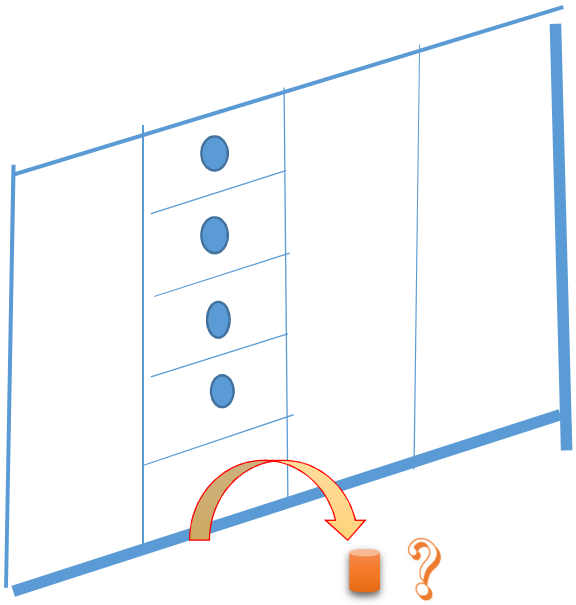
GLP Consulting

http://consultglp.com

基本统计学应用于化学分析系列（III）

离群值检验



目录

3.0 离群值（Outlier）检验概论

离群值是一个统计学里的专业术语。在对同一样本进行多次重复测定时，经常会发现在一组测定数据中有某一个或数个测定值看似比其他测定值明显地偏小或偏大。称这种显著偏离的数据为离群值。

有疑问的观察值可能是测试中随机波动的极度表现，但还处于统计控制的范围之内，是属于同一总体的误差，不算离群值。它也可能是源于个人失误或其他因素如仪器感应器波长飘移所造成的，归类为过失误差。 这些数据须经过一些统计检验方法来判断。在确定为离群值后须从那组数据中剔除，不然会引起较大的系统误差。

离群值的取舍统计方法很多。这里讨论几个较通用的方法，适用于在单次实验中组内离群值的检验。

3.1 巴依达 (Pahta's) 准则

顺序排列一组 n 测定值，如有离群值 x_p，它肯定是出现在极两端的。计算出其平均值 $\bar{x}$ 和标准差 s. 巴依达准则规定若离群值 x_p 与测定平均值 $\bar{x}$ 的偏差的绝对值大于三倍标准差，即

$$| v_p | = | x_p - \bar{x} | > 3s \qquad (3\text{-}1)$$

则认为 x_p 为异常值，应将从该组测定值中剔除掉。

选择 $3s$ 是与显著性水平有关。$3s$ 相当于 99%置信度(显著性水平α = 0.01)。两倍标准差 $2s$ 也可采用，相当于 95%置信度 （$\alpha = 0.05$）。

例 3.1 在例 1.1 的示例中，一实验室重复测定工厂下水道污水样本中的油脂量得如下的结果（mg/L）： 125，142，133，150，129，145 mg/L。 问测定值 125mg/L 可否为离群值？

答

首先求出平均值和标准差： $\mu = 137.3$ mg/L; $s = 9.81$ mg/L; $2s = 19.62$ mg/L

测定值 125mg/L 与平均值的绝对偏差 $v_p = |125 - 137.3| = 12.3\text{mg/L} < 2s$

因此测定值 125mg/L 非离群值，是在 95%统计控制范围内，应该给以保留。

3.2 肖维勒（Chauvenet's）准则

基本理论与上述巴依达准则相似，只是要求离群值与平均值的偏差, v_p 必须是：

$$|v_p| > \omega s \qquad (3\text{-}2)$$

式中 ω 临界值取自于肖维勒表，n 为测定值的总数，如表 3.1。

表 3.1 肖维勒检验的临界值ω

n	ω	n	ω	n	ω
3	1.38	**13**	2.07	**23**	2.3
4	1.53	**14**	2.10	**24**	2.3
5	1.65	**15**	2.13	**25**	2.3
6	1.73	**16**	2.15	**30**	2.4
7	1.80	**17**	2.17	**40**	2.5
8	1.86	**18**	2.20	**50**	2.6
9	1.92	**19**	2.22	**75**	2.7
10	1.96	**20**	2.24	**100**	2.8
11	2.00	**21**	2.26	**200**	3.0
12	2.03	**22**	2.28	**500**	3.2

例 3.2

引用例 3.1 的数据，共有 $n = 6$ 个测定数据，ω_6 临界值 = 1.73

测定值 125mg/L 与平均值的绝对偏差 v_p = |125 – 137.3| = 12.3mg/L < *1.73*x*9.81* 或 17.0mg/L

结论： 测定值 125mg/L 并非为离群值。

3.3 格拉布斯(Grubb's)准则

用格拉布斯准则检验离群值时，当观察值 x_p 的计算值 v_p 大于格拉布斯 G 临界值时，则这数据作为异常值被剔除：

$$|v_p| = |x_p - \bar{x}| > \lambda_{\alpha=0.05,n} s \qquad (3.3)$$

或 $$|v_p| = \frac{|x_p - \bar{x}|}{s} > \lambda_{\alpha=0.05,n}$$

或 $$|v_p| = G_p > \lambda_{\alpha=0.05,n}$$

式中， $$G_p = \frac{|x_p - \bar{x}|}{s}$$

这里的 $\lambda_{\alpha=0.05,n}$ 是与测量数次 n 和显著性水平$\alpha=0.05$ (95%置信度)有关，是依据格拉布斯临界值为准。见表 3.2。

