光谱数据舍弃——极差法

一七〇厂中心试验室　　周顺庆

本文采用差值法导出极差舍弃标准，方法简捷，在日常分析工作中用心算可以决定数据的取舍，保证工厂生产中的实际精度要求。

一、 Chauvenet数值舍弃标准[1]

Chauvenet标准 设在n个读数中任一读数与平均值的偏差是这样大小，凡等于或大于此偏差的所有偏差出现的机率均小于1/2n时，此读数应予弃去。这个标准也可部分借用公式来表达。设一读数与平均值之差大于$\tilde{X}$，$\tilde{X}$由下规定：

$$1-P\tilde{x}=1-\frac{N}{\sqrt{\pi}}\int_{-x}^{\tilde{x}}e^{-h^2x^2}dx=\frac{1}{2n} \quad \cdots\cdots(1)$$

则此读数应弃去。并有表如下：

表 1

n	5	6	7	8	9	10	12	14	16	18
x/σ	1.68	1.73	1.79	1.86	1.92	1.96	2.03	2.10	2.16	2.20

上表的用法：设在一组测量中，观察次数为6次。先算出σ，其中某次读数为X_i，当$X_i>1.73\sigma$，则X_i应舍弃。

由（1）式可知当n减少其精度降低，而在工厂的实际测定中，对一试样摄谱次数一般为3～4次不可能太多，另一方面Chauvenet标准在应用上是不方便的，应该采用更简捷的方法。

二、 极差舍弃

光谱的偶然误差分布是高斯正态分布，其概率密度为

$$P(x)=\frac{1}{\sqrt{2\pi}\,m_{\Delta s}}e^{\frac{-(x-\mu)^2}{2m_{\Delta s}}} \quad \cdots\cdots(2)$$

落在（a、b）间的概率为：

$$P(a<X<b)=\frac{1}{\sqrt{2\pi}m_{\Delta s}}\int_a^b e^{\frac{-(x-\mu)^2}{2m_{\Delta s}}}dx \quad \cdots\cdots(3)$$

经计算知：[2]

（$\mu-m_{\Delta s}$，$\mu+m_{\Delta s}$）的概率为68.3%

（$\mu-2m_{\Delta s}$，$\mu+2m_{\Delta s}$）的概率为95.4%　　……（4）

（$\mu-3m_{\Delta s}$，$\mu+3m_{\Delta s}$）的概率为99.7%

由上可知：超过$\pm3m_{\Delta s}$的误差其出现几率在一千次测定中才出现3次，一般在生产中测定次数是有限的，所以凡超过$\pm3m_{\Delta s}$的误差一定不属于偶然误差，应该舍弃。

我们将测得的ΔS分为b组，每组内各有a个ΔS，同一组的ΔS属于同一个样品观测值。

$$\begin{array}{llllll}
\text{第1组} & \Delta S_{11}、 & \Delta S_{12}\cdots & \Delta S_{1j}\cdots & \Delta S_{1a} \\
 & \vdots & \vdots & \vdots & \vdots \\
\text{第2组} & \Delta S_{21}、 & \Delta S_{22}\cdots & \Delta S_{2j}\cdots & \Delta S_{2a} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{第i组} & \Delta S_{i1}、 & \Delta S_{i2}\cdots & \Delta S_{ij}\cdots & \Delta S_{ia} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{第b组} & \Delta S_{b1} & \Delta S_{b2}\cdots & \Delta S_{bj}\cdots & \Delta S_{ba}
\end{array}$$

其偶然误差分布是高斯正态分布，其概率密度为：

$$f(x)=\frac{1}{\sqrt{2\pi}\,m_{\Delta s}}e^{\frac{-(x-\overline{\Delta S})^2}{2m_{\Delta s}^2}} \quad \cdots\cdots(4)$$

