行政院國家科學委員會專題研究計畫　成果報告

指數排序統計量線性組合之比率的機率問題

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-032-002-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：淡江大學數學系

計畫主持人：林千代

計畫參與人員：吳正新，黃彥龍

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 92 年 11 月 5 日

Exact Distributions of Test Statistics for Multiple Outliers in Exponential Samples

Abstract

A computing algorithm developed by Huffer and Lin (2001) is employed to evaluate the exact null distributions of the test statistics proposed by Kimber (1982) for testing up to k upper outliers in exponential samples.

Keywords: Critical values; Outliers; Spacings.

1 Introduction

Early research on tests for outliers focused on Dixon-type test statistics for normal populations. Later on, tests for outliers in samples from exponential populations were discussed in detail in the literature because of the important role that exponential distribution plays in the areas of life-testing and reliability; see, for example, the book on the exponential distribution by Balakrishnan and Basu (1995). These developments on outliers for the exponential distribution include the works of Epstein (1960), Laurent (1963), Basu (1965), Likeš (1966), Kabe (1970), and Mount and Kale (1973). It should be mentioned that Likeš (1966) obtained the distribution of Dixon's statistic in (disguised) beta function forms, and then tabulated some selected percentage points of these distributions for some cases. Kabe (1970) explicitly expressed these distributions in terms of finite series of beta functions and mentioned that the work of finding the percentage points of these distributions is formidable.

Some other work focused on statistics of the form [see Lewis and Fieller (1979), Kimber and Stevens (1981), Chikkagoudar and Kunchur (1983) and Barnett and Lewis (1984, 1993)]

$$T_k = \frac{X_{(n-k+1)} + \cdots + X_{(n)}}{\sum_{i=1}^{n} X_{(i)}},$$

where $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of a random sample $X_1, \ldots, X_n$ from an exponential distribution with cdf

$$F(x;\theta) = 1 - \exp(-\theta x),\ x > 0,\ \theta > 0. \quad (1)$$

The block test statistics of this kind introduced by Barnett and Lewis (1978) involves specifying k, the number of outliers thought to be present in the data, and then either zero or k outliers are declared discordant on the basis of a suitable hypothesis test. The other approach for testing the discordancy is the sequential or consecutive procedure which applies a single-outlier procedure repeatedly, starting with the most extreme observation, deleting the discordant value at each stage and applying the test again to the reduced sample. This process continues on until a non-significant result is obtained. Rosner (1975) proposed another sequential procedure which avoids the shortcomings of the above method in routinely screening the Gaussian samples. Rosner's method applies

to the reduced data as above except that the test statistic is calculated for the reduced data a fixed number of times, k, and that the k-th most extreme outlier is treated first; if this gives a significant result, then k outliers are declared discordant; if a non-significant result is obtained, then $(k-1)$-th most extreme outlier is tested. This process is continued until either a significant result is obtained or no outliers can be declared discordant. Kimber (1982) adopted Rosner's procedure to propose a test to detect the discordancy of up to k outliers in exponential samples, and then derived the exact null distributions of the test statistics. Because the computation of these distributions involves intricate numerical integration, only certain critical values for $k = 2, 3$ and 4 were tabulated and presented.

In this paper, we utilize the algorithm of Huffer and Lin (2001) to evaluate the exact null probabilities and the exact critical values of Kimber's test statistics. A similar procedure for testing up to k lower outliers is also discussed. Finally, we present two examples to illustrate the exact method of computation developed here. Our method can be easily extended to the corresponding consecutive test statistics [see Chikkagoudar and Kunchur (1987)] based on the Dixon-type statistics

$$D_j = \frac{X_{(n-j+1)} - X_{(n-j)}}{X_{(n-j+1)}},\ j = 1, \ldots, k.$$

2 Kimber's Sequential Procedure

Assume that $X_{(n-j+1)}, \ldots, X_{(n)}$ are suspected to be outliers and have come from an exponential distribution $F(x; \theta\lambda)$. Then an appropriate procedure is to test the null hypothesis

$$H_0 : \lambda = 1,$$

against the *labelled slippage* alternative

$$H_1 : 0 < \lambda < 1.$$

The likelihood ratio test statistics in this case are of the form

$$S_j = \frac{X_{(n-j+1)}}{\sum_{i=1}^{n-j+1} X_{(i)}},\ j = 1, \ldots, n-1. \qquad (2)$$

For a given significance level α and a maximum number of contaminants k, the critical values s_j for $j = 1, \ldots, k$ are determined such that

$$P\left\{ \bigcap_{j=1}^{k} (S_j < s_j)\,; H_0 \right\} = 1 - \alpha$$

and

$$\begin{aligned} P(S_k > s_k; H_0) &= P(S_{k-1} > s_{k-1}; H_0) \\ &= \cdots = P(S_1 > s_1; H_0) = \beta, \text{ say.} \end{aligned}$$

Then, a size α test for up to k outliers proceeds as follows:

(i) if $S_k > s_k$, declare the k largest observations discordant;

(ii) if $S_i < s_i$ for $i = k, k-1, \ldots, j+1$ and $S_j > s_j$, declare the j largest observations discordant $(j = k-1, k-2, \ldots, 1)$;

(iii) if $S_i < s_i$ for $i = 1, \ldots, k$, declare no observations discordant.

