

多级(系统)分组方差分析 用于化探数据质量评价

沈 杰

(安徽省地质矿产局)

对化探数据的影响因素有采样点、采样、和化验分析,当采样点对数据影响显著,而采样和化验分析对其影响不显著时,表明该批数据真正地反映了元素在某地区的起伏变化情况,反之,则说明后两个因素造成数据失真。

关键词: 多级(系统)分组方差分析;化探数据质量评价

问题的提出与分析

在使用某地区的一批化探数据时,首先要对这批数据进行质量评价,即了解所用数据的误差是否合乎要求,如数据严重失真而不能使用,则要进一步了解造成数据失真的原因。

根据获取化探数据的工作过程,可以依次排出3个对数据有影响的因素:定点→采样→化验分析。在这些因素中,前者影响后者,而后者不影响前者。在不同的采样点上所能采集到的样品是不相同的;而在同一个采样点上所采集的样品也有些差异,虽然这种差异受到采样点约束;最后,同一块样品经多次化验分析得出的结果数据也不尽相同。

在以上3个因素中,由于采样点的不同而使数据产生差异是正常的,因为这种差异正好反映了该元素随采样点的变化而变化起伏。同一采样点内不同样品之间有差异也是正常的,但这种差异不应该很大,如果它大于因采样点不同而引起的变化,可能说明采样工作中存在着问题(如采样位置选择不当,混采新鲜和久经风化的岩石,采样人员的素质不高等)。化验分析产生的差异往往是些无法控制的因素,如仪器性能的波动、操作人员的情绪状况和工作态度等造成的,对一批质量较高的化探数据来说,这种差异应

该很小。

由此看来,对某一地区化探数据的质量评价就是要了解两个问题:数据中由采样点的不同而产生的差异是否大于由采样工作和化验分析工作所造成的差异?造成数据失真的主要原因是采样工作还是化验分析工作?

从方差分析的角度看,我们可以把采样点看作因素A,采样工作看作因素B,把化验分析中产生的差异看作是随机误差。以上两个问题可以看作是检验A因素和B因素对数据是否有显著影响。如果A因素的影响显著,则说明这批数据真实地反映了该元素在这个地区的起伏变化情况。如果A因素的影响不显著,B因素的影响显著,则意味着采样工作是造成数据失真的主要原因;反之,则说明化验分析工作是造成数据失真的主要原因。

多级(系统)分组方差分析

在方差分析方法中,多级(系统)分组方差分析适用于解决多种因素之间存在着单向套合关系的问题。

设,在某一地区选定 a 个采样点,每个采样点上进行了 b 次采样(即采集了 b 个样品),每个样品做了 c 次化验分析,得出某一元素的数据 $X_{ijk}(i=1,2,\dots,a,j=1,2,\dots,b,k=1,2,\dots,c)$,它表示第 i 个采

样点上第 j 个样品的第 k 次分析数值。

可以认为化验数据是随机变量, 采样点和采样工作对数据均有影响, 所以可以把不同采样点的不同样品看作来自不同的正态母体 (共有 ab 个), 而把每次化验分析的结果 $X_{ijk} (k=1, 2, \dots, c)$ 看作是这 ab 个母体中抽得一个容量为 c 的子样。因此, 推断因素 A 和因素 B 的影响是否显著就化为判断这 ab 个母体是否有相同的分布。在获取数据的过程中, 除了我们关心的因素采样点和采集的样品外, 化验分析条件总是尽可能地一致, 如同一个样品在同一个人的操作下, 用同一台仪器来完成 c 次分析。这就使我们认为每个母体的方差是相同的。因此, 问题就化为检验具有相同方差的正态母体其均值是否相等。

可以设:

$$X_{ijk} \sim N(u_{ij}, \sigma^2)$$

$$X_{ijk} = u + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

其中 $i=1, 2, \dots, a; j=1, 2, \dots, b; k=1, 2, \dots, c; u$ 是该地区的总体平均值, α_i 表示因素 A 对数据的影响, $\beta_{j(i)}$ 表示在因素 A 的第 i 个水平下因素 B 对数据影响, $\varepsilon_{ijk} \sim N(0, \sigma^2)$ 是随机误差项, 是由化验分析中的未能控制的因素和其他因素造成的。

