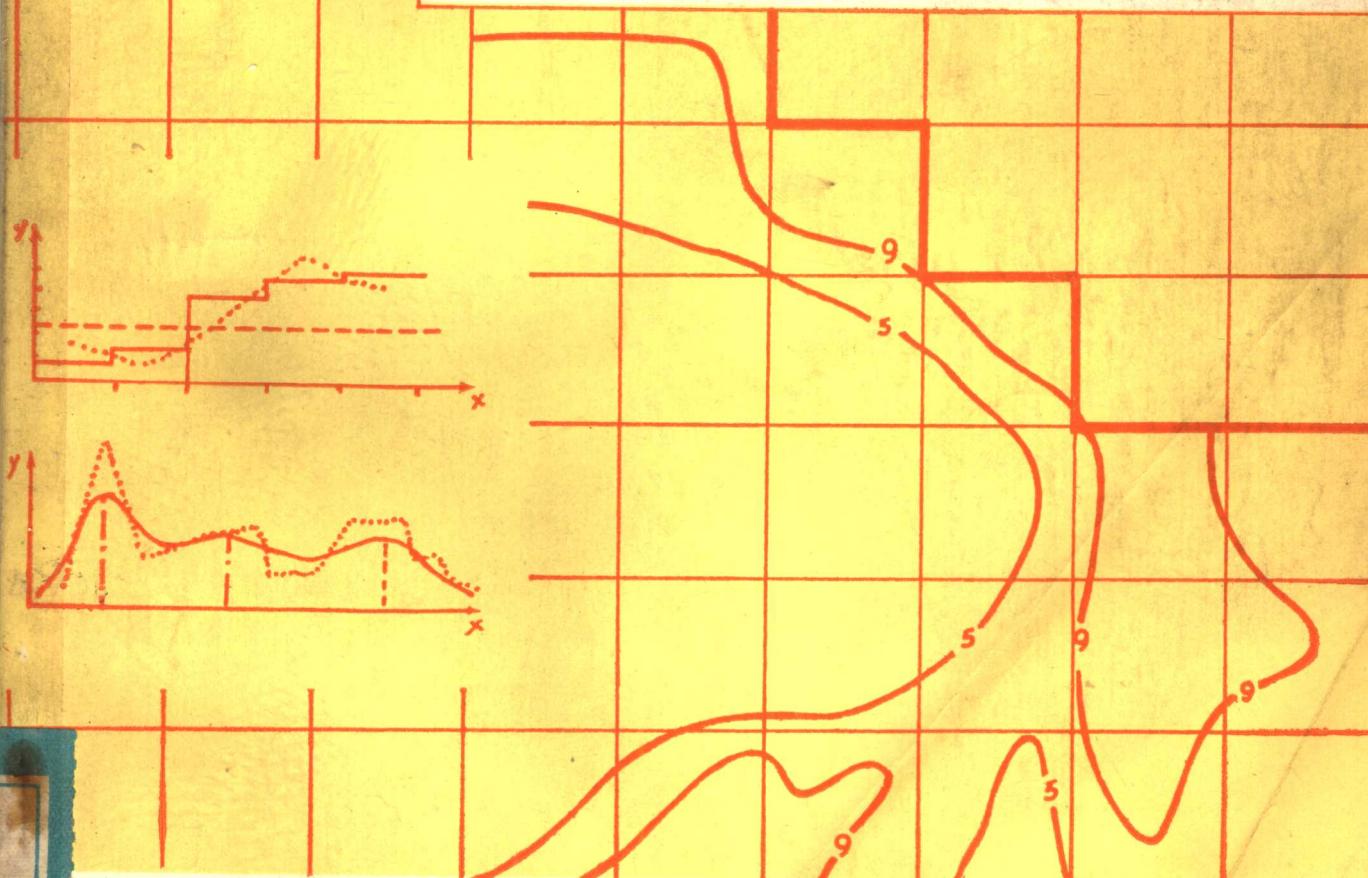


高等学校教学参考书

矿床统计预测

赵鹏大 胡旺亮 李紫金 编著



地质出版社

高等学校教材

矿床统计预测

张德成 张德成 张德成 张德成

地质出版社

高等学校教学参考书

矿床统计预测

赵鹏大 胡旺亮 李紫金 编著

地质出版社

矿床统计预测

赵鹏大 胡旺亮 李紫金 编著

责任编辑 刘绫堂

地质矿产部教材室编辑

地质出版社出版

(北京西四)

