特异值的识别、处理及块段平均品位估计

侯景儒 张树泉 黄竞先
（北京科技大学地质系） （中国有色总公司北京地质所）

提 要 特异值问题可能存在于许多抽样调查之中，特异值在地质勘探及采矿工程中称为特高品位，它一般对应于高品位的矿化作用。因此，特异值对于地质研究、矿石储量计算以及数据处理十分重要。本文主要讨论了以下四个问题：1）特异值及特高品位的概念；2）特异值的统计学意义；3）特异值的识别及处理方法（包括估计邻域法及影响系数法）；4）用于包含特异值数据的非参数统计方法（包括指示克立格、对数正态克立格法、切尾法以及顺序一秩统计量法）。

关键词 特异值 指示克立格法 切尾法 对数正态克立格法 顺序一秩统计量法 变异函数

一、特异值及特高品位

众所周知，许多数学地质方法及矿产储量计算方法都是基于经典的概率统计理论，而且要求所研究的变量的观测值服从正态分布，否则，其统计计算的结果将会出现程度不等的偏差，而影响观测值不服从正态分布的重要因素之一则是在观测值中存在有特异值（outlier）。所谓特异值是指具有以下特点的数值：

1）它比所研究的全部数据的算术平均值或中位数的数值要高得多。

2）它是实实在在地存在于我们所研究的母体之中（例如在某一矿化范围的高品位值），因此，特异值绝非采样或化验分析等所引起的人为的误差。

3）它在全部所研究的数据中只占极少部分，但它对全部数据的统计结果影响极大，例如一些金矿体中的特高品位样品只占全部样品的极少部分，但却占总储量的很大比例；

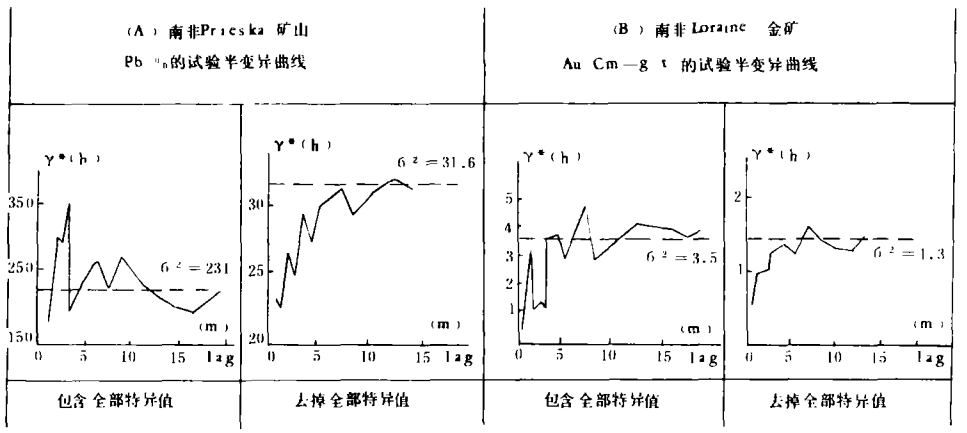
1）它的出现只限于所研究母体的一定空间位置，例如一些高品位的金属矿化仅仅局限于浸染矿化中的火山碎屑角砾岩中。

对一个金矿床的研究表明，Au 在全矿床中平均品位 $m = 10g/t$，中位数 $M = 3.16g/t$，其方差为 $\sigma^2 = 900(g/t)^2$，变化系数 $V = \sigma/m = 3$，而在该矿床中样品 Au 含量大于 100g/t 者只占总样品数的 1%，但这 1%样品 Au 的平均品位却等于 196g/t，它确实存在于该矿床的一定的空向位置，而且对于储量计算至关重要，因此，我们称这 1%的样品的观测值为特异值。

广而言之，在许多抽样问题中，均可能出现特异值，例如产品检验、环境检测分析、社会调查、矿业工作中的化探、物探取样，矿山地质、水文地质、工程地质中的样品分析等，当在地质勘探及矿山地质研究中出现特异值（高值）时，我们称之为特高品位，因此，如果特异值为一统计术语的话，特高品位则为等同于特异值的地质及矿业术语。

