

刍议地球化学勘查中的采样误差

孙 宝 田

(有色金属总公司吉林矿产地质研究所)

本文对勘查地球化学采样误差作了理论探讨。将采样误差分为两次采样的均方差,两次采样有不同的方差;两次采样存在的协方差。指出对多元素测定结果,应考查所有被测元素才能对采样质量作出客观评价。

关键词: 勘查地球化学; 采样误差; 采样质量

概 述

在地球化学勘查中,正确地采集样品对发现和评价地球化学异常至关重要。通常采样的质量是由重复采样后进行化验分析得到的结果评估,实现这个过程包括采样、加工、化验分析三个环节,而其中每个环节都可能带来误差,这就给评估采样质量带来种种复杂问题。

为评估采样质量,在冶金和有色地质系统通常是生产人员野外采样之后的二、三天内,由质量检查人员再重复采取一定数量(约占总工作量的2~5%)的样品,作为评估生产人员采样质量的标准。若两次采取的样品经加工和化验分析得到的结果符合合格率要求,则认为采样工作质量可靠;否则认为采样质量不合格。而化验分析的工作质量则由密码抽查和内验结果来考核。这种做法的缺点是:①若质量检查人员不及时进行重复采样,有时会造成与生产采样的条件不同(如山洪暴发前后采集水系底沉积物样品),会出现诸如采样位置、样品性质等方面的差异。即使生产和质量检查人员都认真地按采样技术规定采集样品,也难以避免上述情况的发生。一旦出现差错,孰是孰非难以客观地评定。②计算采样合格率的公式为 $(A-B)/[(A+B)/2] \times 100\%$ 。式中A、B表示两

次采样某元素的含量测定结果。但不合格界限的划定没有建立在严格的概率论及数理统计学基础上,因而不知道做出某种判断(采样质量合格还是不合格)的信度如何。③如果样品进行了多元素测定,是所有元素的合格率都符合要求才认为采样合格,还是大多数元素测定结果合格就认为采样合格?如果采用后一种作法的话,有多少元素合格才能认为是大多数元素合格?目前对这个问题尚未有得到令人满意的解决。

地矿部系统在开展区域化探扫面工作时,是由不同人不同时进行一定数量的重复采样,将两次采到的样品分别进行两次化验分析。这样,在一个点得到四个数据,然后用“三层套合方差分析”方法,同时给出对采样与化验分析误差的估计。这样做能及时评估采样及化验分析质量的好坏,避免了分别考察采样、化验分析质量造成的耽误时间、抽取密码缺乏代表性(多注重抽取异常样,忽视背景样品的抽取)等弊病。但是,“三层套合方差分析”方法也有不足之处,即同样不能用于讨论多元素测定情况下采样及化验分析质量,且不能说明采样误差的性质,因而无法讨论减少采样误差的途径。需要指出一点:包括“三层套合方差分析”方法在内的各种方差分析方法要求所研究的各个变量是相互独立的,事实上大多数情况下很难满足

这个前提。

针对上述情况,下面对有关勘查地球化学采样误差问题做些理论探讨。

误差分析

从整体而言,地质体本身是不均匀的,在其不同部位(如矿化及无矿地段)成矿元素的含量是不同的。因此人们能通过采样确定找矿有望地段。但是,从组成地质体的每一个小局部而言,采样人又希望在采样之处一定大小范围内的样品相对均匀。否则又何以能保证样品的代表性!所以采样工作是从一个整体上不均匀的地质体的局部,采取相对均匀的样品。针对这个特点,就要知道该地质体中元素含量在整体上的变异性及局部上的平稳性。当事先不能得知这些信息时,则必需通过实践来获取。人们经常根据所研究的地质体中可能存在的矿化类型、矿化规模等特点,先验地确定采样网度、每件样品的形态特征和重量。显然,这很大程度上带有主观因素,往往给样品的客观性带来不良影响。为了使采样更加合理,应先进行实验工作。

