DOI:10.16016/j.1000-5404.1981.03.023

讲 座

科研中偏离过大数值的取舍

卫生系流行病学教研室 涂 瀛

在科学实验或调查研究中，有时出现个别特大或特小的数值，与均数相差很大；或大多数数值的出现有一定规律，个别数值偏离过大，以致对平均数和标准差影响很大或难以配好回归曲线。有的人随意舍弃此种“异常”数值，企图使实验结果更趋一致或使曲线更为平滑，这是不恰当的；也有人不管所得数值是否可靠或异常，毫不怀疑地一并进行统计分析也是不对的。一个异常数值可对结果影响很大，甚至得出相反的结论。例如曾有人研究生花生和炒花生的蛋白质生理价值，若根据统计学原理舍弃一个特小的数值，则两者差别显著；否则差别不显著，而舍弃后得出的结论是符合营养学原理的，故对偏离过大数值的取舍是一值得研究的问题。现将偏离过大数值出现的可能原因及其取舍办法简介如下：

一、偏离过大数值出现的原因

（一）固有差异

宇宙间一切事物无刻不在运动、发展和变化，而这种运动、发展和变化彼此是不平衡的，故事物彼此之间普遍存在着差异，这种差异是固有的、正常的。例如各人的身高长短、身体轻重、血球多少和血压高低等都不相同，有时相差较大。出现的偏离过大的数值也可能就是这种事物中的极大或极小值，反映了客观事物本来的面貌，是难免的，不应随便舍弃。

（二）特殊原因

有时偏离过大数值并不是固有差异所致，而是有特殊原因的。如：1.实验过程中，条件有大的改变或条件稍有改变引起原来不起作用的因子变为重要因子，如仪器偶然发生故障或灵敏度改变因而使所得数值偏离过大，不过我们尚未觉察出实验条件已经改变而已；2.异常数值来自其他总体，和所研究的总体有本质不同；在调查时也可能有张冠李戴的情况；3.实验或调查中发生过失误差，如在称样品时游标尺反读，砝码加错，取样时吸管拿错或刻度看错；操作中试剂加错或忘加；样品被污染或部份丢失，或由于粗枝大叶、过度疲劳以致操作不当，观察不准，判断不确，记录有错或计算有误等。

二、偏离过大数值取舍办法

判断这种“异常”数值可否取舍，首先应仔细回忆整个科研过程并检查全部记录，若有充分理由说明这种数值的出现是由于条件改变、操作错误或其他过失引起，则可根据重复实验的结果改正或舍弃该数值，否则不应凭主观臆断随意更改或舍弃。但可能根据概率理论来决定取舍，因为若偏离过大数值与其他数值的差别是由于固有差异，没有特殊原因的话，则此种差别不会太大，其出现的机会不会太小，就不应舍弃。反之，若偏离过大的数值与其他

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

数值相差很大，出现的机会一定很小，则该数值与其他数值的差别由于固有差异所致的可能性就不大，也就是说由特殊原因引起的可能性较大，则可将此数值舍弃。

早在1838年Bessel就曾考虑过舍弃偏离过大数值的办法，1852年Peirce曾试图用概率来判断取舍异常数值。自后1863年Chauvenet、1884年Wright、1927年Student、1935年Thompson、1950年Dixon、1950年及1959年Grubbs等都提出了异常数值舍弃的方法和标准。大都是以检验可疑数值是否偏离正态为基础，但偏离到什么程度方可舍弃，则是主观人为的，各家不一，所用方法繁简也不同。现将常用者简介如下：

（一）一个样本中只有一个可疑数值时

设一个随机样本含有n个观察值（X），由小到大排列为$X_1 \leqslant X_2 \leqslant \cdots\cdots \leqslant X_n$。如何用统计方法判断$X_1$或$X_n$是否异常数值？