地质出版社印刷厂印刷

(北京海淀区学院路29号)

新华书店北京发行所发行·各地新华书店经售

开本: $787 \times 1092 \frac{1}{16}$ 印张: $17\frac{3}{8}$ 字数: 408,000

1983年3月北京第一版·1983年3月北京第一次印刷

印数: 1-4,847册·定价: 2.40元

统一书号: 15038·教146

前 言

近年来,国际上的一些未来学家预言世界即将面临的几个根本问题是:人口、粮食、能源、资源及环境。第26届国际地质大会主席在一次会议上也强调:“本世纪末的中心问题就是能源和矿产资源问题”。生产和消费的迅速发展,人们对矿物原料的需求量也日益增长。有人估计,到2000年,我国人口将达到12亿,为了实现人均收入1000美元的目标,每人消耗能源以现代化的最低标准1.6吨计,则总共将消耗能源为19.2亿吨。如果其中煤炭占70%,石油占10%,则两种矿产应分别达到年产12亿吨和1.9亿吨。如按一般正常的采储比(设煤为1:60,石油为1:35)加以折算,则煤和石油应分别有780亿吨及70亿吨储量为保证。显然,发现和探明如上数字的储量乃是一项十分艰巨的任务。此外,对于不可再生的矿产资源长期供给保证问题一向是人们所普遍关注的大问题。

另一方面,由于易于发现的露头矿、浅部矿、大型矿和已知类型矿的日趋减少、找矿勘探的难度日益增大。如加拿大,1950年以前,85%矿床由传统方法找出,1951—1955年期间下降至46%,1956—1960年为25%,1961—1965为31%,而1966—1969仅有9%。与此同时,找矿勘探的成本大为提高。1950年加拿大的矿床勘探费用占金属矿床产值的0.8%,1955年增至2.4%,1960年为3.2%,而1965年则增加到4%。不仅如此,由于当今对矿产预测、普查找矿及勘探等工作投入工作量的增加,因而由这些工作的失误所造成的损失也就明显增大了。

为了满足对矿物原料日益增长的需要,必须提高普查勘探工作的效果。以深入研究地质成矿规律为基础的矿产预测工作日趋重要。近30年来,成矿规律学迅速发展并形成了几个独立分支:①全球成矿规律学、②区域成矿规律学、③矿区成矿规律学及④专门成矿规律学或各别元素成矿规律学。与此同时,人们积极探索新的矿床预测理论和方法,其中数学地质及电子计算机在矿床预测中的应用是近十年来最重要的成果之一。目前,数学地质在矿床预测中的应用或矿床统计预测也相应地形成了几个重要方向,即:①矿产资源总量估计及潜力评价、②成矿远景区定量预测及③成矿、找矿地质标志的统计分析及其含矿性评价。

目前,矿床统计预测是数学地质最活跃的一个分支。与26届世界地质大会同时召开的国际数学地质学术讨论会上有关矿床统计预测的论文占论文总数1/3,我国长沙第二届全国数学地质学术讨论会上有关矿床统计预测的论文占1/4,无论国内外,矿床预测都成为数学地质首屈一指的应用领域。

本教材正是根据上述形势的需要而编写的。作为矿床地质专业和矿产地质及勘探专业的学生,学习数学地质的最基本目的不仅在于掌握数学地质的基本理论和方法,而且还应了解如何将这些理论与方法应用到地质勘探工作的最重要领域——矿床预测中去。

本教材是在1978年编写的《地质勘探中的统计分析》基础上经过三年教学试用加以补充修改而完成的。参加1978年教材编写工作的有赵鹏大、胡旺亮、李紫金、冉宗培及邹海卿等同志。

本教材在编写时除根据作者近年来在矿床统计预测研究中的成果外,还广泛应用和参阅了国内外有关单位和学者的资料。主要有中国科学院,中国地质科学院数学地质组,

长春及成都地质学院数学地质研究室及教研室，云南大学数学系及联合国教科文组织北京矿产资源评价讨论班专家报告等资料。在此对以上单位和有关学者表示感谢。

武汉地质学院数学教研室刘绂堂同志曾详细阅读1978年教材第一稿并提出宝贵意见，地质部资料局袁君孚高级工程师及刘绂堂同志又对本教材原稿进行了详细审阅并提出宝贵意见，在此对他们表示衷心感谢。