把同组中任二个ΔS差值$\Delta S_{ij}-\Delta S_{ij'}$ $j\neq j'$看成一个新的统计随机变量，则它亦是

高斯正态分布。其概率密度为：

$$g(x)=\frac{1}{\sqrt{2\pi}\,m'_{\Delta s}}\,e^{\frac{-(x-\mu)^2}{2m'^{2}_{\Delta s}}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

根据误差传递公式得：

$$m'_{\Delta s}=\sqrt{2}\,m_{\Delta s} \quad \cdots\cdots\cdots\cdots (6)$$

对于$\Delta S_{ij}-\Delta S_{ij}'=d_{ij}$来说，同一样品的观察值其真值$\mu=0$，由（5）式知：$d_{ij}$落在

$$\left.\begin{array}{l}(-m'_{\Delta s},+m'_{\Delta s})\text{ 的概率为 }68.3\% \\ (-2m'_{\Delta s},+2m'_{\Delta s})\text{ 的概 率为 }95.4\% \\ (-3m'_{\Delta s},+3m'_{\Delta s})\text{ 的机 率为 }99.7\%\end{array}\right\} \quad \cdots\cdots\cdots\cdots\cdots\cdots (7)$$

引用“差值法”的结论[4]：

$a=2 \quad m_{\Delta s}=\sqrt{\frac{\pi}{2}}\,\bar{d}$，

$a=3 \quad m_{\Delta s}=\sqrt{\frac{\pi}{3}}\,\bar{d}$； $a=4 \quad m_{\Delta s}=\frac{\bar{d}}{2.0}$

a为同一样品摄谱次数，$\bar{d}$为各组极差的平均。

在可信度为$\alpha=95.4\%$时

当$a=2 \quad 2m'_{\Delta s}=2\cdot\sqrt{2}\,m_{\Delta s}$

$=2\cdot\sqrt{2}\cdot\sqrt{\frac{\pi}{2}}\,\bar{d}=2.507\bar{d}$

$a=3 \quad 2m'_{\Delta s}=2\cdot\sqrt{2}\,m'_{\Delta s}$

$=2\sqrt{2}\sqrt{\frac{\pi}{3}}\,\bar{d}=1.671\bar{d}$

$a=4 \quad 2m'_{\Delta s}=2\sqrt{2}\,m_{\Delta s}$

$=2\cdot\sqrt{2}\,\frac{\bar{d}}{2.0}=1.414\,\bar{d}$

同样在$\alpha=99.7\%$时

$a=2 \quad 3m'_{\Delta s}=3.760\bar{d}$

$a=3 \quad 3m'_{\Delta s}=2.507\bar{d}$

$a=4 \quad 3m'_{\Delta s}=2.121\bar{d}$

表 2 极差舍弃的标准

α \ a / $k\bar{d}$	2	3	4
95.4%	$2.507\bar{d}$	$1.671\bar{d}$	$1.414\bar{d}$
99.7%	$3.760\bar{d}$	$2.507\bar{d}$	$2.121\bar{d}$

三、极差舍弃的应用

1．按差值法统计极差的平均值$\bar{d}$；

2．在表2上找出K值；

3．算出$k\bar{d}$，作为生产中测定舍弃数据的标准。

$\bar{d}$可以根据投产前的试验或者日常生产中的数据来统计，只要投产前后的条件不变，其均方差的变动是很小的，所以可以根据预先算得的$k\bar{d}$值来作为日常分析的数据舍取标准。

4．应用示例：若某些元素测定其$\bar{d}=2$，$k\bar{d}=4.28$（四次摄谱）。在日常测定中测得某组值为

ΔS 10，12，15，11 我们马上就可以知道15是偏离最大的数，$d_{ij}=15-10=5$这个值超过4.28应该舍弃15这个数，余下的三个数之间的极差小于4，所以它们应该保存下来。