作统计假设:

$$H_{01}: \alpha_i = 0 \quad i=1, 2, \dots, a$$

$$H_{02}: \beta_{j(i)} = 0 \quad j=1, 2, \dots, b$$

检验子样 X_{ijk} 母体的均值是否相同, 从统计角度看, 就是要检验 H_{01} 和 H_{02} 是否成立。

$$\bar{X}_{ij} \triangleq \frac{1}{c} \sum_{k=1}^c X_{ijk}$$

$$\bar{X}_{i..} \triangleq \frac{1}{b} \sum_{j=1}^b \bar{X}_{ij}$$

$$= \frac{1}{bc} \sum_{j=1}^b \sum_{k=1}^c X_{ijk}$$

$$\bar{X} \triangleq \frac{1}{a} \sum_{i=1}^a \bar{X}_{i..}$$

$$= \frac{1}{abc} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c X_{ijk}$$

数据总变差

$$SS_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X})^2$$

应该反映了数据总的变动, 根据前面的分析, 它包含有因素 A 和因素 B 以及随机因素造成的差异。下面将 SS_B 分解。

$$\begin{aligned} SS_B &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{X}_{i..} - \bar{X})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ij.} - \bar{X}_{i..})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij.})^2 \\ &\quad + 2 \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{X}_{i..} - \bar{X}) \\ &\quad \times (X_{ijk} - \bar{X}_{i..}) \\ &\quad + 2 \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij.}) \\ &\quad \times (\bar{X}_{ij.} - \bar{X}_{i..}) \end{aligned}$$

可以证明上式右边最后两项分别等于零 (略), 并记:

$$SS_1 \triangleq \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{X}_{i..} - \bar{X})^2$$

$$= bc \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2$$

$$SS_2 \triangleq \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{X}_{ij.} - \bar{X}_{i..})^2$$

$$= c \sum_{i=1}^a \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..})^2$$

$$SS_3 \triangleq \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij.})^2$$

那么 $SS_B = SS_1 + SS_2 + SS_3$

在 SS_1 中, $(\bar{X}_{i..} - \bar{X})^2$ 表示了第 i 个采

样点上的数据平均值与全体平均值的偏差,

$\sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2$ 就是把所有采样点上的这种偏差累加起来, 所以 ss_1 反映了采样点 (因素A) 对数据的影响。

在 ss_2 中, $(\bar{X}_{ij.} - \bar{X}_{i..})$ 表示了第 i 个采样点上第 j 个样品的数据平均值与这个采样点上的数据平均值的偏差,

$$\sum_{i=1}^a \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..})^2$$

就是把所有采样点上的所有样品, 与各自采样点上的数据平均值的偏差累加起来, 所以 ss_2 反映了采样工作 (因素B) 对数据影响, 而这种影响是受限于采样点的。

在 ss_3 中, $(X_{ijk} - \bar{X}_{ij.})^2$ 表示了第 i 个采样点上的第 j 个样品的第 k 次化验分析值与第 i 个采样点上第 j 个样品的数据平均值的偏差,

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij.})^2$$

就是把全体数据中的所有这种偏差累加起来, 所以 ss_3 反映了化验分析工作 (随机误差) 对数据影响, 这种影响受限于采样点和采样工作。

从直观上可以大致看出, 若 ss_1/ss_2 较大, 则可能反映了采样点 (因素A) 的影响显著。反之, 则可能说明采样工作 (因素B) 和化验分析 (随机误差) 的影响显著。如果 ss_2/ss_3 较大, 则可能说明采样工作 (因素B) 的影响显著, 误差主要是采样工作造成的。反之, 则说明化验分析的影响显著, 是造成误差的主要原因。

直观上的看法是粗糙的, 以下进行数学推证。

定理1: 设总体 ξ 服从正态分布 $N(u, \sigma^2)$, 则样本平均值 $\bar{X}$ 服从正态分布 $N(u, \sigma^2/n)$ 。

定理2: 设总体 ξ 服从正态分布 $N(u, \sigma^2)$ 则样本平均值 $\bar{X}$ 与样本方差 s^2 相互独

立, 并且统计量 ns^2/σ^2 服从自由度为 $n-1$ 的 χ^2 分布。

如果 H_{01} 和 H_{02} 成立, 则 X_{ijk} 都来自同一个正态母体。由定理1知:

$$\bar{X}_{i..} \sim N(u, \sigma^2/bc)$$

由定理2知:

$$\begin{aligned} \frac{ss_1}{\sigma^2} &= bc \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2 / \sigma^2 \\ &= \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2 / (\sigma^2/bc) \sim \\ &\quad \chi^2(a-1) \end{aligned}$$

由定理1知:

$$\bar{X}_{ij.} \sim N(u, \sigma^2/c)$$

由定理2 又知:

$$\begin{aligned} c \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..})^2 / \sigma^2 \\ &= b \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..})^2 / (\sigma^2/c) \\ &\sim \chi^2(b-1) \end{aligned}$$