在本教材编写及定稿过程中，武汉地院教材科、绘图室及矿产系等单位给予了大力支持和帮助，特别是褚松和、王润斋、邢相勤、张桂珍、黄宏瑞等同志及77级部分学生作了大量辅助工作，在此一并表示感谢。

对本教材中的缺点和错误，敬希读者批评指正。

目 录

第一章 绪论	(1)
第一节 数学地质的任务、对象和发展.....	(1)
第二节 数学地质发展的必然性.....	(3)
第三节 概率统计中的某些基本概念.....	(7)
第四节 地质学中应用统计分析的特点.....	(10)
第二章 地质勘探数据的统计分布特征及其意义	(16)
第一节 研究统计分布特征的意义.....	(16)
第二节 几种地质上重要的分布模型.....	(17)
第三节 混合分布及其地质意义.....	(34)
第四节 分布模型在地质研究中的实际应用.....	(44)
第三章 地质变量的选择、取值及变换	(47)
第一节 地质变量的类型.....	(47)
第二节 地质变量的选择.....	(48)
第三节 地质变量的取值和变换.....	(59)
第四章 矿床统计预测：矿产资源总量估计及潜力评价	(65)
第一节 矿床统计预测的数学分析途径.....	(65)
第二节 矿产资源潜力评价概论.....	(67)
第三节 资源总量估计和潜力评价方法.....	(70)
一、“镶嵌模型”估计矿床产出概率法.....	(70)
二、矿床值概率分布法.....	(71)
三、单位区域价值估计法.....	(76)
四、丰度估计法.....	(79)
五、体积估计法.....	(83)
六、矿床模型法.....	(84)
七、德尔菲法或主观概率法.....	(89)
八、综合法.....	(92)
第五章 矿床统计预测：成矿远景区定量预测	(94)
第一节 预测比例尺、基本单元划分及控制区的确定.....	(95)
第二节 秩相关分析法预测.....	(99)
第三节 信息量计算法预测.....	(104)
第四节 多元回归及逐步回归分析法预测.....	(107)
第五节 判别分析法预测.....	(125)
第六节 聚类分析法预测.....	(142)
第七节 逻辑信息法预测.....	(161)
第六章 矿床统计预测：地质标志的统计分析 & 含矿性评价	(177)

第一节	地质趋势研究.....	(178)
第二节	最优分割与地层划分.....	(200)
第三节	阶梯函数分析与变质岩分层及构造模式的建立问题.....	(209)
第四节	因子分析与地质成因解释.....	(217)
第五节	对应分析.....	(234)
第六节	非线性映射分析.....	(242)
第七节	典型相关分析.....	(253)
第八节	马尔科夫过程与地质过程的随机模型.....	(261)
尾 语		(270)
主要参考书		(272)

第一章 绪 论

第一节 数学地质的任务、对象和发展

作为一门新兴的边缘学科——数学地质近年来在国内外有了很大发展。地质学从传统定性描述和成因推理进入到应用数学作为工具，并且运用现代电子计算技术自动处理地质数据，显示地质成果并解决各类复杂的理论和实际问题，这不能不说是地质学发展中的一次质的飞跃。尽管如此。数学地质作为一门独立的学科，仍处于形成和初创阶段，许多数学地质方法仍具有探索性质，它们解决实际地质问题的有效性还有待于更多的生产实践检验。

关于数学地质的任务、对象和内容至今还没有一致的看法。概括地说，有两类观点。一些学者认为数学地质包括“地球科学中的全部数学应用”(F.P.阿格特伯格1974)，类似观点如“地质数据的定量分析方法”(J.C.戴维斯1973)；“用数学方法研究和解决地质问题”(中国科学院地质所数学地质组《数学地质引论》1977)等，诸如此类可认为是“广义的观点”。另一些学者，如A.B.维斯捷利乌斯(1977)则强调必须严格区分数学地质和地质学中的数学。在谈到这两者的关系时他认为：地质学中需要用数学方法既解决各类带根本性的任务，也解决各种技术性问题，但不是所有的数学应用都属于数学地质范畴。维斯捷利乌

2. 研究地质过程中的各种因素及其相互关系，建立地质过程的数学模型；
3. 研究适合地质任务和地质数据特点的数学分析方法，建立地质工作方法的数学模型。

地质科学中最早应用统计分析方法一般追溯到1833年。当时英国学者莱伊尔首次用统计分析方法对第三纪地层进行划分。尽管地学中应用统计分析已有100多年的历史，但在这当中几经起落和兴衰。

