数理统计在地球化学探矿中的应用简介(一)

地　文

数理统计是应用数学的一个分支，它根据人们对偶然现象的统计规律的认识，制定出解决各种实际问题的具体统计方法。

近年来已有多种新的统计方法引用到化探工作中来，显著地提高了化探资料的整理和解释水平。本“简介”的目的，一方面是介绍数理统计的一些基本概念，另一方面是介绍几种统计方法在化探中的应用，包括：概率分布、统计推断、相关分析、判别分析和趋势分析等。为便于推广，这里只介绍用普通计算工具（如算盘、计算尺、手摇或电动计算机等）即能进行工作的方法。

在化探工作中应用数理统计的作用，在于更科学地、更充分地发掘出隐藏在原始数据中的经验规律，从而提高资料的利用率，取得更大的找矿效果。

但是用统计方法确定的经验规律，只是地质-地球化学作用的现象，而不是本质。所以应用统计方法时必须紧密结合地质-地球化学资料。这里包含两方面的含义：其一是指在进行资料收集、整理及编组时必须认真地考虑地质-地球化学背景；其二是指统计计算结果应尽可能地说明地质-地球化学问题。

一、概率分布

1.基本概念　为了建立一些最基本的统计概念，我们从具体事例出发。设有一袋研磨得极其均匀的试料。由一名分析人员用同一种分析方案对其中某元素的含量进行几次重复测定，得到n个数据：$x_1, x_2, x_3, \cdots\cdots x_n$。这n个测定结果不可能完全相等，但也不会相差太大。对于这样一批数据，统计处理的办法就是用下列公式求出两个特征数：

表1

样品号＼分析号	1	2	3	……	n
1	X_{11}	X_{12}	X_{13}	……	X_{1n}
2	X_{21}	X_{22}	X_{23}	……	X_{2n}
3	X_{31}	X_{32}	X_{33}	……	X_{3n}
⋮	⋮	⋮	⋮		⋮
m	X_{m1}	X_{m2}	X_{m3}	……	X_{mn}

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \cdots (1) \text{和} S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \cdots (2)$$

$\bar{X}$是试料中被测定元素真实含量的近似值，即算术平均值。S为标准误差，它反映分析质量的好坏。

再假设有一个花岗岩体，由岩体不同部位取得m个样品。每个样品都按上述方式重复分析n次。这样就得到了m×n个数据，列如表1。对于这样一套数据，统计处理就比上例复杂一些。但在化探实际工作中一般是不这样做的，而只对每个样品进行一次分析，所以对每一个样品只能得到表中某一行中的某一数值。例如表1中划有横线者，即只得到m个数据。可以想见，这m个数据之间的差别将会更大。如仍按上列的式(1)、式(2)计算，则所得的$\bar{X}$及S值将分别代表花岗岩中被测定元素的平均含量，以及该元素在岩

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

体中分布的不均匀程度（其中包括分析误差引起的含量起伏）。

从上述两个例子中，可对数理统计的一些基本概念（术语）作如下说明：

母体、子样、个体：统计研究对象的总体，叫做**母体**。例如试料中某种元素的真正含量，花岗岩体中某一元素的真实分布。从理论上讲，只有当n和m趋于无穷大时，才能真正认识它们。但实际上n和m都是有限的，所以人们的认识总是相对近似的。对母体取得的每一批实测数据，构成一个**子样**。每一个子样中都包含许多个实际观测数据，每一实测数据称作**个体**。子样中包含的个体数目，叫作**子样容量**，如第一例中容量为n，第二例中容量为m×n。粗略地说，容量大于30～50的子样，可以叫作大子样，而容量少于30～50的则称为小子样。数理统计就是根据子样的实际观测值对母体的特征进行各种估算和对比，从而得出较确切的统计规律。

频数、频率、概率：统计工作的第一步，就是把测得的含量分成若干间隔，统计落在每一间隔内的样品数，这就是**频数**。频数被样品总数除得的商以百分数(%表示)，即**频率**。当样品总数趋于无穷大时，频率围绕某一固定数值的摆动将越来越小，这个固定值叫作**概率**。概率就是含量落在某一给定范围内的可能性的数字度量。例如落在50～100γ/g范围内的概率为0.4，落在100～200γ/g范围内的概率为0.3，则落在50～200γ/g范围内的概率为0.4+0.3=0.7。由此可见，概率的大小与我们所取的含量范围有关。含量范围取得愈大，落在该范围内的概率也就愈大。概率随含量间隔增大而变化的规律，叫作**累积概率分布**，可用曲线或折线表示，也可用数学公式表示。图1为最常见的一种累积概率分布曲线。图中$x_{0.5}$及$x_{0.95}$是两个最常用的特征数，$x_{0.5}$称为50%分位数，即通常所说的中位数，它表示x值大于及小于此数的概率均为50%。$x_{0.95}$称为95%分位数，它表示x值小于此数的概率为95%，大于此数的概率仅为5%。平常把此值叫作信度5%的临界值。

