第13卷 第2期 郑州工学院学报 Vol.13 No.2
1992年 6 月 Journal of Zhengzhou Institute of Technology Jun. 1992

正态样本多个异常值检测的一类半极差型方法*

张新育 刘江国

(郑州工学院数力系)

摘 要: 本文在两种情况下提出了正态样本多个异常值检测的一类半极差型方法，并对这类方法进行了总结.

关键词: 正态样本，异常值，半极差型方法

中国图书分类号: TB114

在正常样本异常值检测中常用的方法之一是半极差型方法，这类方法在单个异常值的检测中功效是高的，但在多个异常值的检测中由于异常值对参数估计有影响，因此"遮蔽"作用相当严重，见〔1〕PP3～13。本文对已知 u 未知 σ 及未知 u 已知 σ 两种情形进行了改进。对 u、σ 均未知的情形〔4〕给出了改进，本文只列出结论。最后我们列出了异常值分布的不对称性对检测结果的影响情况。本文的方法在很大程度上降低了"遮蔽"作用，提高了半极差型方法的稳健性，计算也比较简单。

1 方法的改进与实例

设 $x_1, x_2, \cdots, x_n \quad iid\sim N(u, \sigma^2)$，$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ 是对应的次序统计量，就单测检验而言，对 $x_{(n)}$ 或 $x_{(1)}$ 半级差型检验的拒绝域有如下形式:

①已知u，σ，$\dfrac{x_{(n)}-u}{\sigma} \geqslant c_n$，$\dfrac{u-x_{(1)}}{\sigma} \geqslant c_n$;

②已知u，未知σ，$\dfrac{x_{(n)}-u}{s(u)} \geqslant bn$，$\dfrac{u-x_{(1)}}{s(u)} \geqslant bn$;

③已知σ，未知u，$\dfrac{x_{(n)}-\bar{x}}{\sigma} \geqslant Nn$，$\dfrac{\bar{x}-x_{(1)}}{\sigma} \geqslant N_n'$;

④已知u，σ，$\dfrac{(x_{(n)}-\bar{x})}{s} \geqslant gn$，$\dfrac{(x-x_{(1)})}{s} \geqslant g_n'$;

其中$s(u) = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-u)^2}$，$s = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$，

* 收稿日期: 1991-03-07

当n较大时$N_n' \approx N_n$，$g_n' \approx g_n$。双侧检验的拒绝域暂不考虑。

设 $W=\frac{1}{n-2[np]}\sum_{i=[np]+1}^{n-[np]} x_{(i)}$，$M=\frac{1}{n-2[np]}\sum_{i=[np]+1}^{n-[np]} x_{(i)}^2$，$\sigma^*=\sqrt{M-W^2}$，其中$o<p<\frac{1}{2}$，对（2）、（3）、（4）改进后得如下拒绝域：

②′已知u，未知σ，$\frac{x_{(n)}-u}{\sigma^*}\geqslant Bn$，$\frac{u-x_{(1)}}{\sigma^*}\geqslant Bn$；

③′已知σ，未知u，$\frac{x_{(n)}-W}{\sigma}\geqslant n_n'$，$\frac{W-x_{(1)}}{\sigma}\geqslant n_n'$；

④′未知u，σ，$\frac{x_{(n)}-W}{\sigma^*}\geqslant Gn$，$\frac{W-x_{(1)}}{\sigma^*}\geqslant G_n'$；

当 n 较大时，$n_n' \approx n_n$，$G_n' \approx G_n$。本文末列出了 $p=\frac{1}{8}$ 时，B_n，n_n，G_n 的临介值表。其中（4）′在〔4〕中已提了出来，为了完整本文也列了出来。

在使用此类方法剔除多个异常值时，应采用逐个剔除的方法，即每次都要考虑 W，σ^* 的值，当样本容量较大而异常值不多时 W，σ^* 的值变化不大，可以多次使用。因 $p=\frac{1}{8}$，故一般应有单侧异常值个数 $\leqslant\frac{1}{8}n$