众所周知,所谓采样误差,是所采样品化验分析得到的测定值与真值之间的差别。通常,真值是人们所无法确知的,只有重复多次地对同一采样对象进行等精度测量时,其平均值才能接近真值,因此,可以有条件地把这个平均值视作具有一定精度的真值。在勘查地球化学找矿实践中较多的采用“点采样”的方法,目的在于通过大量采样点获取的样品分别进行化验,求出测区中成矿元素(或找矿指示元素)的背景含量,确定异常下限,进而圈定有望异常。所以既没有必要,也没可能求得每个样品中元素含量的真值,只能利用有限的测定值去讨论上述问题,讨论采样质量时也是如此。这样做是否可靠?采样质量如何?通常用方差分析方法来回答。在这方面已有不少文章介绍,本文则拟从另

一个角度来讨论这个问题。

假定在第 i 个采样点先后采取一个生产样品和一个质量检查样品,设生产样品中某元素的含量为 C_{oi} ;其真值为 C_i ; C_{ki} 为质量检查样品中该元素的含量。那么生产样品和质量检查样品同真值间的误差分别为 $(C_{oi}-C_i)$ 和 $(C_{ki}-C_i)$;质量检查样品同生产样品之间的差值为 $(C_{ki}-C_{oi})$,则有下式成立。

$$(C_{ki}-C_{oi})=(C_{ki}-C_i)-(C_{oi}-C_i) \quad (1)$$

从(1)式可以看出,两次采样的误差是由两次采样各自同真值的误差之差构成。因此只要计算 $C_{ki}-C_{oi}$ 即可估计两次采样的误差,而不必关心采样点真值 C_i 的大小。当测定了不同采样点上的 C_{ki} 及 C_{oi} ,并分别计算出 $C_{ki}-C_{oi}$,将得到一个随机数组。若在直角坐标系 XOY 平面内, X 轴表示生产样品 C_o , Y 轴表示质量检查样品 C_k 。如采样质量合格, C_o 与 C_k 之间应存在较严格的线性关系,大多数采样点应该在直线 $C_k=C_o$ 上或其附近,不会偏离太远。凡是不在这条直线上的点,都有一个非零的采样误差。

$$\text{设 } U^2 = \frac{1}{n} \sum_{i=1}^n (C_{ki}-C_{oi})^2 \quad (2)$$

U^2 可作为估计采样误差的综合标志, U^2 越小则两次采样误差越小;反之亦然。但是,仅从 U^2 的大小并不能揭示产生误差的原因,为了解决这个问题,需将 U^2 分解成几部分。可以证明有下式成立(具体推导从略)。

$$U^2 = (\bar{C}_k - \bar{C}_o)^2 + (S_k - S_o)^2 + 2S_k S_o (1-\gamma) \quad (3)$$

$$\text{式中, } \bar{C}_k = \frac{1}{n} \sum_{i=1}^n C_{ki}$$

$$\bar{C}_o = \frac{1}{n} \sum_{i=1}^n C_{oi}$$

$$S_k^2 = \frac{1}{n} \sum_{i=1}^n (C_{ki} - \bar{C}_k)^2$$

$$S_o^2 = \frac{1}{n} \sum_{i=1}^n (C_{oi} - \bar{C}_o)^2$$

$$\gamma = \frac{\frac{1}{n} \sum_{i=1}^n (C_{ki} - \bar{C}_k)(C_{oi} - \bar{C}_o)}{S_k S_o}$$

$\bar{C}_k$ 、 $\bar{C}_o$ 、 S_k^2 、 S_o^2 的意义很清楚，不必赘述， γ 是相关系数。

从 (3) 式可以看出，两次采样的误差 U^2 被分解成三部分。显然第一部分 $(\bar{C}_k - \bar{C}_o)^2$ 系由两次采样在平均值上的差异所引起。第二部分 $(S_k - S_o)^2$ 系由两次采样有不同的方差造成的。而第三部分是由两次采样存在协方差引起的。