例1：测定生花生蛋白质生理价值为44，56，56，56，59，60，61，61，63，63，试判断44是否异常数值。

1．简便法

（1）除去偏离过大的那个数值外，计算其余数值的算术平均数

$$\bar{X} = \left(\sum_{i=2}^{n} Xi\right) \Big/ \quad (n-1) \qquad (若X_1为可疑数值)$$

$$= \left(\sum_{i=1}^{n-1} Xi\right) \Big/ \quad (n-1) \qquad (若X_n为可疑数值)$$

（2）计算平均差$\delta = \left(\sum_{i=2}^{n} |Xi - \bar{X}|\right) \Big/ \quad (n-1) \qquad (若X_1为可疑数值)$

$$= \left(\sum_{i=1}^{n-1} |Xi - \bar{X}|\right) \Big/ \quad (n-1) \qquad (若X_n为可疑数值)$$

（3）若偏离过大数值（X_1或X_n）与算术平均数之差，等于或大于4倍平均差，

$|X_1 - \bar{X}| \geq 4\delta$或$|Xn - \bar{X}| \geq 4\delta$则可舍弃$X_1$或$X_n$。

例1：$\bar{X} = 535/9 = 59.5$，$\delta = 21.5/9 = 2.4$

$|44 - 59.5|/2.4 = 15.5/2.4 = 6.5$ 故44可被舍弃。

2．肖维纳（Chauvenet）法

（1）包括偏离过大的数值在内，计算算术平均数（$\bar{X}$）和标准差（S）。

$$\bar{X} = \left(\sum_{i=1}^{n} Xi\right) \Big/ n \qquad S = \sqrt{\left(\sum_{i=1}^{n} (Xi - \bar{X})^2\right)/(n-1)}$$

（2）计算偏离过大数值与算术平均数之差，及此差相当于标准差之倍数即

$To = |X - \bar{X}| / S$。

（3）根据n值从表一找出相应的T_c值，若T_o T_c，则此偏离过大数值可舍弃之。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表一 肖维纳数值舍弃标准

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tc	1.15	1.38	1.53	1.65	1.73	1.79	1.86	1.92	1.96	2.00	2.01	2.07	2.10	2.13

n	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	35
Tc	2.16	2.18	2.20	2.22	2.24	2.26	2.28	2.30	2.32	2.33	2.34	2.35	2.37	2.38	2.39	2.45

n	40	50	60	80	100	150	185	200	250	300	400	500	600	1000	2000	5000
Tc	2.50	2.58	2.64	2.74	2.81	2.93	3.00	3.02	3.11	3.14	3.23	3.29	3.34	3.48	3.66	3.89

例1：$\bar{X}=57.9$，$S=5.59$　$T_0=|44-57.9|/5.59=2.49$

查表一：当n=10时，$T_c=1.96<T_0=2.49$，

故44这一数可舍弃之。

3．格拉布斯（Grubbs）法

计算方法和步骤与肖维纳法相同，不过其数值舍弃标准不同，并提出了危险率α（α是实际不是异常数据而判为异常数据的机率）为5%、2.5%及1%的数值舍弃标准，如表二：

表二 格拉布斯（Grubbs）数值舍弃标准（单侧）

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	1.15	1.46	1.67	1.82	1.94	2.03	2.11	2.18	2.23	2.29	2.33	2.37	2.41	2.44	2.47	2.50
Tg	1.15	1.48	1.71	1.89	2.02	2.13	2.21	2.29	2.36	2.41	2.46	2.51	2.55	2.59	2.62	2.65
	1.15	1.49	1.75	1.94	2.10	2.22	2.32	2.41	2.48	2.55	2.61	2.66	2.71	2.75	2.79	2.82

n	19	20	21	22	23	24	25	30	40	45	50	60	70	80	90	100
	2.53	2.56	2.58	2.60	2.62	2.64	2.66	2.75	2.87	2.92	2.96	3.03	3.09	3.14	3.18	3.21
Tg	2.68	2.71	2.73	2.76	2.78	2.80	2.82	2.91	3.04	3.09	3.13	3.20	3.26	3.31	3.35	3.38
	2.85	2.88	2.91	2.94	2.96	2.99	3.01									