由于特异值对统计推断的严重影响，而早为统计学家所注意，著名的瑞士统计学家Bernoulli早已指出特异值在很大程度上影响着基于最小二乘法的许多数学方法的计算结果，Peirce很早就提出了识别和剔除特异值的方法。

基于区域化变量理论的地质统计学是研究既具有随机性又具结构性的自然现象的最好方法，特异值的存在却严重的影响了变异函数的计算（如图1所示），众所周知，变异函数是地质统计学研究中的重要数学工具，变异函数计算的偏差无疑大大影响了地质统计学研究结果的最优及无偏性。



A. 南非 Prieska 矿山 Pb%的试验半变异曲线　　B. 南非 Loraine 金矿 Aug/t 的试验半变异曲线

图中 δ^2 为方差，lag为滞后距，$r^*(h)$ 为试验半变异函数值）

图1　特异值对试验半变异曲线的影响

（据 G. Krige 资料）

Fig 1 Influnce on test semivariation curve by outlier

二、特异值的统计意义

在进行数学地质研究时，首先是对地质数据通过构制频率直方图或统计分布曲线来研究地质数据的统计分布特征，这种对数据分布律研究的重要性不仅在于分布函数是地质体最重要的数学特征之一，而且是进一步进行统计学研究的基础，因此，了解特异值的统计意义对于如何去识别并处理特异值是十分重要的。

假设一个总体是由正常值及特异值组成，$f_V(X)$ 为该总体中正常值的分布密度函数，$f_0(X)$ 为该总体中特异值的分布密度函数（或称为污染分布函数），且该特异值所占总体的比列为P，则该总体观测值的分布密度函数 f (X) 为：

$$f(X) = (1-P)f_V(X) + Pf_0(X) \qquad (1)$$

f (X)、$f_V(X)$ 及 $f_0(X)$ 之间的关系如图 (2a) 所示。图(2a)表示特异值只出现在正常值分布的右

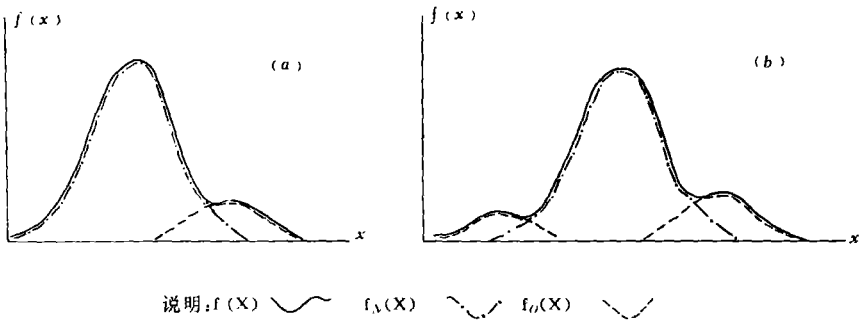


图 2 观测值、正常值及特异值分布函数之间的关系

Fig. 2 Distribution function of observation value, normal values and outliers

端时的情况．当特异值同时也出现在正常值分布的左端，并以 $f_{o'}(X)$ 表示其分布密度函数时，其图形如图(2b)所示，令左端特异值所占总体的比例为 P_1 时，(2)式可改写如下：

$$f(X) = (1 - P - P_1)f_\lambda(X) + Pf_o(X) + P_1 f_{o'}(X) \quad (2)$$

当我们从总体观测值的分布中找到正常值的分布就不难识别出特异值了，所谓特异值的识别就是从母体观测值分布 f(X)中将特异值的分布 $f_o(X)$ 与正常值的分布 $f_\lambda(X)$ 区别开来。在地质科学及采矿工程中重点研究右端特异值的分布 $f_o(X)$，除非在特殊条件下才研究左端特异值的分布 $f_{o'}(X)$，因此，所谓特异值一般是指对于 $f_o(X)$ 而言。