〔例 1〕 某批化纤纤维干收缩率的 25 个观侧值为：3.49，3.49，4.01，4.48，4.61，4.76，4.98，5.25，5.32，5.39，5.42，5.57 5.59，5.59，5.63，5.63，5.65，5.66，5.67，5.69，5.71，6.00，6.03，6.12，6.76（单位%），已知该化纤纤维干收缩率服从 N（u，0.625^2），试检测出该批产品中的所有异常值。

先利用（3），即所谓的 *Nair* 检测法，$\bar{x}=5.3$，$\frac{x_{(25)}-\bar{x}}{\sigma}=\frac{6.76-5.3}{0.65}$

≈ 2.2462，$\frac{(\bar{x}-x_{(1)})}{\sigma}=\frac{5.3-3.4}{0.65}=2.785$ 查〔1〕*pp29Nair* 检测临界值表得 $n=25$ 时，水平 $\alpha=0.05$ 的临界值为 2.815，故 $x_{(1)}$，$x_{(25)}$ 都不是水平为 0.05 的异常值。

再利用本文改进的 *Nair* 方法检测，$n=2$，$p=\frac{1}{8}$，$[np]=\frac{1}{3}$，$W=5.4053$，$\frac{x_{(25)}-W}{\sigma}=2.084$，$\frac{W-x_{(1)}}{\sigma}=2.947$，查本文附表（2），$n=25$ 时，水平 $\alpha=0.01$ 的临界值为 2.29，故 $x_{(1)}$，$x_{(2)}$，$x_{(25)}$ 均为水平 $\alpha=0.01$ 的异常值，可见改进的 *Nair* 检测稳健性较好；在检测中异常值之间的相互影响较小，在多个异常值的检测中新方法有很大的实用价值。

下面举一使用改进方法(2)′的例子，并与有一定稳健性的*Dixon*检测进行比较。

〔例 2〕 某批维尼纶纤维纤度的 10 个测量值为 1.12，1.32，1.35，1.36，1.38，1.39，1.40，1.44，1.55，1.80，已知该种纤维纤度服从 N（1.405，σ * T2 ），试检测出其中上、下侧极端值是否异常。

先利用 *Dixon* 方法，$n=10$，$Y_{11}=\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(2)}}=\frac{1.80-1.55}{1.80-1.32}=0.521$；$n=9$，$r'_{11}=\frac{x_{(2)}-x_{(1)}}{x_{(n-1)}-x_{(1)}}=\frac{1.32-1.12}{1.44-1.12}=0.625$，查〔1〕*P36Dixon* 检测临界值表知，$n=10$，水平为 0.05 的临界值为 0.477，因为 $0.521>0.477$，故 1.85 是水平为 0.05 的异常值，查〔1〕*P36dixon* 检测临界值表，$n=9$，水平为 0.05 的临介值为 0.512，$0.625>0.512$，故 1.07 为水平为 0.05 的异常值。

再用改进的方法 ②′ 检测，$p=\frac{1}{8}$，$n=10$，$[np]=1$，

$$W=\frac{1}{n-2[np]}\sum_{i=[np]+1}^{n-[np]}x_{(i)}=\frac{1}{8}\sum_{i=2}^{9}x_{(i)}=1.399,\quad M=\frac{1}{8}\sum_{i=2}^{9}x^2_{(i)}=1.967$$

$\sigma^*=\sqrt{M-W^2}=0.0662$，$\frac{(x_{(10)}-u)}{\sigma^*}=\frac{1.80-1.405}{0.0662}=5.96$。查本文附表（1），$n=10$，$\alpha=0.01$的临界值为5.299，故1.80是水平为0.01的异常值。剔除后对$n=9$，$p=\frac{1}{8}$，$[np]=1$，$W=\frac{1}{7}\sum_{i=1}^{8}x_i\approx 1.377$，$W=\frac{1}{7}\sum_{i=2}^{8}x^2_{(i)}\approx 1.898$，$\sigma^*=\sqrt{M-W^2}\approx 0.0357$，$\frac{u-x_{(1)}}{\sigma^*}=\frac{1.405-1.12}{0.0357}\approx 7.98$，查附表（2），$n=9$，$\alpha=0.01$ 的临界值为 5.202，$7.98>5.202$，故 1.12 是水平为 0.01 的异常

