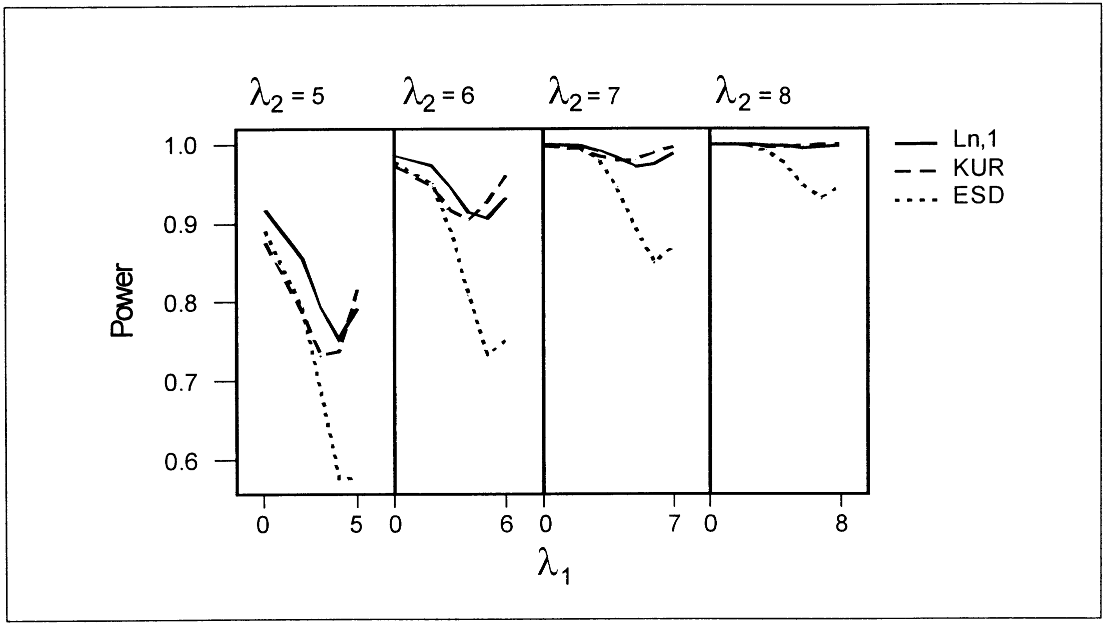


Table 3.3: Empirical Powers of $L_{n,1}$, ESD, and KUR for Detecting One Outlier in Presence of Two Outlier for The Case where n =20 and α = 0.05 (This Table Is Illustrated By Figure 3.2)

λ_1	λ_2	$L_{n,1}$	ESD	KUR
0	5	0.9165	0.8902	0.8752
2		0.8544	0.7934	0.7867
3		0.7933	0.6874	0.7327
4		0.7534	0.5764	0.7378
5		0.7911	0.5781	0.8153
0	6	0.9851	0.9771	0.9731
2		0.9726	0.9497	0.9460
3		0.9447	0.8923	0.9161
4		0.9134	0.8088	0.9046
5		0.9058	0.7330	0.9264
6		0.9313	0.7510	0.9603
0	7	0.9990	0.9986	0.9981
2		0.9976	0.9946	0.9935
3		0.9916	0.9799	0.9844
4		0.9824	0.9446	0.9790
5		0.9717	0.8921	0.9803
6		0.9741	0.8495	0.9885
7		0.9876	0.8694	0.9960
0	8	0.9999	0.9999	0.9999
2		0.9999	0.9997	0.9997
3		0.9996	0.9976	0.9984
4		0.9980	0.9911	0.9966
5		0.9971	0.9754	0.9967
6		0.9936	0.9474	0.9974
7		0.9960	0.9311	0.9991
8		0.9981	0.9459	0.9992

Figure 3.2: Empirical Power of $L_{n,1}$, *ESD*, and *KUR* for Detecting One Outlier in Presence of Two Outlier (n = 20, and $\alpha = 0.05$)



