

288-293

地球化学正态分布悖论

林存山

(地质矿产部物化探研究所)

p632

摘 要

1. 化探中正态数据未必是正态地球化学特征的体现, 正态可能是一种假象。
2. 用分布参数的偏度与峭度作为正态分布的检验准则而不伴随其它检验方法是不充分的。
3. 用对数变换将正偏数据正态化可能导致参数计算的明显不合理。
4. 总体分解方法可因人而异, 结果是多样的使人迷惑的。
5. 将自变量和因变量正态化可实现最优回归预测的论点未必正确。这可通过分析或模拟得到确认。

化探人员从接触地球化学数据起便涉及正态分布的概念及相应的数据处理问题。虽然从统计学书中读到的正态理论和实例是完善和典型的, 但在实践中正态理论同地球化学数据的结合并不都很理想, 处理结果也不尽人意。一些化探人员常常致力于将自己的数据正态化, 而较少地关注到所得效果是否得到明显的改善。实际上, 对几十种化探数据的正态化不是总能成功的。他们可能面对同一个研究目标上不同分布类型的数据的存在而不知是该将它们统一起来, 还是任其共存。对多种处理方法引向不同的解缺乏判定准则。在解释推断上体现出薄弱与不足。在这种情况下作者在本文中阐明的几点反论似乎又增添了纷乱的头绪。但这是问题复杂性的客观存在。在分析问题、解决问题时这些是本来都应当考虑到的。

一、正态性假象问题

摆在化探人员面前的一个基本的问题是地质地球化学特征可能表现出哪些类型的统计性质(例如, 指数分布或对数分布等)? 它们能否从抽样的数据中客观地体现出来? 众所周知, 阿仑斯用实验数据最早地说明了在火成岩中微量元素呈对数正态分布。在五、六十年代对此曾经有过一些争议。表面上争议的是在追问是算术正态、或对数正态, 还是别的甚么。而本质上则是要回答正态的机制是如何导出的。阿仑斯的论点后来在不同程度上被应用。自那以后, 这方面的研究未见有大的进展。地质学中的正态分布律并没有得到实质上的阐明。只是近年来将正态分布用于化探方面仍明显地拓宽了。它的范围已延伸到土壤和水系沉积物等地球化学数据的处理方面。所研究的指标也已不限于微量元素了。在背景值与异常下限的确定, 异常的分类与分析, 以及编图等工作中, 正态分布常常都被联

系到。

与阐明地质学中正态理论机制相平行,对化探人员同等感到重要的问题是,一个外观上具有正态分布形态的数据是不是一定是正态地球化学过程的体现。或者说,自然界是否存在一种非正态律支配的地质地球化学过程,它的分布特征被掩蔽了,而代表它的数据则体现出正态的假象?

笔者认为这种可能性是不能完全排除的。读者只要注意到以下几点事实便可得到理解:1. 地球化学数据是通过采样和分析测试而取得的,而采样和分析都包含有误差。2. 虽然对地球化学采样误差研究得不够充分,但在地球化学分析中,对所产生的误差则有较深入的了解与研究,而且分析偶然误差已被确认为正态分布的。3. 因此从一个地质总体中获得的地球化学数据总是重迭有采样与分析的误差分量。这样可使从该总体中获得的原始分布受到畸变。当误差分量较大时有可能掩盖原始分布的形态,使化探人员产生错觉。

为了对上述的误差分量的迭加效果作观察,笔者在计算机上进行了一种模拟。这种模拟是在原始分布为矩形的总体上进行分量的迭加处理。我们略去了取样误差但保留下分析误差。图1为计算机模拟结果的一个图示。图中的矩形表示被模拟总体中某元素的原始分布(即概率密度函数)。图上横轴代表元素含量,其含量范围为10至100 $\mu\text{g/g}$ 。用对数值即表示为1.0至2.0。设以对数值计算的分析偶然误差服从平均值为0.0、标准差为0.3的正态分布,则迭加上此分析误差以后该总体的分布已经面目全非,而趋于较好的正态分布了。当含量范围加大,分析误差减小时,迭加分布的中央可形成一定宽度的平顶,而含量范围减小,分析误差加大时,此种平顶形态则减小或消失。笔者还试验了矩形以外的其它一些形态的原始分布(如三角分布,梯形分布等),多数情况均得到近似正态的迭加效果。

上述的分析和计算模拟表明,当实际地球化学数据表现为正态分布时,它未必代表真实的原始分布,而可能是一种假象。作为本论题的佐证,可以用一定范围的天然水体中简单金属离子的浓度分布为例。本例中金属离子的真实分布是一条直线(含量区间宽度为零的矩形)。实际得到的偏离直线的任何正态或钟形的分布形态都是其它因素,尤其是分析误差因素迭加上去的。