三、特异值的识别及处理

如前所述，由于特异值的存在而较大的影响了一些重要的统计参数(如均值、方差及半变异函数等)，对于特异值的研究实际上是进行各种统计分析之前对样品进行预处理，目前对这一问题的研究已经取得了较大的进展，例如在地质勘探中对于特异值的识别及处理就有以下几种方法：1)按均方差的倍数来识别特异值，例如，对于变化简单的矿床，其特异值定为样品均值 m 加三倍的均方差 σ：m+3σ；2)，按样品品位的变化系数 (V = σ/m) 来识别特异值，例如，当变化系数 V < 20 时，其特异值的下限定为样品均值的(2～3)倍，当 V > 150 时(例如某些稀有金属矿床或金矿床)，其特异值的下限定为样品均值的(15～20)倍等；3)在分布密度函数曲线上将拐点对应的值作为特异值的下限值(图 2)。处理特异值的方法也多种多样；或将样品中的特异值全部去掉，不参加统计；或用正常值的最大限值代替特异值；或用总体的平均值(包含特异值的或不包含特异值的)代替特异值进行统计等等，所有上述方法在生产实践中也不同程度的发挥了作用，但也存在着某些勿容忽视的缺点，例如只是经验方案而无统计意义，或者未将特异值与正常值置于同一邻域来研究等。下边我们将讨论一些在理论及实践上更为优越的识别及处理特异值的方法。

1、估计邻域法

该法是 D、G、Krige 及 D、M、Hawkins 把地质统计学的基本思想用于识别和处理特异值的方法。地质统计学是以区域化变量理论为基础，以变异函数为基本工具，研究那些在空间分布上既有随机性又有结构性的自然现象的科学，特异值属于具上述性质的区域化变量。地质统计学研究的基本方法是在估计方差极小条件下，通对对待估点（或块段）影响范围之内所有信息值（例如样品品位值）进行加权平均来估计待估点（或块段）的平均品位，从地质统计学的观点来看，一个观测值是否为特异值，不仅要考虑观测值本身，还应考虑与其相邻（即影响范围之内）的若干样品的观侧值，也就是说，识别和处理特异值时既要考虑观测值本身，又要考虑它所处的空间环境，例如，在一个金矿床中，某金含量为 3g/t 的样品观测值在低品位区可能被确定为特异值，但在高品位区，Au 含量为 8g/t 的样品观测值仍属于正常值范围、因此邻域法是把被识别的观测值（称为可疑样品）置于一个空间连续矿化域的背景上进行研究。

Hawkins (1980)提出识别特异值的统计量如下：

$$I=\frac{n(G-M)^2}{(n+1)\sigma^2} \tag{3}$$

(3)式中：I 是识别特异值的统计量，它是服从自由度为 1 和∞的 F 分布，当 I > 3.84 时，可疑值 G 被确定为特异值，即表示在 95%的置信区间上确定 G 为特异值；G 为被研究的可疑值 (Suspect value)；M 为不包含可疑值 G 的邻域内其他样品观测值的算术平均值；n 为不包含 G 的邻域内样品数；σ^2 为邻域内观测值的平均方差。关于式(3)有两点必须指出：

1) 一般来讲，不包含 G 的邻域内的样品数 n 可以按照所研究变量的变异性来确定，当其变异性很小时，可取 n = 4，当其变异性很大时可取 n = 10，当然，也可以用变异函数确定的变程值 a 来确定邻域范围从而确定更加合理的 n 值。

2)，式(3)中的 σ^2 计算公式是：

$$\sigma^2=\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}r_{kl}(h) \tag{4}$$

式(4)中 $r_{kl}(h)$ 为相距 h 的两个样品点 k，l 的观测值 Z (k)与 Z (l)之间的半变异函数值，其计算公式是：

$$\gamma(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(X_i)-Z(X_i+h)]^2 \tag{5}$$

也可以先求出被研究变量的理论半变异函数模型，再在邻域中变换 h 求出不同滞后距 h 的变异函数值，当然，也可以应用求半变异函数更为稳健的方法。式(4)中的 m 为包括可疑值在内的邻域宽度（即样品数）。