将 (3) 式两端除以 U^2 后经整理得到下式：

$$\frac{(\bar{C}_k - \bar{C}_o)^2}{U^2} + \frac{(S_k - S_o)^2}{U^2} + \frac{2S_k S_o (1 - \gamma)}{U^2} = 1 \quad (4)$$

令： $u_m = (\bar{C}_k - \bar{C}_o)^2 / U^2$

$u_s = (S_k - S_o)^2 / U^2$

$u_c = 2S_k S_o (1 - \gamma) / U^2$

则 (4) 式可变为：

$$u_m + u_s + u_c = 1 \quad (5)$$

u_m 、 u_s 和 u_c 分别称为“偏移得分”、“方差得分”和“协方差得分”。如果 u_m 很大，则表示两次采样的平均值彼此强烈地偏离，

很可能是第一次采样不当造成样品测定值的系统误差。为了弥补这一缺陷，要么提醒采样人员注意采样质量，要么有条件地引入“校正系数”来纠偏。因为方差 S_k^2 或 S_o^2 是表示样品中某元素测定值相对于平均值的离散程度，若两次采集的样品性质相同，它们应有相近的方差，否则在方差上也应表现出较大的差别。故当 u_s 很大时，则意味着两次所采样品存在着较大的差别。与 u_m 联系起来考虑，甚至可以推测两次采样来自不同的总体。协方差得分 u_c 提供这样一种信息，即企图靠引入校正系数来修正平均值还是不够的。从误差性质而言，可分为系统误差和偶然误差。后者在一般情况下很难避免，故人们的注意力应集中于避免系统误差上。就一般情况而言，若 $u_c > (u_m + u_s)$ ，则可认为两次采样没有明显的系统误差，其误差主要是偶然误差所致，可认为生产中所采样品是合格的。

如下表列出了在新疆某地开展分散流找矿时两次采样铜的分析结果。

经计算得： $\bar{C}_k = 36.0$ ； $\bar{C}_o = 35.4$ ； $S_k = 14.53$ ； $S_o = 15.70$ ； $\gamma = 0.8943$ ； $U^2 = 49.93$ 。进而求得 $u_m = 0.007$ ； $u_s = 0.027$ ； $u_c = 0.966$ 。

由计算结果可以看出，无论是偏移得分 u_m ，还是方差得分 u_s ，与协方差得分 u_c 相比都是微不足道的。因此认为两次采样化验结

新疆某地水系底沉积物两次采样Cu的含量 (ppm)

样号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C_{ki}	64.6	29.7	35.3	33.7	33.4	34.8	28.4	20.9	22.9	22.9	37.1	31.4	29.0	44.7	25.2	22.1
C_{oi}	45.7	25.7	61.9	37.0	35.4	33.7	30.3	22.8	20.6	22.4	35.2	31.7	27.5	41.2	23.7	26.1

样号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
C_{ki}	42.9	22.5	10.8	36.6	24.1	29.1	100.1	50.7	33.9	53.7	35.7	46.9	38.5	33.1	26.4	31.7
C_{oi}	45.4	25.3	26.2	32.3	23.0	32.3	93.3	50.2	30.6	51.7	36.5	54.5	43.4	30.3	22.2	32.8

果所得误差 U^2 ，主要是由协方差引起，而两次采样各自的平均值和方差无很大差别，故

认为生产采样质量合格。

式 (3) 是对 U^2 的对称分解方法，即无

论是对 $\bar{C}_k$ 、 $\bar{C}_o$ ，还是对 S_k 、 S_o ，都是对称的，事实上， U^2 还有其他分解方法，式(6)给出了 U^2 的一种不对称分解方法(推导步骤略)。

$$U^2 = \frac{1}{n} \sum_{i=1}^n (C_{ki} - C_{oi})^2 \\ = (\bar{C}_k - \bar{C}_o)^2 + (S_k - \gamma S_o)^2 + (1 - \gamma^2) S_o^2 \quad (6)$$