3.5 Power Comparisons Between $L_{n,k}$ and Sequential Applications of $L_{n,1}$, ESD, and KUR for Detecting Multiple Outliers

In this section, we compare the empirical performance of the new multi-stage multiple outliers procedure with the sequential application of the tests based on the statistics $L_{n,1}$, ESD, and KUR. The processes of generating the random samples and introducing the outliers used in this section are similar to those used in Section 3.4. In this simulation study, we used $n = 25, \alpha = 0.05$. The performance of each procedure when at most 3 outliers were present in the sample was evaluated. The empirical power of a test is the number of random samples in which the test detects the correct number of outliers divided by the total numbers of samples. The results of this study are presented in Table 3.4.

The numbers presented in Table 3.4 are the percentages of trials in which the indicated number of outliers (m) was detected by each procedure. The patterns of contamination including a no-outlier case used in this study are also shown in Table 3.4. The effect of using different initial values of k in the multistage procedure was examined by repeating the procedure for $k = 1, 2, 3, 4$, and 5.

From this simulation study we conclude several results as follows. First, the empirical size of the test of the new procedure is almost equal to the specified significance level, viz. $\alpha = 0.05$, whatever initial value is chosen for k. This indicates that the new procedure adequately controls Type I error. Second, for small λ_1(i.e., for the case (5,0,0)), the power of the new procedure, when only one outlier is present, is slightly less than that of sequential $L_{n,1}$,ESD, and KUR only when the true number of outliers are overestimated, (i.e., when using $k \geq 2$) and it decreases slightly for large values of k. For large values of λ_1(i.e., for the case (10,0,0)), the power of the new procedure is almost greater or equal to those of sequential $L_{n,1}$, ESD, and KUR for all values of k. Third, the power of detecting multiple outliers of the $L_{n,k}$ procedure is always larger than those of ESD, and KUR when the initial values of k, are chosen equal to or larger than the true number of outliers. Fourth, the power of detecting multiple outliers based on the sequential $L_{n,1}$ procedure is always larger than those of ESD, and KUR. Fifth, when r is correctly estimated or slightly overestimated, the $L_{n,k}$ multistage procedure is always more powerful than the sequential $L_{n,1}$

procedure, while the sequential $L_{n,1}$ is slightly more powerful only when r is much overestimated. Sixth, when k is correctly chosen, the new procedure is more powerful than the sequential $L_{n,1}$, ESD, and KUR for all cases. Seventh, the new procedure is not effected by the problem of swamping, whereas the sequential procedures based on ESD and KUR highly suffer from this problem, see for example the cases (10,5,0), (10,10,0) and (15,10,5). Eighth, in detecting multiple outliers and when r is correctly estimated or slightly overestimated, the $L_{n,k}$ multistage procedure is generally less effected by the problem of masking than the sequential procedures based on $L_{n,1}$, ESD and KUR, see for example the cases (5,5,5) and (5,5,-5). Finally, this result shows that the power of detecting multiple outliers using the new multistage procedure remains high when the initial value of k is chosen equal to or larger than r. Moreover, the decrease in the power when the initial value of k is chosen slightly larger than r is small.

Table 3.4: Performance of Multiple outliers Procedures ($n = 25, \alpha = 0.05$)