当我们用式(3)识别出特异值之后，下一步就是如何处理该特异值，一种方法是用特异值的下限值 GL 代替特异值。根据式(3)，当 I = 3.84 时，可疑值 G 即为 GL，这时，式(3)改写为：

$$\frac{n(GL-M)^2}{(n+1)\sigma^2}=3.84 \tag{6}$$

于是：

$$GL=\sqrt{\frac{3.84\sigma^2(n+1)}{n}}+M \tag{7}$$

当然,也可以用 GL 代替特异值之后再求邻域内包括特异值样品的邻域平均值来代替特异值,即用〔n·M+GL〕/(n+1)来代替特异值;也可以用包含可疑值的邻域平均值代替特异值,即用(n·M+G)/(n+1)来代替特异值。

表 1 估计邻域法 I 值计算结果

Tadle 1 I values of estimation neighborhood calculation

样 品 号	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
坐标(m)	1	2	3	4	5	6	7	8	9	10
观测值(g/t)	4	3	2	8	5	2	4	6	7	10
M	5.2	5.3	5.4	4.7	5.1	5.4	5.2	5.0	4.8	4.5
I	0.21	0.78	1.69	1.48	0.002	1.69	0.21	0.14	0.64	4.23

实例:某矿区沿穿脉取样,样品间距为 1m,每一样品的观测值如表 1 所示,用估计邻域法识别并处理特异值的计算如下:首先计算 σ^2 值,按式(4)计算得 $\sigma^2 = 6.29$。其次按式(3)计算每一样品的 I 值(如表 2 所示),从 I 值知 v_{10} 样品的观测值为 $I = 4.23 > 3.84$,故为特异值,用式(7)计算特异值的下限值 GL = 9.74g/t,因此,可以用 9.74g/t 代替样品 v_{10} 的观测值 10g/t。

2、影响系数法

该法既不是在统计理论基础上推演出来的,也不像估计邻域法那样充分考虑被研究样品存在的空间环境,但这个方法却能在研究矿化变异程度的基础上适当地抑制特异值影响的程度,这在生产矿山中也许用途更广,当我们掌握了某矿山若干有用的资料及数据后,发现根据地质资料预计的矿石品位与实际生产的矿石品位相差很大,这时就可以用影响系数法找到一种较为理想的识别及处理特异值的方案。其方法原理如下:

设有一组观测值,其样品数为 n

令 M 为包含特异值在内的 n 个观测值的均值,令 m 为去掉特异值的 (n−1) 个观测值的均值,则当

$$M/m \leqslant K+1 \tag{8}$$

时,认为该组观测值均为正常样品,即无特异值,式(8)中的 K 是根据所研究的地质变量在空间的变异性人为赋给的。例如,当 K = 0.1 时,说明特异值对全部样品值的影响不得超过 10%,即当包含该特异值时,全部样品的平均品位最多只允许提高 10%,如果影响超过 10%,则该值被识别为特异值。

设第 n 个样品为特异值 GL,则

$$\frac{M}{m} = \frac{\sum_{i=1}^{n} Z(X_i)/n}{(\sum_{i=1}^{n} Z(X_i) - GL)/n - 1} = \frac{(n-1)\sum_{i=1}^{n} Z(Xi)}{n(\sum_{i=1}^{n} Z(Xi) - GL)}$$

$$= M\frac{(n-1)}{\sum_{i=1}^{n} Z(Xi) - GL}$$

根据(8)式，当第 n 个样品值恰好等于特异值下限时，则

$$M\frac{(n-1)}{\sum_{i=1}^{n}Z(Xi)-GL}=K+1$$

最后得特异值的下限值 GL 为：

$$GL=M(\frac{n\cdot K+1}{K+1}) \tag{9}$$

我们可以用特异值的下限值来代替特异值。

实例：我们用表 1 给出的 10 个样品观测值来进行影响系数法的计算，识别特异值的计算结果见表 2：当取 K = 0.1 时，K+1 = 1.10，从表 3 可知 V_{10} 样品的 M/m = 1.12 > 1.10 故为特异值，用式(9)计算特异值的下限值 GL = 9.27，这样可以用 9.27g/t 代替 V_{10} 样品的观测值 10g/t。

