A New Approximation for the Null Distribution of the Likelihood Ratio Test Statistics for k Upper Outliers in a Normal Sample

Mahmoudvand, Rahim and Hassani, Hossein

Payame Noor University, IRAN , Cardiff University, UK

01. September 2005

مجله‌ی پژوهش‌های آماری ایران

J. Statist. Res. Iran **2** (2005): 155-161

پژوهشکده‌ی آمار دوره‌ی ۲، شماره‌ی ۲، پاییز و زمستان ۱۳۸۴، صص ۱۶۴-۱۷۰

A New Approximation for the Null Distribution of the Likelihood Ratio Test Statistics for *k* Outliers in a Normal Sample

Rahim Mahmoudvand[†,*] and Hossein Hassani[‡]

[†] Payam-e Noor University of Toyserkan

[‡] Central Bank of the Islamic Republic of Iran, and Cardiff University

1 Summary

Usually when performing a statistical test or estimation procedure, we assume the data are all observations of i.i.d. random variables, often from a normal distribution. Sometimes, however, we notice in a sample one or more observations that stand out from the crowd. These observation(s) are commonly called outlier(s). Outlier tests are more formal procedures which have been developed for detecting outliers when a sample comes from a normal distribution (Thode, 2002). A lot of work has been done for testing outliers in a univariate sample, most of which corresponds to the normal and exponential distribution. Barnett and Lewis (1994) have presented a summary of tests for outliers and their critical values, many of which are specific to the detection of outliers in normal samples. The theoretical solution for the exact null distribution of the likelihood ratio test for $k = 1$ is solved by Barnett and Lewis (1994) and Zhang and Yu (2006) for the case $k = 2$, but the problem is still open for $k \geqslant 3$. In this paper we introduce a new approximation for the null distribution of the likelihood ratio test for the general case. We compare the critical values obtained by the new approximation to the values, which are obtained by the exact distribution for the cases $k = 1, 2$ to test the accuracy of the new approximation. Also, the results are compared to another approximation method (which is known by Barnett and Lewis (1994)) for the cases $k = 3, 4$.

[*] Corresponding author

2 The Null Distribution of $T_n^{(k)}$

Let $x_1, \ldots, x_n$ be the random sample taken from the normal distribution N(μ, σ^2) (under the null hypothesis), where both μ and σ^2 are unknown, and $x_{(1)}, \ldots, x_{(n)}$ be the corresponding ordered statistics. Suppose the aim is to examine that whether $x_{(n-k+1)}, \ldots, x_{(n)}$ are the k upper outliers in this sample. Under this assumption the alternative hypothesis can be formulated as $x_{(1)}, \ldots, x_{(n-k)}$ belong to the N(μ, σ^2), and $x_{(n-k+1)}, \ldots, x_{(n)}$ (the k upper data) belong to the N($\mu + c, \sigma^2$), where $c > 0$ is a constant. To answer whether the k upper data are outliers, we can build the hypothesis $H_0 : c = 0$ against $H_1 : c > 0$.

The likelihood ratio test statistic for testing k upper outliers ($T_n^{(k)}$) and k lower outliers ($T_{n,(k)}$) are (Barnett and Lewis, 1994)

$$T_n^{(k)} = \frac{x_{(n)} + \cdots + x_{(n-k+1)} - k\bar{x}}{s}, \qquad T_{n,(k)} = \frac{k\bar{x} - x_{(1)} - \cdots - x_{(k)}}{s}, \quad (1)$$

where $\bar{x}$ and s are the sample mean and the sample standard deviation, respectively. Large values of $T_n^{(k)}$ reject the null hypothesis H_0, or recognizing that the set of ordered observations $\{x_{(n-k+1)}, \ldots, x_{(n)}\}$ as discordant observations (which are named as k outliers). Please note that $(x_{(n)}, \ldots, x_{(1)}) \overset{d}{=} (-x_{(1)}, \ldots, -x_{(n)})$ then $T_n^{(k)} \overset{d}{=} T_{n,(k)}, (k = 1, \ldots, n-1)$. In addition, one can see that $T_n^{(k)} \overset{d}{=} T_{n,(n-k)}$ and $T_n^{(k)} \overset{d}{=} T_n^{(n-k)}$. So, without loss of generality, we need just to discuss the sample distribution of $T_n^{(k)}$.