No. of Outliers (r)	Outlier Pattern $(\lambda_1, \lambda_2, \lambda_3)$	No. of Outliers Found (m)	Sequential procedures			Multistage procedure Using $L_{n,k}$				
			ESD	*KUR*	$L_{n,1}$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0	(0,0,0)	0	95.0	.94.8	95.0	95.0	95.1	95.2	95.1	95.0
		≥ 1	5.0	5.2	5.0	5.0	4.9	4.8	4.9	5.0
1	(5,0,0)	0	9.8	10.8	8.0	8.0	11.2	17.0	22.1	26.6
		1	85.5	84.0	87.3	92.0	83.8	78.0	72.6	68.1
		≥ 2	4.7	5.2	4.7	—	5.0	5.0	5.3	5.3
1	(10,0,0)	0	0	0	0	0	0	0	0	0
		1	95.2	95.2	95.2	100	95.2	95.6	95.2	95.1
		≥ 2	4.8	4.8	4.8	—	4.8	4.4	4.8	4.9
2	(5,5,0)	0	23.1	8.5	11.3	11.3	1.9	2.8	4.1	6.1
		1	11.3	13.2	11.6	88.7	13.7	19.2	23.4	28.5
		2	40.6	48.8	73.0	—	84.4	73.1	67.6	60.6
		≥ 3	25.0	29.5	4.1	—	—	4.9	4.9	4.8
2	(10,5,0)	0	0	0	0	0	0	0	0	0
		1	10.1	11.4	8.5	100	8.5	12.5	16.9	22.8
		2	51.7	51.5	86.8	—	91.5	82.6	78.3	72.7
		≥ 3	38.2	37.1	4.7	—	—	4.9	4.8	4.5
2	(10,10,0)	0	0	0	0	0	0	0	0	0
		1	0	0	0	100	0	0	0	0
		2	60.5	60.6	94.9	—	100	94.9	95.1	95.5
		≥ 3	39.5	39.4	5.1	—	—	5.1	4.9	4.5
2	(5,-5,0)	0	9.1	2.2	11.0	11.0	1.8	2.5	3.6	5.2
		1	14.4	17.9	11.5	89.0	13.6	19.0	24.4	29.2
		2	47.7	49.7	73.6	—	84.6	73.6	66.8	61.0
		≥ 3	28.8	30.2	3.9	—	—	4.9	5.2	4.6
2	(10,-10,0)	0	0	0	0	0	0	0	0	0
		1	0	0	0	100	0	0	0	0
		2	59.7	59.6	94.3	—	100	94.3	94.7	95.0
		≥ 3	40.3	40.4	5.7	—	—	5.7	5.3	5.0
3	(5,5,5)	0	69.0	25.7	46.2	46.2	4.2	1.1	1.1	1.6
		1	10.1	5.8	10.3	53.8	21.1	4.1	4.6	6.6
		2	3.9	8.5	7.2	—	74.7	15.4	21.5	29.7
		3	7.3	31.8	34.4	—	—	79.4	67.8	56.7
		≥ 4	9.7	28.2	1.9	—	—	—	5.0	5.4
3	(10,5,5)	0	0.2	0	0	0	0	0	0	0
		1	25.4	9.5	13.6	100	13.6	2.1	2.6	4.0
		2	9.6	10.4	10.4	—	86.4	12.6	18.1	25.9
		3	31.0	38.2	72.3	—	—	85.3	74.6	65.1
		≥ 4	33.8	41.9	3.7	—	—	—	4.7	5.0
3	(10,10,5)	0	0	0	0	0	0	0	0	0
		1	0	0	0	100	0	0	0	0
		2	6.9	7.4	8.3	—	100	8.3	12.3	18.7
		3	3.6	4.0	87.2	—	—	91.7	83.0	76.1
		≥ 4	89.5	88.6	4.5	—	—	—	4.7	5.2
3	(5,5,-5)	0	44.9	4.0	46.4	46.4	4.7	1.0	1.2	1.8
		1	13.2	8.5	10.3	53.6	21.2	4.4	5.0	6.9
		2	7.0	12.7	7.5	—	74.1	15.6	22.2	29.7
		3	15.8	35.8	34.0	—	—	79.0	66.6	56.3
		≥ 4	19.1	39.0	1.8	—	—	—	5.0	5.3
3	(5,5,-10)	0	0	0	0	0	0	0	0	0
		1	23.8	9.6	12.8	100	12.8	2.1	2.7	4.1
		2	9.9	10.5	10.7	—	87.2	13.0	18.3	25.6
		3	32.5	38.4	72.4	—	—	84.9	73.9	65.0
		≥ 4	33.7	41.6	4.0	—	—	—	5.0	5.3
3	(10,5,-10)	0	0	0	0	0	0	0	0	0
		1	0	0	0	100	0	0	0	0
		2	. 8.3	9.2	8.0	—	100	8.0	11.9	18.1
		3	18.4	18.4	87.5	—	—	92.0	83.3	76.8
		≥ 4	73.3	72.4	4.5	—	—	—	4.8	5.1
3	(15,15,-15)	0	7.4	0	9.4	9.4	0	0	0	0
		1	0	0	0	90.6	0	0	0	0
		2	0	0	0	—	100	0	0	0
		3	50.2	55.7	86.2	—	—	100	95.1	94.9
		≥ 4	42.4	44.3	4.4	—	—	—	4.9	5.1