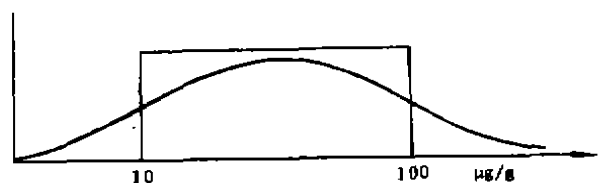


图1 伪正态分布的模拟

由于分析偶然误差的迭加使矩形的
原始分布被畸变成近似的正态分布

二、不充分的正成析验法

上述论点也许会对思维产生一点扰乱,但这对化探人员是一种提醒,对解释推断并无坏处。当前正态分布的概念在多数化探人员中是确立的。在一些情况下把数据这样对待是有利于、也方便于数据处理的。然而由于实际数据所体现的非典型正态的外观,使人们感到有客观判定的必要性。於是便有对数据进行正态假设检验的处理。通常可有几种检验正态性的方法,例如 χ^2 法。这是对比实际分布直方图与所欲拟合的理论正态分布直方图的频数差异而计算出检验用的统计量。而柯尔莫哥洛夫法是用实际与理论累积分布直

方图上的最大累积频数差来计算检验用的统计量。这两种方法都是比较有效的。

使人感到奇怪的是近年来有些化探人员在检验数据正态性时却很少用到 χ^2 法或柯氏法。他们只是去计算数据分布的偏度与峭度(偏度定义为一个分布的三阶中心矩与其标准差三次方之比,而峭度定义为一个分布的四阶中心矩与其标准差四次方之比再减 3)。众所周知,对于一个分布而言,它的偏度与峭度均为零,这只是正态的必要条件。使用偏度与峭度是否接近于零作为正态性的判断准则时如不伴随其它的检验,则是不充分的。可以指出,偏度与峭度同时为零未必都能反推出数据的正态性。对此我们只需举出一个反证即可。

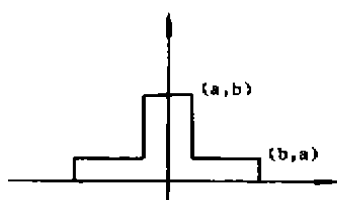


图 2 偏度、峭度均为零的“凸形”分布

图中 $a=0.245, b=1.143$

如果只按偏度=0.0,峭度=0.0 作为正态性的检验准则而不伴随其它的检验,则本图形将被错判为正态分布。

(当 $a=0.35, b=0.89$ 时,峭度=-0.55;当 $a=0.15, b=1.74$ 时,峭度=0.40)

笔者曾经计算一个“凸”形分布总体的偏度与峭度。计算结果示于图 2 中。我们看到随着这个凸形分布形态的变化,其峭度可由小于零变到等于和大于零(对于左右对称的分布,偏度总是零,因此我们不必去计算偏度)。这就是说我们已经找到了一种分布,它同时具备有偏度为零和峭度为零,但它不是正态分布。

三、可能导致不合理结果的正态化

对于直方图或分布曲线明显偏离正态的数据,采用一些数学变换有可能改善分布形态。例如对分布右端拖有“长尾”的正偏化探数据,将它们转换成对数可使这种长尾形态变成对称钟形。近些年来,国外有人还提出一种包括对数变换在内的所谓广义幂函数变换等。在我国使用对数变换的做法则较为普遍。这样一些变换大都不

涉及数据的内涵,它仅仅是一种数学运算。一些化探人员在将自己的数据施行变换并参加随后的计算处理时,未必都深刻地建立以下的概念,即一种变换越是具有对数据作大幅度变动(压缩或放大)能力的,则在作反变换时由于同等程度的变动能力有可能还原出不合情理的结果来。广义幂函数(包括对数)变换很可能就属于这样的一种。举例来说,对于一组“长尾”数据,将它们转变为对数后,由于分布形态的改善及通过了有关正态检验,它已被确认为接近于正态分布而被允许参加计算。於是求得这一组数据在对数变换下的一些参数,例如平均值为 $\bar{x}=2.30$,标准差 $s=0.48$,异常下限 $xa=2.3+2 \times 0.48=3.26$ 。对这些结果取反对数便得到相应的真数,即平均值 $=200\mu\text{g/g}$,标准差 $=3\mu\text{g/g}$ 及异常下限 $=1800\mu\text{g/g}$ 。显然化探人员很难接受高于背景值如此之大的异常下限。於是他们返回去检查数据处理过程,并没有发现什么错误。这样,他们就只好将所采用的置信限系数 2 减小,使异常下限降低到他们的经验所能接受的程度为止。不少化探人员都有这样的经验:使用对数变换后的数据去确定异常下限时,多数情况结果偏高。幂/对数变换可能引起的不良后果对于上述这种简单的参数计算况且如此(幸亏直观上它是可察觉、因而可弥补的),更不用说将它作为预处理并继以复杂的数学分析时,所得结果将是如何难以预料的了。