表 2　影响系数法 M/m 计算结果

Table 2　M/m values of influnce coefiecient calculation

样品号	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
观测值 g/t	4	3	2	8	5	2	4	6	7	10
M	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
m	5.22	5.33	5.44	4.78	5.11	5.44	5.22	5.00	4.89	4.56
M/m	0.98	0.96	0.94	1.07	0.998	0.94	0.98	1.02	1.04	1.12

必须指出：用影响系数法对样品组的观测值进行多次扫描方能识别出所有特异值，而后一次扫描可能识别出新的特异值(它们可能是前一次扫描未识别出的，也可能是已识别出的)，给新识别出的特异值应赋于式(9)定义的特异值的下限值 GL，这样反复扫描直至再也识别不出新的特异值为止。

四、用于特异值数据的若干非参数统计方法及块段平均品位的估计

对于一个总体的研究，统计学常用两个参数来表征其分布特点：一个是分布的中心位置，另一个是偏离该中心位置的离散的量，在经典统计学中这两个参数是算术平均值

$$m=\frac{1}{n}\sum_{i=1}^{n}Xi \tag{10}$$

及样本方差

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(Xi-m)^2 \tag{11}$$

但是，当所研究的总体不服从正态分布时，m 及 σ^2 表现出不同程度的不稳健性，为了解决这个

问题有两个途径：一是将原始数据先进行预处理使之服从正态分布（例如对数变换及前边所述的对特异值的识别及处理方法），然后再进行统计研究；另一个途径是无需假设数据服从某种分布，也无需对数据进行变换，即非参数统计方法，下边将介绍几种用于估计中心位置及离散性的较稳健的非参数统计方法。

1、**指示克立格法**（Indicator Kriging）

该法是一种不必去掉实际存在的特异值，在一定风险条件下估计未知值 Z（X）的估计量 $Z^*(X)$ 及其空间分布的非参数方法，指示克立格法把数据中特异值的影响限制到该特异值只能影响的范围之内，但却以概率的形式考虑了特异值的存在。该法的基本做法归纳如下。

1）设矿床 D 化验了一组样品的某金属品位，并且约定其边界品位为 Z，在每一样品点 X∈D 上定义一个 Z 的如下阶梯函数：

$$i(X;Z)=\begin{cases}1 & \text{当}X\text{点上金属品位}Z(X)\leqslant Z\\ 0 & \text{当}X\text{点上金属品位}Z(X)>Z\end{cases} \tag{12}$$

在矿床 D 上任一待估域 A∈D 内，低于边界品位 Z 的品位值 Z（X）占待估域 A 的比例表示如下：

$$\Phi(A;Z)=\frac{1}{A}\int_A i(X;Z)d_X \in[0,1] \tag{13}$$

2）设在待估域 A∈D 的邻域内有 n 个有效数据：{Z（Xα），Xα∈A，α = 1，2，…n}，在约定的边界品位值 Z 下，得到样品的指示函数空间：{i（Xα；Z），α = 1，2，…n}，则 Φ（A；Z）的线性估计量 $\Phi^*(A;Z)$ 是：

$$\Phi^*(A;Z)=\sum_{\alpha=1}^{n}\lambda_\alpha(Z)\cdot i(X\alpha;Z) \tag{14}$$

在给定一系列边界值 Z_l：{Z_l，l = 1，2，…，L}则（14）式可改写成：

$$\Phi^*(A;Z_l)=\sum_{\alpha=1}^{n}\lambda_\alpha(Z_l)\cdot i(X_\alpha;Z_l) \tag{15}$$

3）为了求解（14）式或（15）式中的权系数 $\lambda_\alpha(Z)$ 或 $\lambda_\alpha(Z_l)$，必须求解在无偏条件及估计方差极小条件下构制的下列指示克立格方程组：

$$\begin{cases}\sum_{\beta=1}^{n}\lambda_\beta(Z)\overline{C}\,i(X_\alpha,X_\beta;Z)-\mu=\overline{C}\,i(X_\alpha,A;Z)\\ \sum_{\alpha=1}^{n}\lambda_\alpha(Z)=1\end{cases} \tag{16}$$