再由 χ^2 分布的可加性得

$$ss_2/\sigma^2 \sim \chi^2[a(b-1)].$$

由定理2 知, 对每个 i 和每个 j

$$\sum_{k=1}^c (X_{ijk} - \bar{X}_{ij.})^2 / \sigma^2 \sim \chi^2(c-1)$$

再由 χ^2 分布的可加性得

$$ss_3/\sigma^2 \sim \chi^2[ab(c-1)].$$

同理可知: $ss_B/\sigma^2 \sim \chi^2(abc-1)$ 。

因此, 作统计量

$$F_1 = \frac{ss_1/\sigma^2/(a-1)}{ss_2/\sigma^2/a(b-1)}$$

$$= \frac{ss_1/(a-1)}{ss_2/a(b-1)}$$

$$F_2 = \frac{ss_2/\sigma^2/a(b-1)}{ss_3/\sigma^2/ab(c-1)}$$

表 1

变差源	平方和	自由度	均方	F 值
采样点间	$ss_1 = \sum_{i=1}^a \left(\sum_{j=1}^b \sum_{k=1}^c X_{ijk} \right)^2 / bc - \left(\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c X_{ijk} \right)^2 / abc$	$a-1$	$V_1 = ss_1 / (a-1)$	$F_1 = V_1 / V_3$
样品间 (采样工作)	$ss_2 = \sum_{i=1}^a \sum_{j=1}^b \left(\sum_{k=1}^c X_{ijk} \right)^2 / c - \sum_{i=1}^a \left(\sum_{j=1}^b \sum_{k=1}^c X_{ijk} \right)^2 / bc$	$a(b-1)$	$V_2 = ss_2 / a(b-1)$	$F_2 = V_2 / V_3$
化验分析	$ss_3 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c X_{ijk}^2 - \sum_{i=1}^a \sum_{j=1}^b \left(\sum_{k=1}^c X_{ijk} \right)^2 / c$	$ab(c-1)$	$V_3 = ss_3 / ab(c-1)$	
总变差	$ss_{总} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c X_{ijk}^2 - \left(\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c X_{ijk} \right)^2 / abc$	$abc-1$		

$$= \frac{ss_2/a(b-1)}{ss_3/ab(c-1)}$$

显然 $F_1 \sim F[a-1, a(b-1)]$

$$F_2 \sim F[a(b-1), ab(c-1)]$$

对统计量运用 F 检验法。给定一显著水平, 由 F 分布表查得临界值 F'_1 和 F'_2 。若 $F_1 > F'_1$, 则拒绝 H_{01} , 即采样点(因素 A)对数据有显著影响。反之, 采样点对数据的影响不显著, 这时再看如果 $F_2 > F'_2$, 则拒绝 H_{02} , 即因素 B(采样工作)对数据有显著影响, 它是造成误差的主要原因, 反之, 化验分析(随机因素)对数据有显著影响, 它是造成误差的主要原因。

实际上, 从数学上可以证明(因证明的过程较复杂且所用篇幅较大, 略去)

ss_1 是 $\sigma^2 + bc\sigma_A^2 + c\sigma_{B(A)}^2$ 的无偏估计;

ss_2 是 $\sigma^2 + c\sigma_{B(A)}^2$ 的无偏估计;

ss_3 是 σ^2 的无偏估计。

即

$$E(ss_1) = \sigma^2 + bc\sigma_A^2 + c\sigma_{B(A)}^2$$

$$E(ss_2) = \sigma^2 + c\sigma_{B(A)}^2$$

$$E(ss_3) = \sigma^2$$

其中

$$\sigma_A^2 = \sum_{i=1}^a (u_i - u)^2 / (a-1)$$

$$\sigma_{B(A)}^2 = \frac{\sum_{i=1}^a \sum_{j=1}^b (u_{ij} - u_i)^2}{a(b-1)}$$

σ_A^2 表示因素 A 的影响, $\sigma_{B(A)}^2$ 表示 B 因素的影响, $E(ss_1/ss_2)$

$$= E(ss_1) / E(ss_2)$$

$$= 1 + bc\sigma_A^2 / (\sigma^2 + c\sigma_{B(A)}^2)$$

可见当 H_{01} 不成立即因素 A(采样点)对数据的影响显著时, ss_1 与 ss_2 之比的期望值有偏大的趋势。所以当 $F_1 > F'_1$ 时就拒绝 H_{01} , 反之接受 H_{01} 。

$$E\left(\frac{ss_2}{ss_3}\right) = \frac{E(ss_2)}{E(ss_3)} = 1 + c\sigma_{B(A)}^2 / \sigma^2$$