实际上,在有些情况下不如回到比较安全且简单的处理方法去,所得效果未必就很差。例如在上例中,如果以中位数为基础并弃除含量分布高端的小部分(如百分之几的)数据点,所算出的异常下限必定更为切合实际。

四、多总体分解处理中的多解性

对于上述的正偏的长尾分布数据,另一种处理方法是所谓的总体分解法。这种方法假设一组数据所代表的总体中包含有两个或多个正态分布的子总体,其中平均值较低、容量(样品数目比例)较大的正态子总体代表了背景总体,而平均值较高、容量较小的正态子总体则代表异常总体。它们在空间上的迭加可构成形态不一的各种偏离正态的混合分布。应用直方图、概率纸或计算方法均可进行这样的总体分解。对于原始混合分布呈现明显双峰等形态的数据,总体分解法可能比较适用。从文献中已见到这样的总体分解的方法与实例。

总体分解与正态变换两种方法显然很不相同。在正态变换中处理的前提是单总体偏态分布。它可以将长尾的偏态分布调整成接近于正态,但对于双峰(或多峰)的长尾分布却无法通过变换将它们简化成单峰的。另一方面,即使确实存在多总体迭加的混合分布,也不是每一个子总体的峰值都能在混合分布中体现出来。特别是当其中的某一个子总体占很小的比例时,它的峰值在混合分布中就可能被掩盖。这就是说,外观上看起来属于单总体长尾形态的分布可能是两个或多个正态子总体的迭加。因此,面对着一组偏离正态的数据,化探人员就无所适从了。他们可以有多种理解和处理方法。而对于同一方法而言,又可由于做法上的差别,产生出不同的结果。笔者曾经使用了两个正态子总体和三个正态子总体分别去拟合同一个正偏长尾分布,都能获得满意的结果。那么究竟哪一种拟合更为正确呢?再有,对于这样一组正偏(如果是单峰)的数据施行对数变换,所得结果如也能通过正态假设检验的话,则我们又有了单总体的结果。这三种结果差别很大,在方法上缺乏取舍准则,使化探人员不得不更多地走向经验、主观、和定性的方法。

五、正态化一定能改善回归预测吗?

与正态分布相联系的还有一些统计预测方法。在线性回归分析中,国内有人曾经做过一些试验并断言,如将自变量和因变量各自进行正态化变换,再对变换后的数据进行线性回归分析,则所得效果将比不施行正态变换的为优。这种断言是在进行了很有限的试验之后作出的,而且理论支持不足。这样的命题随后还被进一步作了引伸。同前类似,笔者只需举出一两个反证例子便可说明这个断言的不妥。在图 3 中,自变量 X 与因变量 Y 均已经是正态的了,但不能进行回归预测。在这里 X 与 Y 的相关系数近于零。这是回归分析中联系最弱、相互预测能力最差的一种变量对。读者可见到在图 3 中画有两条回归线,一条是 Y 对 X 的,另一条是 X 对 Y 的。假设我们现在研究 Y 对 X 的回归,即 Y 是因变量而 X 是自变量。这对应于图中斜率较小的那一条直线。

我们以下来说明,如将图 3 中的数据施行一种“去正态化”的调整时所得结果将反而能改善预测能力。图 4 示这样的调整过程。我们根据以下两条原则对数据点进行移动:1. 各点的 X 值保持不变,2. 各点 Y 值的秩保持不变。具体的处理是,在保持各点 Y 值大小的排列顺序不变的条件下,将各点平行于 Y 轴方向朝回归直线靠近。在图 4 中只画出图 3 中某一段 Y 值相邻的三个点 1、2 和 3(为醒目起见,本图作了放大)。将点 1、2 和 3 向下移到点 1'、2' 和 3'。我们看到在移动以后这三个点的排列顺序没有变化,但它们沿 Y 轴方向

与回归线相交的距离比移动以前缩短了。这相当于回归的剩差减小了。如果按某种设计步骤将图 3 中的全部点进行这样的调整,则在调整后各点的剩差均会有不同幅度的下降。此时如再作回归分析,所得的回归线虽然不会与原来这条相重合,但剩余平方和将取极小值并优于用调整后的点按原回归线(图 4)计算出的剩余平方和。当然也就更优于原来(图 3)的剩余平方和了。换句话说,回归效果得到了改善。类似地,如保持图中各点 X 值大小的排列顺序,将各点沿着水平方向移动并靠近回归直线。所得效果同样也是使回归预测得到改善。