3 Some Approximations for the Null Distribution of $T_n^{(k)}$

In this section we introduce a new approximation for the exact distribution of $T_n^{(k)}$. Split the index set $I = \{1, \ldots, n\}$ into

$$[I]_k = \{i_1 \leqslant \cdots \leqslant i_k\}, \qquad 1 \leqslant i_1 < \cdots < i_k \leqslant n,$$

and

$$(I)_k = I \backslash [I]_k.$$

Set the notation

$$T_{[I]_k} = \frac{\sum_{i \in [I]_k} x_i - k\bar{x}}{s}. \quad (2)$$

By definition (2), we can rewrite $T_n^{(k)} = \max_{[I]_k} T_{[I]_k}$, so the distribution of the $\max_{[I]_k} T_{[I]_k}$ is the upper bound for the null distribution of $T_n^{(k)}$. It can be simply proved that the $\binom{n}{k}$ random variables $T_{[I]_k}$ have the distribution

$$f_{T_{[I]_k}}(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\sqrt{\frac{n}{k(n-k)(n-1)}}\left(1-\frac{n}{k(n-k)(n-1)}x^2\right)^{\frac{n-2}{2}-1}, \tag{3}$$

where $|x| \leq \sqrt{\frac{k(n-k)(n-1)}{n}}$.

3.1 Approximation I

Let $A_1, \ldots, A_m$ be m arbitrary events. Based on the probability laws, one can write

$$P\left(\bigcap_{k=1}^{m} A_i\right) \leqslant P(A_k), \quad k = 1, 2, \ldots, m$$

It is easy to show that

$$\begin{aligned} P\left(T_n^{(k)} \leqslant t\right) &= P\left(\max_{[I]_k} T_{[I]_k} \leqslant t\right) \\ &= P\left(\bigcap_{[I]_k}\{T_{[I]_k} \leqslant t\}\right) \leqslant P\left(T_{[I]_k} \leqslant t\right). \end{aligned} \tag{4}$$

Based on the Bonferroni general inequality we can write that

$$P\left(\bigcap_{k=1}^{m} A_k\right) \geqslant \sum_{k=1}^{m} P(A_k) - m + 1.$$

Barnet and Lewis (1994) introduced an approximation to calculate the critical values of the null distribution of the likelihood ratio test for $k \geqslant 2$ based on the above inequality, as below:

$$P\left(T_n^{(k)} \leqslant t\right) \geqslant m \int_L^t f_{T_{[I]_k}}(x)\, dx - m + 1, \qquad L = -\sqrt{\frac{k(n-k)(n-1)}{n}}, \tag{5}$$

where $m = \binom{n}{k}$.

3.2 Approximation II

To introduce a new approximation, we prove that $T_n^{(k)}$ are asymptotically independent. To do that, we just need to consider two conditions: the first one is to show that $T_n^{(k)}$ are asymptotically normally distributed and the second one is that $T_n^{(k)}$ are asymptotically uncorrelated. The first one (normality) is easily obtained by taking the limit

$$\lim_{n\to\infty}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\sqrt{\frac{n}{k(n-k)(n-1)}}\left(1-\frac{n}{k(n-k)(n-1)}x^2\right)^{\frac{n-2}{2}-1}$$
$$=\frac{1}{\sqrt{2\pi k}}\exp\left\{-\frac{x^2}{2k}\right\}.$$