注　Tg：上行：α为5%，中行：α为2.5%；下行：α为1%，

例1：$\bar{X}=57.9$，$S=5.58$，$T_0=2.49$。

查表一：当n=10时，$T_0>Tg$，0.01=2.41。

故舍弃44这一数值错误的概念不到1%。

4．狄克逊（Dixon）法

（1）计算极差（range）和次极差（subrange）的比值（r）。

① 当$n\leq 7$　最小值可疑时，$r_{10}=(X_2-X_1)/(X_n-X_1)$

最大值可疑时，$r_{10}=(X_n-X_{n-1})/(X_n-X_1)$

② 当$n=8\sim10$　最小值可疑时，$r_{11}=(x_2-x_1)/(X_{n-1}-X_1)$

最大值可疑时，$r_{11}=(X_n-X_{n-1})/(X_n-X_2)$

③ 当$n=11\sim13$　最小值可疑时，$r_{21}=(X_3-X_1)/(X_{n-1}-X_1)$

最大值可疑时，$r_{21}=(X_n-X_{n-2})/(X_n-X_2)$

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

④ 当n 14 最小值可疑时，$r_{22} = (X_3 - X_1)/(X_{n-2} - X_1)$

最大值可疑值，$r_{22} = (X_n - X_{n-2})/(X_n - X_3)$

（2）查表三：根据n值找出相应r临界值，若计算所得之r值大于1%或5%危险率的r临界值，则可舍弃之。

表三 狄克逊 (Dixon) 检验的 r 值的临界值

n	3	4	5	6	7	8	9	10	11	12	13	14
	0.886	0.679	0.557	0.482	0.434	0.479	0.441	0.409	0.517	0.490	0.467	0.492
r	0.941	0.765	0.642	0.560	0.507	0.554	0.512	0.477	0.576	0.546	0.521	0.546
	0.988	0.889	0.780	0.698	0.736	0.683	0.635	0.597	0.679	0.642	0.615	0.641

n	15	16	17	18	19	20	21	22	23	24	25
	0.472	0.454	0.438	0.424	0.412	0.401	0.391	0.382	0.374	0.367	0.360
r	0.525	0.507	0.490	0.475	0.462	0.450	0.440	0.430	0.421	0.413	0.406
	0.616	0.595	0.577	0.561	0.547	0.535	0.524	0.514	0.505	0.497	0.489

注：γ值：上行α为10%，中行α为5%，下行α为1%。

例1：n=10，$X_1=44$，$X_2=56$，$X_{n-1}=63$

则 $r_{11} = (X_2 - X_1)/(X_{n-1} - X_1) = (56-44)/(63-44) = 12/19 = 0.6316$

查表三，得n=10时，r_{11} $r_{0.01} = 0.597$

可说有99%以上的把握可舍弃44这个数值。

（二）一个样本中有过大及过小各一个数值时

例2：有一样本由小到大排列如下：—1.40，—0.44，—0.30，—0.24，—0.22，—0.13，—0.05，0.06，0.10，0.18，0.20，0.39，0.48，0.63，1.01。其中最小值—1.40及最大值1.01均怀疑为偏离过大的数值，试用统计方法判断其取舍。