4. AN EXAMPLE

To demonstrate the use of the multistage procedure proposed for detecting multiple outliers, we present the following data. Daniel (1959) reported the results of a 2^5 factorial experiment where the 31 contrasts arranged in order of increasing absolute value as follows.

0.0000	0.0281	−0.0561	−0.0842	−0.0982	0.1263	0.1684	0.1964
0.2245	−0.2526	0.2947	−0.3087	0.3929	0.4069	0.4209	0.4350
0.4630	−0.4771	0.5472	0.6595	0.7437	−0.7437	−0.7577	−0.8138
−0.8138	−0.8980	1.0800	−1.3050	2.1470	−2.6660	−3.1430	

The half-normal probability plot of this data is shown in Figure 4.1. Table 4.1 summarizes the calculations for four possible selections of k_0 and k_1 based on the half-normal probability plot. We realize that the results of the test coincide for all cases in declaring three outliers at $\alpha = 0.10$. The three outliers declared are 2.147, -2.666, and -3.143. This result supports our claim that slightly overestimating the true number of outliers, r, by the initial estimate k will not cause a much decrease in the power of the test.

Figure 4.1: The half-normal probability plot

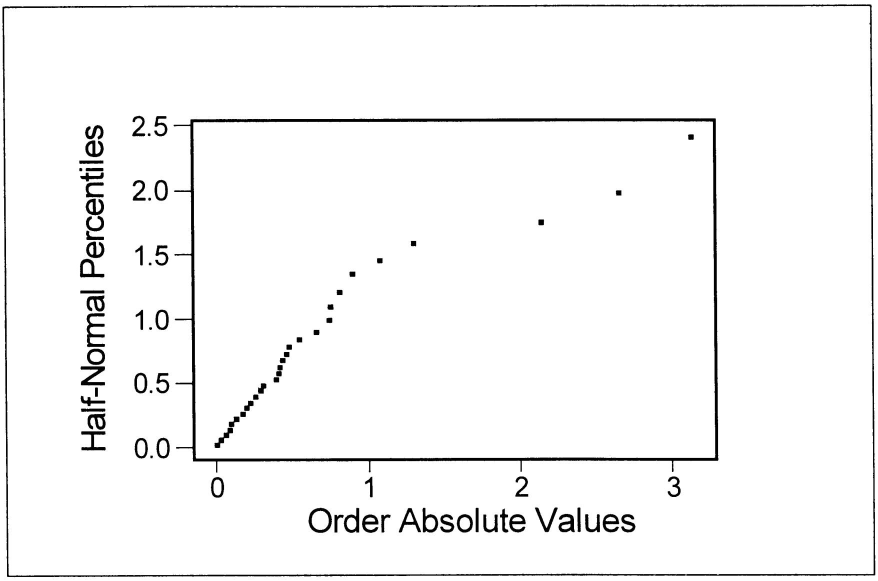


Table 4.1: Calculation of the test

	n	k	$L_{n,k}$	Simulated Critical Values of $L_{n,k}$ α =0.05	$\alpha = 0.10$
a. $k_0 = 3$	31	3	22.5376**	10.760	9.467
$k_1 = 0$	30	2	18.3424**	11.043	9.625
$k = 3$	29	1	14.4320**	12.026	10.151
b. $k_0 = 3$	31	4	21.7211**	10.848	9.572
$k_1 = 1$	30	3	16.6821**	10.754	9.491
$k = 4$	29	2	11.7704**	11.075	9.699
	28	1	6.3509	11.999	10.131
c. $k_0 = 3$	31	5	20.9453**	11.103	9.830
$k_1 = 2$	30	4	15.6101**	10.879	9.611
$k = 5$	29	3	10.6718*	10.830	9.555
	28	2	6.1415	11.055	9.701
	27	1	4.9930	12.004	10.165
d. $k_0 = 3$	31	6	19.9904**	11.496	10.187
$k_1 = 3$	30	5	14.6115**	11.240	9.888
$k = 6$	29	4	9.8310*	10.995	9.658
	28	3	5.8154	10.896	9.618
	27	2	4.6816	11.050	9.679
	26	1	3.8273	11.991	10.122

**The corresponding observation is declared as an outlier at $\alpha = 0.05$.

*The corresponding observation is declared as an outlier at $\alpha = 0.10$.

REFERENCES

1. V. Barnett and T. Lewis, *Outliers in Statistical Data*, New York: John Wiley 1978.

2. R.J. Beckman and R.D. Cook, *Outlier ....s*, Technometrics **25**(1983), 119-149.