3 Computational Algorithm

Denote the spacings by

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)}),\ i = 1, \ldots, n.$$

It is well known that, in the case of an exponential sample, the spacings D_1, D_2, ...,

D_n are all independent and identically distributed (i.i.d.) as exponential with parameter θ; see, for example, Arnold, Balakrishnan and Nagaraja (1992, pp. 72–73). This particular distributional property of spacings and the algorithm of Huffer and Lin (2001) described below have been utilized effectively here to compute the exact null distributions of the test statistics S_j in (2). We now describe Huffer and Lin's algorithm and relate its application to the outlier testing procedure considered here.

Suppose $Z_1, Z_2, \ldots, Z_n$ are i.i.d. exponential random variables with mean 1. Let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)'$. Given the level of the significance α under the null hypothesis and maximum number of contaminants k, the probabilities we are interested in are

$$\begin{aligned} & P(S_j > s_j) \\ = \; & P\left(\frac{X_{(n-j+1)}}{\sum_{i=1}^{n-j+1} X_{(i)}} > s_j \right) \\ = \; & P\left(\frac{\sum_{i=1}^{n-j+1} \frac{1}{n-i+1} D_{(i)}}{\sum_{i=1}^{n-j+1} \frac{n-j-i+2}{n-i+1} D_{(i)}} > s_j \right) \\ = \; & P\left(\sum_{i=1}^{n-j+1} \frac{1}{n-i+1} Z_i \right. \\ & \left. - \sum_{i=1}^{n-j+1} \frac{n-j-i+2}{n-i+1} s_j Z_i > 0 \right) \end{aligned} \quad (3)$$

for $j = 1, \ldots, k$. Thus, they are simply probabilities involving linear combinations of i.i.d. exponential random variables with arbitrary rational coefficients of the form

$$P(\boldsymbol{a}\boldsymbol{Z} > 0), \quad (4)$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)$ and a_i's are rational values. The recursion given in (6) below, which simplifies the general algorithm proposed by Huffer (1988), Lin (1993) and Huffer and Lin (1999, 2001), can help us to find s_j such that the probabilities

$$P(S_j > s_j) = \beta = \alpha/k. \quad (5)$$

Result 1 *Let $\boldsymbol{a}$ be an arbitrary row vector. Let n be the number of entries of $\boldsymbol{a}$. For any value x, we define $\boldsymbol{a}_{i,x}$ to be the vector obtained by replacing the i^{th} entry of $\boldsymbol{a}$ by x. Let $\boldsymbol{c} = (c_1, \ldots, c_n)'$ be any $n \times 1$ vector satisfying $\sum_{i=1}^n c_i = 1$, and let $\xi = \boldsymbol{a}\boldsymbol{c}$. Then*

$$P(\boldsymbol{a}\boldsymbol{Z} > 0) = \sum_{i=1}^{n} c_i P(\boldsymbol{a}_{i,\xi}\boldsymbol{Z} > 0). \quad (6)$$

A simple example is given below to illustrate the use of this algorithm. Consider a sample with size $n = 3$ and maximum number of contaminants $k = 2$ in (1). Then, the probabilities in (3) for $j = 1$ have the form in (4) with the vector $\boldsymbol{a}$ given by

$$\left(\frac{1}{3} - s_1, \frac{1}{2} - s_1, 1 - s_1 \right).$$

Using the expression (6) in Recursion 1, we then obtain the probability

$$\begin{aligned} & P(S_1 > s_1) \\ = \; & P[(\frac{1}{3} - s_1)Z_1 + (\frac{1}{2} - s_1)Z_2 \\ & +(1 - s_1)Z_3 > 0] \\ = \; & (3 - 6s_1)\{(2 - 2s_1)P[(1 - s_1)Z_1 > 0] \\ & +(-1 + 2s_1)P[(\frac{1}{2} - s_1)Z_1 > 0]\} \\ & +(-2 + 6s_1)\left\{ \left(\frac{3}{2} - \frac{3s_1}{2} \right) \right. \\ & \qquad P[(1 - s_1)Z_1 > 0] \\ & \left. + \left(-\frac{1}{2} + \frac{3s_1}{2} \right) P[(\frac{1}{3} - s_1)Z_1 > 0] \right\} \end{aligned}$$

as a function of s_1. Examining the value of s_1 between the intervals $(0, 1/3)$, $(1/3, 1/2)$, and $(1/2, 1)$, we obtain the exact 2.5% critical value for S_1 to be 0.90871.