可用戴维（David）法：1.计算其极差$W = X_n - X_1$

2．计算样本标准差S；

3．计算W/S值；

4．查表四，若W/S值大于表四中相同n值的W/S临界值，则可舍去。

表四 戴维 (David) W/S临界值

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	2.00	2.43	2.75	3.01	3.22	3.40	3.55	3.68	3.80	3.91	4.00	4.09	4.17	4.24
W/S	2.00	2.44	2.80	3.10	3.34	3.54	3.72	3.88	4.01	4.13	4.24	4.34	4.43	4.51
	2.00	2.45	2.81	3.12	3.37	3.58	3.77	3.94	4.08	4.21	4.32	4.43	4.53	4.62

n	17	18	19	20	30	40	50	60	80	100	150	200	500	1000
	4.31	4.38	4.43	4.49	4.89	5.15	5.35	5.50	5.73	5.90	6.18	6.38	6.94	7.33
W/S	4.59	4.66	4.73	4.79	5.25	5.54	5.77	5.93	6.18	6.36	6.64	6.85	7.42	7.80
	4.69	4.77	4.84	4.91	5.39	5.69	5.91	6.09	6.35	6.54	6.84	7.03	7.60	7.99

注：W/S值上行α为5%，中行α为1%，下行α为0.5%。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

例 2：$\bar{X}=0.018$，$S=0.5509$，$X_1=-1.40$，$X_n=1.01$，

$W=X_n-X_1=1.01-(-1.40)=2.41$，$W/S=2.41/0.5509=4.374$

查表四：$n=15$时，4.374落在5%和1%危险率之间，故若我们只愿冒1%可能舍弃错误的风险，则应保留此两数值。若愿冒5%机会舍弃错误的风险，则此两数值至少有一个可舍弃，若最大和最小值与平均数相差的绝对值相等，则两个数值均可舍弃，否则可先舍弃一个与平均数相差较大者，再检验另一个是否可舍弃。今两数值与平均数相差分别为1.418及0.992，故宜先舍弃—1.40，再按$n=14$时用格拉布斯法来检验1.01是否可舍弃：结果$n=14$，$\bar{X}=0.119$，$S=0.401$，$T_0=2.22$，查表二，当$n=14$时，T_0 $T_g0.05$，这样大的数值有5%以上的机会出现，故应保留1.01这个数值。

（三）一个样本中有两个最大值或两个最小值为可疑值时

例 3：有一个样本从小到大排列如下：2.02，2.22，3.04，3.23，3.59，3.73，3.94，4.05，4.11，4.13。其中最小的两个数值2.02及2.22可疑为偏离过大的数值，试用统计方法判断其取舍。

可用格拉布斯（Grubbs）另一方法检验之。

1．舍弃两个可疑值后，计算样本离均差平方和$S^2_{1,2}=\sum_{i=3}^{n}\left(Xi-\bar{X}_{1,2}\right)^2$，

其中$X_{1,2}=\left(\sum_{i=3}^{n}Xi\right)\Big/(n-2)$ 或$S^2_{n-1,n}=\sum_{i=1}^{n-2}\left(Xi-\bar{X}_{n-1,n}\right)^2$，

其中$\bar{X}_{n-1,n}=\left(\sum_{i=1}^{n-2}Xi\right)\Big/(n-2)$

2．计算包括两个可疑值在内的样本离均差平方和$S^2=\sum_{i=1}^{n}\left(Xi-\bar{X}\right)^2$

其中$\bar{X}=\left(\sum_{i=1}^{n}xi\right)/n$

3．计算$S^2_{1,2}/S^2$值或$S^2_{n-1,n}/S^2$值。

4．查表五，根据n值大小找出相应的$\frac{S^2_{1,2}}{S^2}$或$\frac{S^2_{n-1,n}}{S^2}$的临界比值，若计算的样本$S^2_{,2}/S^2$或$S^2_{n-1,n}/S^2$值“小”于表中相应临界比值，则可认为此两数值与其他数值有显著差别，可以舍弃。

表五　格拉布斯$S^2_{1,2}/S^2$或$S^2_{n-1,n}/S^2$的临界值

n	4	5	6	7	8	9	10	11	12
$\frac{S^2_{1,2}}{S_2}$	0.0000	0.0035	0.0186	0.0440	0.0750	0.1082	0.1415	0.1736	0.2044
	0.0002	0.0090	0.0349	0.0708	0.1101	0.1492	0.1865	0.2212	0.2536
或 $\frac{S^2_{n-1,n}}{S^2}$	0.0008	0.0183	0.0565	0.1020	0.1478	0.1909	0.2305	0.2666	0.2996
	0.0031	0.0376	0.0921	0.1479	0.1994	0.2454	0.2863	0.3226	0.3552