3. R.D. Cook and R.J. Beckman, *Using M-estimators to identify outliers*, Los Alamos Report LA-UR 80-1418, Los Alamos, New Mexico, 1980.

4. C. Daniel, *Use of half-normal plots in interpreting factorial two-level experiments*, Technometrics **1**(1959), 311-341.

5. W.J. Dixon, *Analysis of extreme values*, Annals of Mathematical Statistics **21**(1950), 488-506.

6. W.J. Dixon, *Processing data for outliers*, Biometrics **9**(1953), 74-89.

7. T.S. Ferguson, *On the rejection of outliers*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability Vol. 1, Berkeley and Los Angeles: University of California Press (1961), 253-287.

8. F.E. Grubbs, *Sample criteria for testing outlying observations*, Annals of Mathematical Statistics **21**(1950), 27-58.

9. D.M. Hawkins, *Letter to the editor*, Technometrics, **20**(1978), 218.

10. D.M. Hawkins, *Fractiles of an extended multiple outlier test*, Journal of Statistical Computation and Simulation **8**(1979), 227-236.

11. D.M. Hawkins, *Identification of outliers*, London: Chapman and Hall, 1980.

12. J.A. John and N.R. Draper, *On testing for two outliers or one outlier in two-way tables*, Technometrics **20**(1978), 69-78.

13. A. Kudo, *On testing outlying observations*, Sankhya **17**(1956), 67-76.

14. M.G. Marasinghe, *A multistage procedure for detecting several outliers in linear regression*, Technometrics **27**(1985), 395-399.

15. E. Paulson, *An optimum solution to the K-sample slippage problem for the normal distribution*, Annals of Mathematical Statistics **23**(1952), 610-616.

16. E.S. Pearson and Sekar C. Chandra, *The efficiency of statistical tools and a criterion for the rejection of outlying observations*, Biometrika **28**(1936), 308-320.

17. B. Rosner, *On the detection of many outliers*, Technometrics **17**(1975), 221-227.

18. G.L. Tietsjen and R.H. Moore, *Some Grubbs-type statistics for the detection of several outliers*, Technometrics **14**(1972), 583-597.

Department of Statistics and Operations Research, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia
e-mail: aalshiha@ksu.edu.sa

Department of Statistics and Statistical Laboratory, Kansas State University, Manhattan, Kansas, U.S.A.
e-mail: ssyang@stat.ksu.edu

Date received July 26, 2000.