可见当 H_{02} 不成立即因素 B(采样工作)对数据的影响显著时, ss_2 与 ss_3 之比的期望值有偏大的趋势。所以当 $F_2 > F'_2$ 时就拒绝 H_{02} , 反之接受。

在具体的计算中, 可以将求 ss_1 、 ss_2 和 ss_3 及 $ss_{总}$ 简化为表 1 中形式, (简化步骤略)。方差分析计算可参照表 1 表 2 进行。

运用实例

在安徽省 H 地区的 1:5 万化探工作中, 为了研究该地区的地质情况, 在地区内选定了 57 个采样点 ($a=57$), 在每个采样点上各取了两块样品 ($b=2$), 每块样品各作了两次化验分析 ($c=2$), 得出了 Cu 等十个元素的化探数据。为了评价这批数据的质量, 对

每个元素的数据作了系统分组方差分析。用BASIC语言编制了计算程序,在PIED PIER微机上完成整个计算。

下面是元素Cu、Cr、V的方差分析结果(表3、4、5)。将数据取对数后再计算。取显著水平 $\alpha=0.005$,其临界值为 $F'_1=1.55$, $F'_2=1.43$ 。

表 2

层次	差 值	样本大小	方差分量
采样点	V_1-V_2	bc	$s_a^2 = (V_1-V_2)/bc$
采样工作	V_2-V_3	c	$s_b^2 = (V_2-V_3)/c$
化验分析	V_3	1	$s_v^2 = V_3$

Cu的方差分析结果 表 3

变差源	均 方	方差分量	F值
采样点间	$V_1 = 0.323502$	$s_a^2 = 0.0797817$	$F_1 = 73.95$ $F_2 = 2.23$
采样工作	$V_2 = 0.00437472$	$s_b^2 = 0.00120625$	
化验分析	$V_3 = 0.00196223$	$s_v^2 = 0.00196223$	

Cr的方差分析结果 表 4

变差源	均 方	方差分析	F值
采样点间	$V_1 = 0.00292751$	$s_a^2 = 0.00017212$	$F_1 = 1.31$ $F_2 = 1.22$
采样工作	$V_2 = 0.00223903$	$s_b^2 = 0.000198364$	
化验分析	$V_3 = 0.0018423$	$s_v^2 = 0.0018423$	

V的方差分析结果 表 5

变差源	均 方	方差分量	F值
采样点间	$V_1 = 0.00435856$	$s_a^2 = 0.000409956$	$F_1 = 1.60$ $F_2 = 1.11$
采样工作	$V_2 = 0.00271874$	$s_b^2 = 0.000132511$	
化验分析	$V_3 = 0.00245372$	$s_v^2 = 0.00245372$	

可以看出:

$$\text{Cu: } F_1 = 73.95 > 1.55 = F'_1$$

$$F_2 = 2.23 > 1.43 = F'_2$$

F_1 值远远超过临界值,说明采样点对数据的影响显著,即Cu元素在该地区中的变化起伏不致于被采样工作和化验分析的误差掩蔽或歪曲,因而Cu的数据是可利用的。 F_2 大于临界值说明采样工作的误差大于化验分析的误差。

$$\text{Cr: } F_1 = 1.31 < 1.55 = F'_1$$

$$F_2 = 1.22 < 1.43 = F'_2$$

F_1 值小于临界值意味着采样点对数据的影响不显著,说明Cr在该地区中的变化被采样工作和化验分析误差掩蔽了,Cr的数据存在着不可忽视的质量问题。而 F_2 值小于临界值意味着采样工作的影响不显著,从各层方差分量的数值来看,化验分析是造成数据失真的主要原因。

$$\text{V: } F_1 = 1.60 > 1.55 = F'_1$$

$$F_2 = 1.11 < 1.43 = F'_2$$

F_1 值大于临界值说明数据是可利用的。 F_2 值小于临界值说明采样工作误差并不大于化验分析误差。

参 考 文 献

- [1] 复旦大学:《概率论》(第二册第一分册),人民教育出版社,1979年。
- [2] 谢学锦:《区域化探》,地质出版社,1979年。
- [3] 沈恒范:《概率论讲义》(第二版),高等教育出版社,1985年。
- [4] 李茂年、周兆麟:《数学统计学》,天津人民出版社,1982年。
- [5] Hald, A.: Statistical Theory With Engineering Application.

Multistage (systematic) Grouping Variance Analysis for Quality Evaluation of Geochemical Exploration Data

Shen Jie

Factors that affect the quality of the geochemical exploration data are sample point, sampling procedure and geochemical analysis. Among these three factors if the first one affects the data more significantly while the other two are not, it shows that the data of this batch will certainly reflect the real variations of the elements measured in this district. On the contrary, it means that the data are distorted by the last two factors.