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

n	13	14	15	16	17	18	19	20
$\frac{S^2_{1,2}}{S^2}$	0.2333	0.2605	0.2859	0.3098	0.3321	0.3530	0.3725	0.3909
	0.2836	0.3112	0.3367	0.3603	0.3822	0.4025	0.4214	0.4391
或 $\frac{S_{n-1,n}}{S^2}$	0.3295	0.3568	0.3818	0.4048	0.4259	0.4455	0.4636	0.4804
	0.3843	0.4106	0.4345	0.4562	0.4761	0.4944	0.5113	0.5269

注：$S^2_{1,2}/S^2$或$S^2_{n-1,n}/S^2$值：第一行α为1%，第二行α为2.5%，第三行α为5%，第四行α为10%。

例3：$S^2_{1,2}=1.197$，$S^2=5.351$，$S^2_{1,2}/S^2=1.197/5.351=0.224$

查表五：当n=10时，0.224小于危险率为5%的临界值（0.2305），大于危险率为1%的临界值（0.1415），若只要求有95%把握舍弃是对的，则可舍弃此两数值；若要求有99%把握舍弃正确，则不应舍弃。再用格拉布斯法检验最小的一个是否可舍弃。

若n太大，计算离均差平方和较麻烦，可用狄克逊法检验，也可先检验最大或最小的一个，再检验另一个次大或次小的。

三、几点注意事项

1．上述几种方法1974年曾有人对混有另一总体数值的各种情况，进行了一万次以上的模拟试验，结果认为格拉布斯方法检出率最高，效果最好。故若同一资料用两种方法检验结果不一致时，一般以格拉布斯方法为准。但当资料太多，n较大时，可用狄克逊法，以减少计算工作量。

2．在采取舍弃办法时，还应三思而而行，因为即使经过计算符合舍弃标准，而舍弃的可疑数值属于固有差异的可能性虽小但毕竟不等于零。且大多数取舍办法是以资料属正态分布为基础的，若资料不属正态分布，舍弃更应慎重，最好先将其变换成正态分布后再行计算。

3．即使偏离很大，也应该追查其出现的原因。历史上不少重大发现，是在追查异常数值出现的原因中获得的，如雷利（Rayleigh）1894年在测定氮的密度时，发现从大气提取的氮和由化学制备的氮有很大差别（相当23个标准差），但他并没有归结于某种差错而舍弃其中的一个数据，从而导致后来发现了一种新元素氩。故凡有舍弃数值时，应在报告中注明被舍弃的具体数值及舍弃的依据。

4．若既找不出舍弃可疑数值的充分理由，又不符合统计学舍弃标准，就不应该舍弃，此时应继续研究，加大样本，重复实验，若偏离过大的数值确属固有差异范围，就可能出现与可疑数值相近、而均差符号相反的数值，即可排除疑点，加强结论的可靠性。

参 考 文 献

（1）冯师颜：误差理论与实验数据处理。P 36～38 科学出版社，1965
（2）张世箕：测量误差及数据处理。P 42～46，科学出版社，1979
（3）何国伟：误差分析法。P 67～68，国防工业出版社，1978
（4）FJ Anscombe，Technometrics vol 2，P 123～147，1960
（5）FE Grubbs，Annals of mathematical Statistics. vol 21，P27～58，1950
（6）FE Grubbs：Technometrics 11(1):1～21，1969
（7）WJ Dixon：Annals of mathematical statitics vol 21，P 488～506，1950
（8）HA DaVid：Biometrics vol 43，P 449～451，1956

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net