在测定中有时会有大于$K\alpha=99.7\%\cdot\bar{d}$的数据，如果只有那么几个，应该再测定一下以证实不是测量的错误，如确实存在，那么按示例舍弃。如果这种弃舍太多，这说明条件变化了，或者是一系列的操作中有错误，应该重新做试验找出原因并加以克服。本方法应用方便简捷，可以一边记录一边检查，心算舍弃数据并确定试验成功与否。顺便提一下，当$a=2$时只能判别这组数据是否可取，不能判别应该舍哪个数，在这种情况下只能重做试验，不过一般在日常生产中并不采用$a=2$的摄谱法。

四、极差舍弃统计

表3的统计是从日常生产中统计来的，由表3得出下面的结论：

1．$\alpha=99.7\%$时超过$k\bar{d}$的极差几乎不存在，凡超过者必须按示例舍弃一个偏差最大的数，因为它不是偶然误差所致。

2．$\alpha=95.4\%$时超过$k\bar{d}$的极差几乎每块谱板上都存在，但它们出现的几率不会很大，在5%左右，它可能是偶然误差所致，应该复查，一般不要舍去。

表 3

元素含量 %	a	$\bar{d}$	单板测定						
			$\alpha=95.4\%$			$\alpha=99.7\%$			$\sum\limits_{ij} d_{ij}$*
			$k\bar{d}$	$d_{ij}>k\bar{d}$ 个数	$1-\alpha\%$	$k\bar{d}$	$d_{ij}>k\bar{d}$ 个数	$1-\alpha\%$	
B 0.012 ～0.022	3	3.098	5.18	2	2.3%	7.77	0	0 %	72
				4	5.6%		0	0 %	
	4	3.098	5.63	3	2.7%	9.97	0	0 %	108
				1	1.6%		0	0 %	60
Co 14～16	4	1.71	2.42	2	2 %	3.63	0	0 %	96
				0	0 %		0	0 %	84
Mo 2.5～3.5	4	2.09	2.96	2	2 %	4.43	0	0 %	96
				2	2.7%		1	1.39%	72
V 0.6～0.9	4	2.75	3.89	1	1.1%	5.89	0	0 %	90
				0	0 %		0	0 %	96
Cr 8.5～9.5	4	1.70	2.40	3	3 %	3.61	0	0 %	90
				0	0 %		0	0 %	72
Ni 8～11	4	1.91	2.70	5	5.2%	4.05	0	0 %	96
				2	1.85%		0	0 %	108
W 1.5～2.0	4	1.62	2.29	3	2.8%	3.44	0	0 %	108
				1	0.9%		0	0 %	108
Ti 1.4～1.9	4	3.10	4.38	2	1.85%	6.58	0	0 %	84
				4	4.76%		0	0 %	102
Al 3.7～4.4	4	2.39	3.38	1	0.9%	5.07	0	0 %	102
				4	3.9%		1	0.98%	102

* $\sum\limits_{ij} d_{ij}$ 为一谱板上的所有差值总数　　a＝3则　3×n样品数　　a＝4则　4×n样品数

五、结　论

上面采用的$m_{\Delta}s \doteq \bar{d}/2.0$ (a＝4) 为近似的经验公式但应用简单完全满足工厂测定的精度要求[6]。另一个是 $m_{\Delta}s$是要涨落的，所以统计$\bar{d}$时要数量足够大的数据[5]，要有代表性，在应用时可以适当地根据其可能的涨落放大一些，只要使用合理是能保证数据的真实性的。此法简便应用时一目了然快速，有一定的数学根据可防止经验上的误断。

参 考 文 献

〔1〕误差理论与实验数据处理P38冯师颜编科学出版社出版　1964年11月第二次印刷

〔2〕常用数理统计方法P6　中国科学院 数学 研究所统计组编　科学出版社1973 年

〔3〕同〔2〕　P9

〔4〕〔5〕〔6〕光谱分析偶然误差的简化统计方法。　黎明机械厂　许国敷著　1965年11月交流资料