4）在（15）式的情况下，L 型个矿化的品位平均值可根据待估域 A 的邻域内 $n_l(l=1,2,\cdots,L)$ 个 L 型个矿化的信息值 Z（$X\alpha_l$）（$\alpha_l=1,2,\cdots n_l$）求得：

$$[Z(A)\mid A\in l\text{型矿化}]^*=\sum_{\alpha_l=1}^{n_l}b\alpha_l\cdot Z(X\alpha_l) \tag{17}$$

（17）式中的 $b\alpha_l$ 可以用普通克立格法或其他方法求得。

5）待估域 A 的平均品位的估计量 $[Z(A)]^*$ 是：

$$[Z(A)]^*=\sum_{l=1}^{L}[\Phi^*(A;Z_l)]\cdot[Z(A)\mid A\in l\text{型矿化}]^* \tag{18}$$

下边给出该法在一个金—银矿床上的应用实例。在该矿床中有 51.6%的样品的 Au 品位大于最低边界品位值 0.2799g/t，其中有 10.5%的样品 Au 品位大于 0.9331g/t（经济边界品位），这 10.5%的样品代表的金量占总储量的 61%，其平均品位为 4.3234g/t。为了应用指示克立格法估计在具有特异值数据的条件下待估块段的金的平均品位，首先按分位数确定了 9 个边界品位并计算了相应的指示半变异函数 $\gamma_I^*(h;Z_l)$ $(l=1,2,\cdots,9)$，研究表明边界品位为 0.2799g/t、0.4666g/t、0.6221g/t 及 1.0264g/t 的 $\gamma_l(h;Z_l)$ 明显不同，所以用上述 4 个边界品位确定了 5 个品位级别进行研究，用待估域 A 的估计邻域内 40 个有效信息来估计待估域 A 的 Au 平均品位值 $[Z(A)]^*$ 时，可用式(15)，(16)，(17)，(18)进行计算，对于待估域 A 而言，4 个边界品位下的 $\Phi^*(A;Zl)$ $(l=1,2,\cdots 4)$ 及 $[Z(A)\mid A\in\ |$ 型矿化 $]^*$ $(l=1,2,\cdots 4)$ 如表 3 所示：

表 3 指示克立格法 $\Phi^*(A;Zl)$ 和 $[Z(A)\mid A\in l]^*$ 计算结果

Table 3 Indicator kriging calculation values

Zl	～0.2799	～0.4666	～0.6221	～1.0264	～
$\Phi^*(A;Zl)$	566.0837	202.1728	0.0	1797.7823	544.3113
$[Z(A)\mid A\in l]^*$	0.1244	0.3888	0.6221	0.8709	3.8879

$$[Z(A)]^* = (5.6608\times 0.1244)+(2.0217\times 0.3888)+(17.9778\times 0.8709)+(5.4431\times 3.8879)$$
$$= 38.3093\ g/t$$

数据中的特异值 16.1427g/t对于 $[Z(A)]^*$ 值的影响被限制到它只能影响的品位级别之中，而该品位级别在 $[Z(A)]^*$ 中仅占 17.5%。

2、对数正态克立格法

设矿床 D 的样品值及块段品位值均服从对数正态分布（或三参数对数正态分布），其平均品位为 Z，矿床 D 内某待估块段为 $V\in D$，其平均品位为 Z_V，Z_V 的估计值为 Z_V^*，其对数值 $\ln Z_V^*$ 可用 n 个已知数据 $\ln(x_\alpha)$ $(\alpha=1,2,\cdots,n)$ 的线性组合来表示：

$$InZ_V^* = C+\sum_{\alpha=1}^{n}\lambda_\alpha In(X_\alpha) \tag{19}$$

式中 C、λ_α 为待定系数，X_α 是定义于信息支撑 V_α $(\alpha=1,2,\cdots,n)$ 的 n 个信息值。

为了求诸权系数 λ_α 及 C 值应解如下的对数正态克立格方程组：

$$\begin{cases}\sum_{\beta=1}^{n}\lambda_\beta\overline{C}_e(V_\alpha,V_\beta)-\mu=\overline{C}e(V_\alpha,V)\\ \sum_{\alpha=1}^{n}\lambda_\alpha=1 \qquad (\alpha=1,2,\cdots,n)\end{cases} \tag{20}$$