此例说明②′可以与*Dixon*检测相互补充使用，②′充分使用了期望值u的信息。

以上方法 ②′③′④′ 使用了 u，σ 的改进矩法估计量 W，σ^* 由 [3] 知 W,σ^* 分别是 u，$\lambda\sigma$ 的强相合估计〔2〕，$\lambda>0$ 是一常数。故样本容量 n 越大，此类方法越精确，且不论 n 大小，此类方法均呈现出很好的稳健性。

2　对改进的半极差型方法的几点说明

2.1　W，σ^* 的计算方法

设 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$ 为样本 $x_1,x_2\cdots x_3$ 的顺序统计量，则 W，σ^* 分别是 $x_{([np]+1)}$，$x_{([np]+2)}$，…，$x_{(n-[np])}$ 的均值和方差。

2.2　估计量 W，σ^* 受异常值影响的定性分析

下述事件发生的概率较大 (1) 当异常值分布在顺序统计量上、下侧的个数相等时，估计量 W，σ^* 基本不受异常值的影响。(2) 当上侧异常值较多时，W,σ^* 分别是 $u\lambda\sigma$ 的过高估计。(3) 当下侧异常值较多时，W 是 u 的过低估计，σ^* 是 $\lambda\sigma$ 的过高估计。

2.3　使用②′ ③′ ④′ 时应注意的问题

由②的分析可知：当异常值在顺序统计量上、下侧存在的个数基本相等时，方法②′ ③′ ④′ 基本可以消除遮蔽现象。当异常值存在于上、下侧的个数不等时，方法②′ ③′

④′有一定的遮蔽作用，检出的异常值为真正异常值与分布对称情况对应概率比较如下：

P	方法					
	②′		③′		④′	
异常值所在的侧	上侧	下侧	上侧	下侧	上侧	下侧
上侧异常值较多	较大	较大	较大	较小	较大	不定
下侧异常值较多	较大	较大	较小	较大	不定	较大

3 关于用计算机模拟随机数的说明

为了检验本文提出的方法，我们先后在IBM—PC机和Wang VS300机上进行了计算机模拟产生随机数及计算临界值的工作，并对程序运行情况及所得数值结果进行了比较。模拟产生随机数是通过随机函数实现的，相比较而言，VS300机产生随机数的周期较PC机长，但效果要好一些。严格说来，所用的随机数是伪随机函数，但从计算结果看来，这种模拟方法还是有一定的有效性。如能采用硬件发生器产生随机数，效果将更好。对本文中的问题，给定参数 n 的值后，通过随机函数产生一批服从正态分布的随机数 x_1，x_2，…，x_n，为使其成为次序统计量，采用排序方法对这批随机数进行排序，使得 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$。据此计算出 W，M，和 σ^* 以及检验统计量的值 g_n $= \dfrac{x_{(n)} - W}{\sigma^*}$（或 $g_n = \dfrac{x_{(n)} - u}{\sigma^*}$ 或 $g_n = \dfrac{x_{(n)} - W}{\sigma^*}$）。对 n 的同一个值进行多次随机模拟，求得多个 g_n，再对所有的 g_n 进行排序，求得临界值。本文末给出了 $p = \dfrac{1}{8}$ 时的参数 N 取值3～100的临界值（在 $Wang\ \ VS$—300机上的计算结果）

附表（1）B_n的临界值表($p=1/8$)