早期阶段，大致在本世纪20—40年代初期，数学方法主要应用于一些实用地质学的领域（如矿床勘探和评价）、岩石学（如沉积岩粒度分析、岩浆岩分类等）以及地层古生物学方面（如化石种属变异特征的研究等）。这个阶段的特点是应用单变量统计方法，利用个别的统计标志解决某些具体的地质问题。另外，这一阶段也是对统计分析在地学中应用的可能性问题进行争论最激烈的时期。

40年代后期，数学方法在地质学中有了更广泛的应用。特别值得提出的是多元统计分析开始得到应用。如1949年古生物学家Burma在《多元分析——地质学和古生物学中的一种新型分析工具》一文中指出了多元统计方法是一种最有远景的计量生物学方法。1956年克伦宾在研究岩石的矿物、岩性和化学成分时应用了多元分析方法，把岩石成分作为N维空间中的一个点或向量来进行统计处理。在这一阶段，法国和非洲南部的一些地质和统计学家创立了“地质统计学”的理论和方法，它和传统的统计学不同，地质数值被看成是空间变化具有貌似连续性的变量的值，这种变量称为“区域化变量”。通过研究区域化变量特征建立变差图（或变异函数）。这一方法主要应用于矿床储量、品位估计及矿石评价方面。近年来，美国、加拿大和澳大利亚等国加强了这方面的研究和应用。

从60年代开始，数学地质进入了其发展的极盛时期。仅1968—1970年有关数学地质的论文超过3000篇。1968年成立了国际数学地质协会并从1969年起出版了专门的数学地质期刊。近年来，更出版了不少数学地质方面的教科书和专著，反映了几乎地质科学一切领域中应用数学地质方法的成果和经验。

这一阶段开始的高速电子计算机的应用，不仅扩大了数学地质的应用范围和引入更多的多元统计分析方法，而且出现了地质过程计算机模拟这一富有重大意义和生命力的方

表 1-1

应 用 领 域	论 文 篇 数
矿床勘探、矿产预测和资源分析	31
沉积、地层及古生物研究	26
地质图书情报检索	14
预测地质学（人类未来环境预测）	12
图形分析及模型识别	10
地球化学	10
地质数据库	9
构造地质学	6
自动绘制地质图	5
水文地质学	3
遥感数据处理	2
物探	2

向。此外,地质数据的计算机存储、检索、显示等数据管理和资料文献计算机管理方面的进展也为地质科学的现代化作出了重要贡献。

近年来,与数学地质有关的国际学术会议频繁召开。如1971年在海德堡召开第一次“沉积分析定量技术”学术讨论会,1975年在法国举行了第二次同名的学术讨论会。在数学方法用于资源评价方面,1976年在挪威召开了第一次“资源研究中计算机应用标准化”讨论会,1977年在肯尼亚和1979年在墨西哥召开了第二、第三次上述问题讨论会。又比如,1977年在澳大利亚召开了第15届国际“计算机在矿业中的应用”讨论会等等。1980年在巴黎召开第26届国际地质大会的同时召开了国际数学地质学术讨论会,会上提交论文100多篇,其应用领域和论文篇数的情况如表1-1所示:

可以毫不夸张地说,当今地球科学的每一个领域都或多或少应用了数学地质的方法,当然,应用效果的显著性是各不相同的。

第二节 数学地质发展的必然性

任何科学的产生和发展都离不开人类最基本的生产实践活动。科学理论和方法都必须通过生产实践的反复检验才能显示其有效性和生命力。数学地质的产生和发展也有其客观的必然性。

1. 地质科学已经发展到定量地质学的新阶段

随着地质生产和研究工作的大量开展,各种新技术新方法的普遍应用,地质工作者已开始面临“数字的海洋”。正确而迅速地处理数据,最大限度地获取有用信息已成为地质学科所面临的新问题。另一方面,地质科学的任务随时间的推移而发生了很大的变化。正如美国数学地质学家麦克卡门(1980)所说“如果说在过去地质学家可以安静地研究和冥思地球的历史,那么现在则要求地质学家对一系列社会问题的解决作出贡献。例如:及时地提供有关影响人类活动事件的性质、地点、规模和时间的信息,评价人类活动的环境影响以及评价自然资源量的大小等等”。上述任务的解决都要求地质学具有更大的预测性。近十年来,地质学的预测能力有了很快的增长,其中一项有关的发展便是地质学的量化问题。