(20) 式中 $\overline{C}e(V_\alpha,V_\beta)$ 代表矢量 h 的两个端点分别在样品支撑 V_α 及 V_β 内的所有对点的平均对数协方差；$\overline{C}e(V_\alpha,V)$ 代表矢量 h 的两个端点分别在样品支撑 V_α 和待估域 V 内的所有对点的平均对数协方差；μ 为拉格朗日乘子

InZ_V^* 的对数正态克立格方差是：

$$\sigma_{ke}^2 = \overline{C}e(V,V) - \sum_{\alpha=1}^{n}\lambda_\alpha \overline{C}e(V_\alpha, V) + \mu \tag{21}$$

(19)式中的 C 值是:

$$C = \frac{1}{2}\left\{\sum_{\alpha=1}^{n}\lambda_\alpha[\overline{C}e(V_\alpha, V_\alpha) - \overline{C}e(V_\alpha, V)] - \mu\right\} \tag{22}$$

则 Z_V 的无偏最优线性估计量 Z_V^* 是:

$$Z_V^* = e^{\sum_{\alpha=1}^{n}\lambda_\alpha\{[\ln(X_\alpha)+\frac{1}{2}\overline{C}e(V_\alpha,V_\alpha)]-[\frac{1}{2}\overline{C}e(V_\alpha,V)+\frac{1}{2}\mu]\}} \tag{23}$$

Z_V^* 的克立格方差 σ_K^2 是:

$$\sigma_K^2 = Z_V^{*2}[e^{Ce(V,V)} + e^{\sum_{\alpha\beta=1}^{n}\lambda_\gamma\overline{C}(V_\alpha,V_\beta)} - 2e^{\sum_{\alpha=1}^{n}\lambda_\alpha\overline{C}(V_\alpha,V)}] \tag{24}$$

Z_V^* 也可根据 $\ln Z_V^*$ 值及估计该值的样品数 n 及估计方差 σ_e^2 值利用 H. S. Sichel 给出的估计因子表由(25)式求出:

$$Z_V^* = e^{\ln Z_V^* \cdot \gamma_n(\sigma_e^2)} \tag{25}$$

式中 $\gamma_n(\sigma_e^2)$ 是估计因子表给出的系数,而 σ_k^2 也可根据 σ_{ke}^2 利用 H. S. Sichel 表求出估计精度的上、下限表示之。

当未采用对数正态克立格法对全矿床进行局部估计时,在大子样的情况下,可用下式求出全矿床的平均品位:

$$Z = e^{[Z_e+\frac{1}{2}\overline{C}e(D,D)\pm\frac{1.65}{\sqrt{n}}]} \tag{26}$$

式中 $\overline{C}e(D,D)$ 是定义于全矿床 D 上的 $y = \ln(X_\alpha)$ 的方差, Z_e 是 $\ln(X_\alpha)(\alpha = 1,2,\cdots,n)$ 的平均值,n 是样品数

如果品位值服从三参数对数正态分布时,可用下式求出全矿床的平均品位:

$$Z = e^{[Z_e+\frac{1}{2}\overline{C}e(D,D)\pm\frac{1.65}{\sqrt{n}}]} - a \tag{27}$$

式(27)中 a 是第三参数,其他符号同式(26)。

以上讨论均是在对数正态守恒假设下进行的,即:当样品值呈对数正态分布时,其样品值的平均值也呈对数正态分布,而且它们的联合分布也保持对数正态分布。

3、**切尾法**(trimmed means)

该法是去掉几个特高值及特低值,再用剩余数据计算平均值,即为"α 一切尾平均值",α 是两边去掉的特高值及特低值所占全部数据的比例,这时,切尾平均值 m (α)为:

$$m(\alpha) = \frac{1}{h}\sum_{i=g+1}^{n-g} Xi \tag{28}$$

(28)式中:$g = n\alpha$ (29)