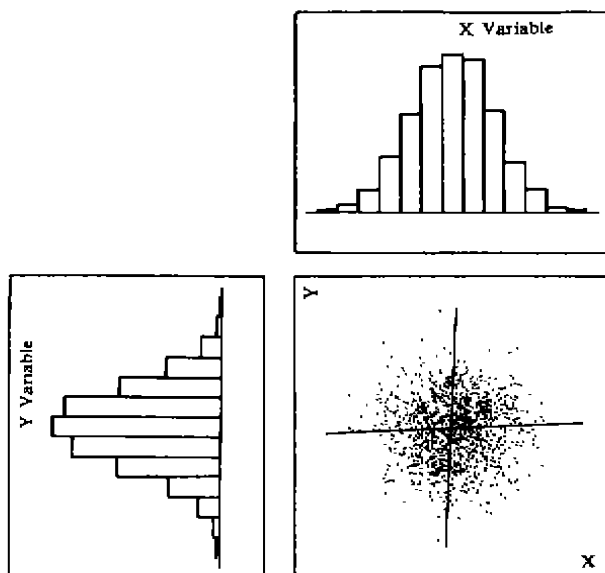


图 3 不能相互进行回归预测的一对正态变量

图示用计算机模拟出的由 3000 个点构成的两组正态分布的变量 Y 和 X。它们的相关系数趋近于零。图中两根直线分别代表 Y 对 X 和 X 对 Y 的回归直线。

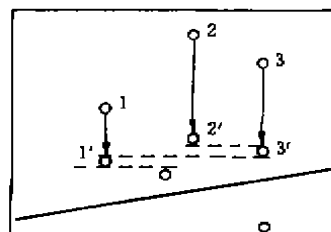


图 4 “去正态化”导致回归预测效果的改善的图示

这是作为回归预测一个命题的反证例子。本图是图 3 一部分的放大。图中显示将 Y 值相邻的三个数据点 1、2、3 在保持 Y 值的秩不变条件下向回归直线移动,分别到达 1'、2'、3'。表明 Y 的正态性受到改变或破坏。但此时回归剩差下降,表明回归效果得到了改善。

综上所述,我们对数据点所作的调整使得原有数据分布的正态性受到改变或破坏,同时却改善了两

组数据之间的线性拟合关系。由此我们已经不很严格地给出了反证。

结 语

地质学中化学元素统计分布的特征需要研究工作中理论与实践的结合来认识,尤其是要通过取样、分析和数据的处理与解释来实现。这种认识的过程是复杂的,需要研究人员多方面的思考和谨慎小心的工作。为找矿的目的而研究数据分布,其目标偏重于确定地球化学异常;而在基础地质研究中,目标则偏重于确定地球化学背景。实际情况中背景和异常的确定很有一些伸缩性,而严格处理的结果又常常由于经验和直观的修正而大大地变了样。即使撤掉这种主观方法的介入,当前的化探人员似乎不去精确追究数据是否具有正态性而照样也能工作得很好。这是因为例如使用一个梯形分布甚至三角形分布的模型去拟合化探数据,计算出的背景和异常下限也可以同用正态分布模型计算出的结果相差不多(如对比用主观修改所产生的大变动的)。除此以外,正态性所带来的信息并没有产

生研究工作中的飞跃或提高。假如化探人员同意这种看法,则冒着将要收到难以预料的结果去追求各种复杂变换处理的作法就不很值得的了。其实这样的做法 在不同程度上已被重复了多年和无数次。累计工作量之大是空前的,然而我们至今还没有得到能同所付出的这么大劳动相匹配的收获和学术新知。从应用上向理论靠近一步,地球化学正态性的探索看来要留待理论地球化学人员和勘查地球化学人员去继续努力。

参 考 文 献 (略)

PARADOX OF GEOCHEMICAL NORMAL DISTRIBUTION

Lin Cunshan

(Institute of Geophysical and Geochemical Exploration)

Abstract

Geochemical data of normal distribution may not indicate the normality of geochemical characteristics. It may be a false appearance. It is insufficient to use skewness and kurtosis as criteria for normality test without additional tests procedures. The use of log-transform to normalize skew data may cause very unreasonable results in parameter calculation. Decomposition of multi-mode populations relies on persons. The results are varied and confusing. The argument that to normalize both the dependent and independent variables will implement an optimum regression in prediction may not be true. This can be demonstrated by analysis or simulation.

【作者简况】林存山,地矿部物化探研究所副所长,高级工程师(教授级),中国地质学会勘查地球化学专委会主任委员,国际地科联 COGEODATA 专业组成员。

~~~~~

# 庆祝本刊 创刊十五周年!