将(6)式两端除以 U^2 ，整理后得到：

$$\frac{(\bar{C}_k - \bar{C}_o)^2}{U^2} + \frac{(S_k - \gamma S_o)^2}{U^2} + \frac{(1 - \gamma^2) S_o^2}{U^2} = 1 \quad (7)$$

从(7)式可以看出，等式左端第一项就是偏离得分 u_m ；等式左端第二项我们称作“回归得分”，用 u_R 表示；左第三项称作“扰动得分”，用 u_D 表示。

u_D 表示 U^2 中非系统误差所起的作用，这部分非系统误差是用回归分析所不能处理的；而 u_R 则表示了用回归分析能弥合的误差所占的比例。现仍以前表所列的数据求出各种得分： $u_m = 0.007$ ； $u_R = 0.0048$ ； $u_D = 0.9885$ 。由计算结果可见， u_D 很大而 u_R 很小，表明两次采样的系统误差很小，可以认为两次采样的结果都是可靠的。

我们通过将总误差进行分解的方法了解所产生误差的性质，进而结合实例对两次采样结果进行了讨论，并假定第二次所采样品(质检采样)是可靠的。那么，两次采样究竟是否采自同一总体(如来自同一总体，则认为采样质量合乎要求)，则要进行假设检查，通常用 t 检验来检查两种样本的平均值是否有显著差异。其原理在有关数理统计的

教科书中已有详述，本文不再重复，下面只给出运算的步骤及结果。

假若每次采样的测定结果都是遵循正态分布，统计量 t 可用下式求得：

$$t = \frac{|\bar{C}_k - \bar{C}_o|}{\sqrt{S_k^2 + S_o^2}} \sqrt{n} \quad (8)$$

统计量 t 服从自由度为 $f = 2n - 2 = 64 - 2 = 62$ 的Student氏分布，用上面计算的数据得：

$$t = \frac{(36.0 - 35.4)}{\sqrt{14.53^2 + 15.70^2}} \sqrt{32} = 0.16$$

取信度 $\alpha = 0.05$ ，查 t 分布表得知，当自由度 $f = 62$ 时， t 统计量的临界值为：

$$t_{2\alpha} = t_{0.1} = 1.67$$

因为 $t = 0.16$ ，显然小于临界值 $t_{2\alpha} = 1.67$ ，故判两次采样的平均值无显著差异。

比较两次采样的均方差是否有显著差异，可用 F 检验回答。问题归结于检验假设 $H_0: S_k^2 = S_o^2$ ，取信度 $\alpha = 0.05$ ，否定域则为

$$F = S_o^2 / S_k^2 > F_{0.05}$$

对两个自由度皆为 $f_1 = f_2 = n - 1 = 32 - 1 = 31$ 的情况，查 F 分布表得 F 的临界值 $F_{0.05} = 1.84$ ，统计量 $F = S_o^2 / S_k^2 = 15.70^2 / 14.53^2 = 1.16$ ，小于 $F_{0.05} = 1.84$ ，落在肯定域中，认为两者有相同的均方差。

统计检验结果表明，两次采样有相同的平均值和均方差，故可推论两次采样来自同一总体。因而可认为采样质量可靠。

事实上，每个勘查地球化学样品经常要做多元含量测定，这就需将上面对单元素测量结果的讨论推广到多元的情况。若用统计检验方法只要将一元正态分布的 t 检验方法代之以多元正态总体的广义 T^2 检验方法即可。

A Preliminary Discussion on Sampling Error in Geochemical Survey

Sun Baotian

A theoretical discussion on sampling error in geochemical survey is made in this paper. A sampling error consists of three parts: the mean value error, variance and covariance of two successive sample collections. It is indicated that for a multi-element geochemical survey the measuring results of each element should be taken into consideration in order to get an objective evaluation of the sampling quality.