$h = n-2g$ (30)

例如,有一个样品数 n = 10 的一组数据如下:

0.3, 0.4, 0.1, 0.2, 0.5, 0.6, 0.7, 0.6, 0.4, 0.5,

按升值将其排序为：

0.1，0.2，0.3，0.4，0.4，0.5，0.5，0.6，0.6，0.7

当两边各去掉 2 个数值时，$\alpha = 2/10 = 0.2$，按(29)式及(30)式，$g = 10\times0.2 = 2$，$h = 10-2\times2 = 6$

则 $m(0.2) = \frac{1}{6}\sum_{i=3}^{8}(Xi) = 0.45$

当我们用全部 10 个数据时，其平均值 m' = 0.43，从理论上来讲 m (0.2)比 m'更为稳健，即该值受特异值的影响为小。

切尾法在生产实践中经常使用，其优点是计算简便。

4、**顺序—秩统计量**(order—Rank Statistics)

设数据样本为：

$X_1, X_2, \cdots, X_2, \cdots X_n$

按升值排序如下，称之为顺序统计量：

$X_{(1)}, X_{(2)}, \cdots, X_{(i)}, \cdots, X_{(n)}$

其中(i)是对应于顺序统计量的秩。显然：

$X_{(i)} \leqslant X_{(i+1)}$

$X_{Min} = X_{(1)}$ （最小顺序统计量）

$X_{MAX} = X_{(n)}$ （最大顺序统计量）

一个观测值的秩就是它在顺序统计量中的位置，记为：$q(X\alpha) = i$，这 n 个顺序统计量将母体分成 n+1 个部分，每一部分的区间为：

$\{X_{(i)}, X_{(i+1)}\}$

平均而言，每一区间包含母体中的相等部分，且均为 1/n，秩 i 在区间 (0，n+1) 内服从均匀分布。在上述顺序统计量中，样品的中位数 (0.5 分位数) 就是中心位置，而样品的变化范围 $[X_{(n)} - X_{(1)}]$ 是其离散性的测度。

显然，该法的实质是通过研究顺序统计量和秩来研究原始数据的。

参考文献

[1]侯景儒，黄竞先.地质统计学及其矿产储量计算中的应用，地质出版社，北京，1982

[2] A. G. Journel 等(侯景儒，黄竞先译)，矿业地质统计学，冶金工业出版社 北京，1982

[3]侯景儒、张树泉、黄竞先，对数正态克立格法及其在矿石储量估计中的应用，北京科技大学学报，No. 5，1989

[4]侯景儒，指示克立格法的理论及方法，地质与勘探，1990

[5] D. M. Hawkins，Identification of outliers，Chapman and Hall，London，1980

[6] D. G. Krige，Studies of fhe effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body，Math. Geol，14 (6) 1982

[7] A. G. Journel，A. Arik，Dealing with outlier high grade data in precious metals deposits. proceedings of NATO ASI，GEOSTAT TAHOE 83，Reidel Publishing co.，1983

[8] Cressie，N.，Hawkins，D. M.，Robust estimation of the variogram，《Math. Geol.》，vol. 12，1980

IDENTIFICATION AND HANDLING OF OUTLIERS AS WELL AS ESTIMATION OF MEAN GRADE OF BOLCKS

Hou Jingru, Zhang Shuquan, Huang Jingxian

(Depart. of Geology, Beijing University of Sciences and technology)

Abstract

Outliers problems may be present in sampling surveys. Outliers, called extreme high grade data in geological prospecting and mining engineering, usually correspond to very high grade mineralization. So, outliers is important to geologic studing, ore reserves calculation and data processing. The paper dealt mainly with four questions:

1) Concept of outliers and extreme high grade;

2) statistical meaning of outliers;

3) identification and handling for outliers (including: estimation neighbourhood method and influence coefficient method);

4) some nonparametric statistical methods are used for data with ortliers and estimation of mean grade of blocks (including: indicator kriging, lognormal kriging, trimmed means and order—rank statistics).