表 3.2 格拉布斯检验的临界值$\lambda_{\alpha=0.05}$

n	*λ(0.05)*	*n*	*λ(0.05)*	*n*	*λ(0.05)*
3	1.15	**18**	2.65	**33**	2.95
4	1.48	**19**	2.68	**34**	2.97
5	1.71	**20**	2.71	**35**	2.98
6	1.89	**21**	2.73	**36**	2.99
7	2.02	**22**	2.76	**37**	3.00
8	2.13	**23**	2.78	**38**	3.01

9	2.21	**24**	2.80	**39**	3.03
10	2.29	**25**	2.82	**40**	3.04
11	2.34	**26**	2.84	**50**	3.13
12	2.41	**27**	2.86	**60**	3.20
13	2.46	**28**	2.88	**70**	3.26
14	2.51	**29**	2.89	**80**	3.31
15	2.55	**30**	2.91	**90**	3.35
16	2.59	**31**	2.92	**100**	3.38
17	2.62	**32**	2.94	**120**	3.44

例 3.3:

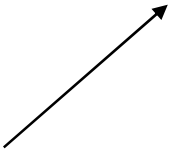
以下一组 *n=10* 测定值中,数据 55.2 表现特低，可用 Grubb’s 统计测定法来确定它是否为离群值。除了检验每个数据的 G-值外，也可以用 Student’s *t*-检验来评定这严重偏低的数据为离群值的置信度，即 *p*-值。*t* 的计算公式为：

$$t = \sqrt{\frac{n(n-2)G^2}{(n-1)^2 - nG^2}} \qquad (3.4)$$

表 3.3　检验一组 10 个重复测定数据中的离群值

复测数 *n*	观察值 *x*	*G*-值	*t*- 值	双边 *p*-值	置信度 *p*-值 (*n* x 双边 *p*-值)
1	56.5	0.1446			
2	56.2	0.398			
3	56.8	0.687			
4	56.5	0.145			
5	56.3	0.217			
6	57.0	1.048			
7	56.4	0.036			
8	57.2	1.410			
9	56.1	0.578			

10	* 55.2	2.205	3.465	0.00851	0.0851
均值 =	56.4				
标准差 s =	0.553				
n =	10	用 Excel 计算	=		
G 临界值 =	2. 29	双边 p 值	TDIST(t,DF,2)		



基于以上的统计分析中，观察 G-值 2.205 是小于 $G_{n=10}$ 临界值 2.29，而且置信度 p 值为 0.085 是大于$\alpha = 0.05$，这说明了数据 55.2 没有显著偏离随机波动的范围，非离群值，因此不能被剔除。

3.4 狄克孫(Dixon's)准则

狄克孙准则要求先按小到大顺序排列一组测定值：$x(1), x(2), \ldots, x(n-1), x(n)$。异常值会出现在数据两端,即 $x(1)$ 或 $x(n)$ 。检验 $x(1)$ 或 $x(n)$ 时，使用表 3.4 中所列的公式得到统计量 D_o。

若 $D_o > D_{(\alpha,n)}$, 则应剔除 $x(1)$ 或 $x(n)$。这里的临界值 $D(\alpha,n)$与显著性水平α及测定次数 n 有关。见表 3.4。

例 3.4 同一样品测得 8 个数据(%)：44.2, 43.3, 43.6, 41. 6 , 43.7, 44.0, 43.1, 44.4。试问在α=0.05 置信度下，41.9%是否为异常值要剔除？

解： 将数据按小到大顺序排列：41. 6 , 43.1, 43.3, 43.6, 43.7, 44.0, 44.2, 44.4 (%) 。由表 3.4 可得：

$$D_o = \frac{43.1 - 41.6}{44.2 - 41.6} = \frac{1.5}{2.6} = 0.58$$

$D_{(\alpha=0.05,8)}$临界值 $= 0.608$

$D_o = 0.58 < D_{(\alpha=0.05,8)}$ 故 $x(1) = 41.6\%$数据在统计学上不是个偏离值，不可剔除。

表 3.4　狄克孙检验的临界值 D(α,n)值及 D_o 计算公式

D_o 计算公式	n	α=0.05	α=0.01
	3	0.970	0.994
$D(3...7) = [x(2) - x(1)] / [x(n) - x(1)]$	4	0.829	0.926
或	5	0.710	0.821
$D(3...7) = [x(n) - x(n-1)]) / [x(n) - x(1)]$	6	0.628	0.740
(选择计算公式是根据 x(1) 或 x(n)有可疑时)	7	0.569	0.680
	8	0.608	0.717
$D(8...12) = [x(2) - x(1)] / [x(n-1) - x(1)]$	9	0.564	0.672
或	10	0.530	0.635
$D(8...12) = [x(n) - x(n-1)]) / [x(n) - x(2)]$	11	0.502	0.605
(选择计算公式是根据 x(1) 或 x(n)有可疑时)	12	0.479	0.579
	13	0.611	0.697
	14	0.586	0.670
	15	0.565	0.647
	16	0.546	0.633
	17	0.529	0.610
	18	0.514	0.594
	19	0.501	0.580
	20	0.489	0.567
	21	0.478	0.555

	22	0.468	0.544
	23	0.459	0.535
D(13…27) = [x(3) – x(1)] / [x(n-2) – x(1)]	24	0.451	0.526
或	25	0.443	0.517
D(13…27) = [x(n) – x(n-2)] / [x(n) – x(3)]	26	0.436	0.510
(选择计算公式是根据 x(1) 或 x(n)有可疑时)	27	0.429	0.502