N	显著性水平			N	显著性水平			N	显著性水平			N	显著性水平		
	0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01
				11	4.705	4.925	5.273	21	4.952	5.070	5.302	50	5.173	5.309	5.453
				12	4.789	5.018	5.349	22	4.966	5.136	5.369	100	5.194	5.325	5.429
				13	4.833	5.063	5.342	23	5.001	5.174	5.402				
				14	4.862	5.050	5.386	24	5.007	5.147	5.411				
				15	4.884	5.085	5.309	25	4.964	5.138	5.350				
				16	4.938	5.091	5.419	26	4.903	5.080	5.297				
				17	4.913	5.112	5.391	27	4.998	5.100	5.319				
8	4.693	4.940	5.227	18	5.007	5.195	5.537	28	4.946	5.135	5.253				
9	4.772	4.881	5.202	19	4.974	5.171	5.447	29	4.941	5.082	5.343				
10	4.708	4.979	5.299	20	4.990	5.169	5.400	30	4.960	5.102	5.340				

附表(2): N_e 的临界值表 (P=1/8)

N	显著性水平			N	显著性水平			N	显著性水平			N	显著性水平		
	0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01
				11	1.439	1.623	1.947	21	1.475	1.563	1.803	31	1.484	1.551	1.732
				12	1.385	1.565	2.126	22	1.543	1.672	2.061	32	1.673	1.763	1.940
				13	1.304	1.419	1.687	23	1.433	1.527	1.719	50	1.672	1.750	1.896
				14	1.230	1.327	1.505	24	1.738	1.874	2.189				
				15	1.363	1.518	1.897	25	1.681	1.781	2.207				
				16	1.773	1.929	2.205	26	1.633	1.723	1.897				
				17	1.648	1.832	2.074	27	1.613	1.734	1.957				
8	1.794	1.956	2.644	18	1.618	1.759	2.112	28	1.565	1.649	1.824				
9	1.560	1.730	2.248	19	1.639	1.833	2.344	29	1.574	1.707	2.097				
10	1.421	1.589	1.867	20	1.496	1.610	1.859	30	1.605	1.726	2.105				

附表 (3): g_e 的临界值表 (p=1/8)

N	显著性水平			N	显著性水平			N	显著性水平			N	显著性水平		
	0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01
				11	2.601	2.787	3.063	21	2.638	2.739	2.902	31	2.706	2.774	2.924
				12	2.622	2.841	3.441	22	2.639	2.755	2.850	32	3.083	3.205	3.693
3	1.148	1.153	1.155	13	2.451	2.589	2.849	23	2.601	2.728	2.943	33	2.918	3.022	3.198
4	1.425	1.463	1.492	14	2.431	2.543	2.817	24	2.901	2.911	3.264	34	2.944	3.059	3.279
5	1.602	1.672	1.749	15	2.439	2.559	2.785	25	2.945	3.084	3.712	35	3.012	3.192	3.535
6	1.729	1.822	1.944	16	2.908	3.030	3.309	26	2.901	3.036	3.426	80	3.093	3.192	3.457
7	1.828	2.032	2.221	17	2.952	3.133	3.800	27	2.779	2.897	3.025	90	3.267	3.405	3.624
8	2.809	3.029	3.392	18	2.798	2.907	3.130	28	2.765	2.860	2.979	100	3.026	3.115	3.382
9	2.669	2.881	3.159	19	2.716	2.879	3.059	29	2.743	2.833	2.974				
10	2.695	2.895	3.614	20	2.691	2.825	3.044	30	2.724	2.838	3.204				

参 考 文 献

〔1〕 异常值处理，89年华东师大数理统计系研究生讲义

〔2〕 张新育.改进的分布参数 σ，u 的矩法估计量及其应用.应用概率统计（待发表）

〔3〕 朱宏.用样本分位数方法同时检测正态样本多个异常值.数理统计与应用概率，Vol.4，No.1，

〔4〕 成平、陈希孺著，参数估计，上海科技出版社，1985

[5] 中科院计算中心编著,概率统计计算,1979.4

An Extreme-Mean Difference Method for the Rejecting of Outliers of Normal Samples

Zhang Xinyu,LiuJiangguo

(Dept of Maths & Mechanics)

Abstract: In this paper, we present an extreme-mean difference method for the rejecting of outliers of normal samples in two cases.

Keywords: Normal samples, Outliers, Extreme-Mean difference method.