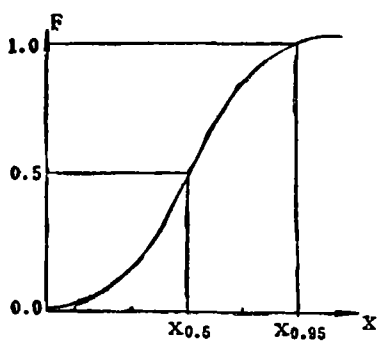

图1

2.正态分布 在数理统计中最重要、最基本的一种概率分布是正态分布，它的理论公式为：

$$P(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdots\cdots(3)$$

式中π及e为两个数学常数，分别等于3.14159及2.71828。γ及σ是两个参数，它们决定P(x)曲线。在数理统计中，γ称为**数学期望**，它决定曲线峯值的位置，即x=γ时P(x)达到极大值。σ称为**方差**，它决定曲线的平缓程度，即σ愈大，曲线愈平缓(图2)。

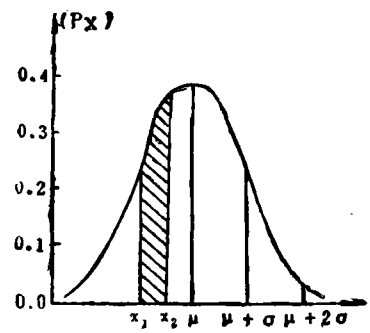

图2

P(x)曲线的意义在于：在任一给定的(x_1、x_2)范围内，曲线与横轴所包的面积，等于x落在该范围内的概率。由此可见，落在-∞μ到范围内的概率等于0.5。因此，μ就是前面所说的50%分位数。用积分分位数的方法可以计算出μ+σ等于84%分位数，μ+2σ等于97.7%分位数，μ+3σ等于99.9%分位数。推

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

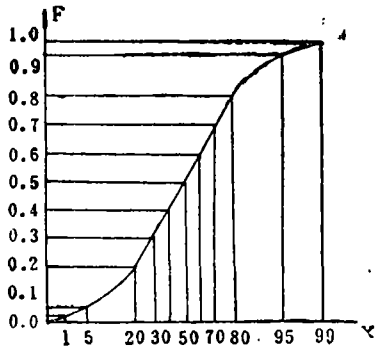

图3

而广之，任一百分比的分位数都可以由μ及σ两个参数值算出。所以可以简单的符号N(μ、σ)来表示正态分布。下面要介绍的几种统计方法其根据都是正态分布的这些性质。

3.概率格纸 正态分布的累积频率分布曲线，以普通比例尺绘出时如图3所示。如按图3中横座标所示的比例尺划分，则累积频率分布曲线将转化成一条直线(图4)。以这种不等距方式划分的座标纸，即为概率格纸。在概率格纸上进行统计作业比常用的频率直方图方法有不少缺点，还可以避免分组所引起的麻烦。为了便于精确地绘制概率格纸，现将有关数据列于表2。

利用表2制作概率格纸的步骤如下：取一张普通的方格纸，大小不限，在中间的厘米整数线上作一横线，注上“50%”，以此线为基准向上、下量取表2所示距离（距离比例尺依所取纸张大小而定），注上相应的百分率，并引出平行线即成。于概率格纸横座标的比例尺没有什么限制，可用普通比例尺，也可用对数比例尺。

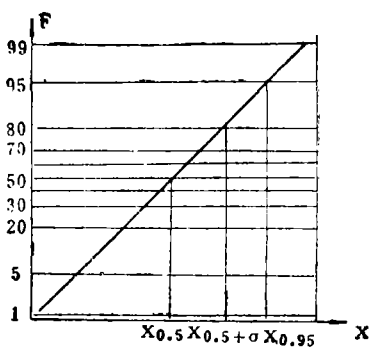

图4

表2

累积频率	50	52	54	56	58	60	62	64	66	68	70	72	74	76	78	80	84.1	90	95	98	99	99.9
至50%线的距离	0	0.05	0.1	0.15	0.2	0.25	0.3	0.36	0.41	0.47	0.53	0.58	0.61	0.71	0.77	0.84	1.00	1.28	1.64	2.05	2.33	3.10
累积频率	50	48	46	44	42	40	38	36	34	32	30	28	26	24	22	20	15.9	10	5	2	1	0.1

由于在概率格纸上正态分布表现为一条直线，所以大大简化了统计作业，只要根据实际数据作出这条直线，则任何一种百分率的分位数都可由图中直接读出。例如直线与50%线交点的横座标即为50%分位数。直线的斜率则反映所求正态分布的方差。直线与84.1%线交点的横座标与50%分位数之差，即为方差。

如果统计工作是为了求出某个元素的异常下限，则需根据具体地质情况，经验性地在85%分位数及99%分位数之间选择。

根据一批化探原始数据，在概率格纸上作出的分布曲线，如果在全部含量范围内成一条直线，即表明这一个子样来自单一母体。但在实际工作中很少遇到这种简单情况。作出的分布曲线往往是折线或是弯曲的。这表明子样来自复合母体，即包含着两个以上具不同参数的分布。如何分解它们，如何在化探资料整理时应用它们，将在下期介绍。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net