4 Examples

We consider two examples given by Kimber and Stevens (1981) and Barnett and Lewis (1984, Example 6.1).

Example 1: The first example is a set of $n = 21$ observations:

$$25, 5, 7, 61, 446, 34, 87, 76, 4, 17, 19,$$
$$240, 116, 45, 64, 141, 31, 503, 10, 181, 101.$$

The observed test statistics S_j $(j = 1, 2, 3, 4)$ and the corresponding p-values evaluated by the approach detailed in the last section are as follows:

$$\begin{aligned} 503/2260 &= 0.22257, \\ P(S_1 > 0.22257) &= 0.13500, \\ 446/1757 &= 0.25384, \\ P(S_2 > 0.25384) &= 0.01090, \\ 240/1311 &= 0.18307, \\ P(S_3 > 0.18307) &= 0.10958, \\ 181/1071 &= 0.16900, \\ P(S_4 > 0.16900) &= 0.17548. \end{aligned}$$

From these values, we can suggest that the two largest values, 446 and 503, may be outliers at 5% significance level when using the sequential procedure with either $k = 2, 3$, or 4. The exact overall 5% critical values s_j for the sequential procedure when $n = 21$ are $s_2 = 0.23308$ and $s_1 = 0.28584$ with $k = 2$, $s_3 = 0.22463, s_2 = 0.24327$ and $s_1 = 0.30018$ with $k = 3$, and $s_4 = 0.22374, s_3 = 0.23076, s_2 = 0.25044$ and $s_1 = 0.31018$ with $k = 4$, respectively. The inequality $0.01202 < P(S_2 > 0.2538) < 0.01204$ given by Kimber (1982) is away from the exact probability of 0.01092, which indicates that some computational rounding errors may have occurred in Kimber's (1982) method.

References

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*, New York: John Wiley & Sons.

Balakrishnan, N. and Basu, A. P. (Eds.) (1995). *The Exponential Distribution: Theory, Methods and Applications*, Newark: Gordon and Breach Science Publishers.

Barnett, V. and Lewis, T. (1978, 1984, 1993). *Outliers in Statistical Data*, First, Second and Third editions, Chichester: John Wiley & Sons.

Basu, A. P. (1965). On some tests of hypotheses relating to the exponential distribution when some outliers are present. *J. Amer. Statist. Assoc.* **60**, 548–559.

Chikkagoudar, M. S. and Kunchur, S. H. (1983). Distributions of test statistics for multiple outliers in exponential samples. *Commun. Statist. –Theory Meth.* **12**, 2127–2142.

Chikkagoudar, M. S. and Kunchur, S. H. (1987). Comparison of many outliers procedures for exponential samples. *Commun. Statist. –Theory Meth.* **16**, 627–645.

Epstein, B. (1960). Tests for the validity of the assumption that the underlying distribution of life is exponential: Part II. *Technometrics* **2**, 167–183.

Huffer, F. W. and Lin, C. T. (2001). Computing the joint distribution of general linear combinations of spacings or exponential variates. *Statistica Sinica* **11**, 1141–1157.

Kabe, D. G. (1970). Testing outliers from an exponential population. *Metrika* **15**, 15–18.

Kimber, A. C. (1982). Tests for many outliers

in an exponential sample. *Appl. Statist.* **31**, 263–271.

Kimber, A. C. and Stevens, H. J. (1981). The null distribution of a test for two upper outliers in an exponential sample. *Appl. Statist.* **30**, 153–157.

Laurent, A. G. (1963). Conditional distribution of order statistics and distribution of the reduced i-th order statistic of the exponential model. *Ann. Math. Statist.* **34**, 652–657.

Lewis, T. and Fieller, N. R. J. (1979). A recursive algorithm for null distribution for outliers: I. Gamma samples. *Technometrics* **21**, 371–376.

Likeš, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika* **11**, 46–54.

Mount, K. S. and Kale, B. K. (1973). On selecting a spurious observation. *Cand. Math. Bull.* **16**, 75–78.

Rosner, B. (1975). On the detection of many outliers. *Technometrics* **19**, 307–312.