To consider the second one, define $\left(T_{[I]_k},T_{[J]_k}\right)$, $(T_{[S]_k}=T_{\{s_1,\ldots,s_k\}}$, for $S=I,J)$. By this definition, the coefficient correlation between the pair $\left(T_{[I]_k},T_{[J]_k}\right)$ is

$$\rho\left(T_{\{i_1,\ldots,i_k\}},T_{\{j_1,\ldots,j_k\}}\right)=\frac{zn-k^2}{kn-k},\qquad z=0,\ldots,k-1 \tag{6}$$

where z is the number of the equal indices in the pair $\left(T_{[I]_k},T_{[J]_k}\right)$. Then,

$$\lim_{n\to\infty}\rho\left(T_{\{i_1,\ldots,i_k\}},T_{\{j_1,\ldots,j_k\}}\right)=\frac{z}{k},\qquad z=0,\ldots,k-1 \tag{7}$$

On the other hand, the frequency distribution of the discrete variable Z is

z	0	1	2	$\ldots$	$k-1$
f_i	$\binom{n}{k}\binom{k}{0}\binom{n-k}{k}$	$\binom{n}{k}\binom{k}{1}\binom{n-k}{k-1}$	$\binom{n}{k}\binom{k}{2}\binom{n-k}{k-2}$	$\ldots$	$\binom{n}{k}\binom{k}{k-1}\binom{n-k}{1}$

By simple calculation, it can be shown that $P(Z=0)\to 1$ as $n\to\infty$. We then accept that $T_{[I]_k}$ are asymptotically uncorrelated when n goes to infinity. We can now assume that $T_{[I]_k}$ are asymptotically independent. Based on the discussion above, we introduced a new approximation for the exact distribution of the $T_{[I]_k}$ for the general case as below.

$$P\left(T_n^{(k)}\leqslant t\right)=P\left(\max_{[I]_k}T_{[I]_k}\leqslant t\right)=P\left(\bigcap_{[I]_k}\left(T_{[I]_k}\leqslant t\right)\right)$$
$$\approx\left(\int_L^t f_{T_{[I]_k}}(x)\,dx\right)^{\binom{n}{k}},\qquad k=1,2,\ldots,n-1 \tag{8}$$

It must be mentioned that the identical distribution of the exact distribution is kept for the third approximation formulae. It is easy to show that $T_{[I]_k} \overset{d}{=} T_{[I]_{n-k}}$, and then

$$\begin{aligned} P\left(T_n^{(n-k)} \leqslant t\right) &= P\left(\max_{[I]_{n-k}} T_{[I]_{n-k}} \leqslant t\right) = P\left(\bigcap_{[I]_{n-k}} (T_{[I]_{n-k}} \leqslant t)\right) \\ &\approx \left(\int_L^t f_{T_{[I]_{n-k}}}(x)\, dx\right)^{\binom{n}{n-k}} \\ &= \left(\int_L^t f_{T_{[I]_k}}(x)\, dx\right)^{\binom{n}{k}}, \qquad k = 1, \ldots, n-1 \end{aligned} \tag{9}$$

We then conclude that the presented approximation distribution is the same for $T_n^{(k)}$ and $T_n^{(n-k)}$.

4 Accuracy

In this section we consider the accuracy of the new and previous approximation (approximation I and II). At the first step we consider the discrepancy between the obtained critical values by approximations I and II and the exact distribution for the cases $k = 1, 2$. Tables 1 and 2 show some critical values for $T_n^{(k)}$ at $\alpha = 0.01$ and $\alpha = 0.05$, respectively, which are obtained by different methods. The bold font indicates the better approximation method which produces the closest values to those which are obtained from the exact distribution. There are some interesting results, as is evident from Tables 1 and 2: (a) The values of the new and previous approximation methods are very close to the values of the exact distribution; (b) The new approximation method is better than the previous method. It must be emphasized also, that there is no scientific

Table 1. Critical values for $T_n^{(k)}$ at level $\alpha = 0.01$ for exact distribution

	n					
	5	10	20	30	50	100
$k=1$						
Exact	1.749	2.410	2.884	3.103	3.337	3.600
Approx I	1.751	2.413	2.886	3.106	3.340	3.650
Approx II	**1.749**	**2.410**	**2.883**	**3.102**	**3.335**	**3.600**
$k=2$						
Exact	2.160	3.402	4.437	4.946	5.497	6.118
Approx I	2.164	3.406	4.465	4.890	5.552	6.193
Approx II	**2.160**	**3.402**	**4.435**	**4.951**	**5.516**	**6.136**

Table 2. Critical values for $T_n^{(k)}$ at level $\alpha = 0.05$ for exact distribution

	n					
	5	10	20	30	50	100
$k = 1$						
Exact	1.671	2.176	2.557	2.745	2.945	3.207
Approx I	1.672	2.179	2.559	2.749	2.967	3.225
Approx II	**1.670**	**2.173**	**2.553**	**2.743**	**2.995**	**3.207**
$k = 2$						
Exact	2.010	3.197	4.110	4.651	5.058	5.638
Approx I	2.103	3.198	4.123	4.600	5.123	5.745
Approx II	**2.100**	**3.193**	**4.113**	**4.584**	**5.094**	**5.657**

Table 3. Critical values for $T_n^{(k)}$ at level $\alpha = 0.01$ by using simulation

	n				
	10	20	30	50	100
$k = 3$					
Simulation	3.997	5.612	6.431	7.329	8.388
Approx I	4.004	5.630	6.467	7.399	8.493
Approx II	**3.998**	**5.614**	**6.451**	**7.388**	**8.474**
$k = 4$					
Simulation	4.323	6.530	7.660	8.935	10.309
Approx I	4.331	6.544	7.717	9.052	10.621
Approx II	**4.323**	**6.529**	**7.700**	**9.035**	**10.599**

Table 4. Critical values for $T_n^{(k)}$ at level $\alpha = 0.05$ by using simulation

	n				
	10	20	30	50	100
$k = 3$					
Simulation	3.813	5.311	6.051	6.871	7.855
Approx I	3.818	5.321	6.111	6.999	8.056
Approx II	**3.814**	**5.314**	**6.100**	**6.992**	**8.044**
$k = 4$					
Simulation	4.155	6.249	7.235	8.408	9.772
Approx I	4.161	6.261	7.379	8.661	10.182
Approx II	**4.155**	**6.253**	**7.370**	**8.651**	**10.172**

discrepancy between the two approximations (I and II) for the small value of n, but for the large value of n the new approximation is better than the previous one; (c) The accuracy for the case $k = 1$ is more than for the case $k = 2$; because, for the case $k = 1$ the coefficient correlation is more closer to zero than for $k = 2$, and thus converges to zero faster; (d) The accuracy of the approximation methods reduce when the α increases, and the approximation I does not work for some value of the α.

At the second step, we consider the accuracy of the approximation methods by the critical values which are obtained for the cases $k = 3, 4$. Tables 3 and 4 show the results. For simulation part, we used S-Plus software and the number 10,000 for simulation. As it appears from Tables 3 and 4, the new approximation is again better than the previous approximation. Also it must be mentioned that conditions (a)-(d) hold for these tables as well.

Keywords. outlier; normal sample; likelihood ratio test; approximation.

References

Barnett, V.; Lewis, T. (1994). *Outliers in Statistical Data*, 3rd ed. Wiley, Chichester.

Thode, H.C. (2002). *Testing for Normality.* Marcel Dekker, New York.

Zhang, J.; Yu, k. (2006). The null distribution of the likelihood-ratio test for one or two outliers in a normal sample. *Test* **15**, 141-150.

Rahim Mahmoudvand
Statistics Department,
Payam-e Noor University of Toyserkan,
Toyserkan,
Iran.
e-mail: *r_mahmodvand@yahoo.com*

Hossein Hassani
Centeral Bank of the Islamic Republic of Iran,
207.1, Pasdaran Avenue,
Tehran,
Iran.

School of Mathematics,
Statistics Department,
Cardiff University,
Cardiff, UK.
e-mail: *hassanih@cf.ac.uk*

The full version of the paper, in Persian, appears on pages 167-184.