作为定量地质学,最基本的要求应体现在以下几方面:

① 准确地定义地质体或地质现象

由于地质学的描述性质,许多地质体或地质现象缺乏严格定义,有时甚至因概念含糊不清造成理解上因人而异,这给地质分析、对比和解释带来很大困难。例如:深大断裂、次火山岩、绿岩、层控矿床、斑岩矿床、勘探类型……等等都没有公认的明确定义或是定义不够严密而易造成理解上的差异,这种例子在地质学中是不胜枚举的。从地质学的量化角度来看,统一名词术语并给以准确严格的定义是十分必要的。当然,这项工作是非常艰巨的,但却是不可回避的。

② 在解决各类地质问题中,给出数量的准则

以矿产预测为例,所谓“定量预测”,不仅要求确定成矿远景区的空间位置,而且应给出远景区可能发现矿床的个数或资源数量,给出在该远景区可能发现矿床的概率,并且查明控矿地质因素和找矿标志的数量规律性,如各因素或标志的信息量(信息权),各因素

标志的找矿概率,各因素标志最有利成矿或找矿的数值区间等。例如岩层产状、岩石孔隙度、岩性组合特征值等各有其不同的变化范围,但不是所有这些值与矿化有同样的关系,应区分:①该因素对成矿的最优值区;②有利数值界限;③中常界限及④不利界限。大部分控矿因素最优值并不相当于它们的最大值或最小值。

其它如在解决分类、对比、判别等性质的问题中也都应有数量的准则。

③ 通过建立数学模型,检验地质理论和假说

有人认为地质理论和假说是“无法检验的”。从地质科学未来的发展来说,这种看法未必正确。地质科学与数学的进一步结合将有可能使地质理论和假说建立在更可靠,而且是可以检验的基础上。数学模型的最大用处恰在于此。例如1972年苏联数学地质学家维斯捷利乌斯提出“理想花岗岩”概念,他应用 $Or-Ab-Q$ 三种矿物组分在实验系统中结晶的可靠基础,计算穿过矿物集合体任意直线上的颗粒转移概率,证明 Or 、 Ab 及 Q 三相必定具有一种简单的马尔科夫性质,然后用实际花岗岩样品所测颗粒转移的经验数据与之进行比较,从而判断或检验花岗岩的各种成岩假说。

④ 正确处理地质数据

地质数据不仅在数量上是大的,而且很多属于多元数据。另外,大多数地质数据具有噪音强、混合性强、区域性强等特点。正确地处理地质数据,并从中发掘最有用的信息,恰为定量地质学的最重要任务之一。

2. 地质学中概率法则的重要性

地质学研究的对象,无论是各种地质体或地质现象以及各种地质观测结果,都普遍地受概率法则支配或影响。这就是说,地质现象、地质过程及其产物大多可视为随机事件,而各种地质观测结果则具随机变量性质。维斯捷利乌斯(1977)曾指出:“地质对象是由一些单个单元联合起来的,这种联合是遵循概率法则的”。的确,地质科学研究的对象都是在广阔空间、漫长时间和复杂介质中发生和发展的地质作用和产物,它们既受确定性法则支配,又在很大程度上受偶然因素的影响,即受概率法则支配。因此,大多数地质观测值都可看成是既受系统性或规律性变化因素的影响,又受局部性或偶然性因素影响的综合叠加结果,用数学模型可表达为:

$$x_i = I_i + Y_i + \varepsilon_i$$

其中 x_i ——第 i 点观测值;

I_i ——受系统因素影响的规律部分或趋势分量;

Y_i ——受局部因素影响的异常部分或异常分量;

ε_i ——受随机因素影响的偶然误差。

由上可见,地质规律大多具有“统计规律”的性质。就是说,它们服从一定的分布律,各种地质事件及观测结果在不同条件下发生的客观可能性或概率不同,要研究掌握它们发生发展和变化的规律,就必须借助于研究随机现象的数学工具——概率论及数理统计。

下面我们看两个实例:

① 找矿工作与概率法则

矿床的形成和分布虽然由于受一定地质条件控制而具有某些规律性,但仍可将其视为受多种随机因素影响的随机事件。因此,在地质条件有利地区进行的找矿工作也必然受概率法则支配而具有一定的风险。根据国外一些统计资料,我们可以看到找矿失败或成功概率

的一般状况：

——据国际原子能机构（1973）统计

对美国10万个异常进行了检查，从中获得4000个远景地（矿点），其中700个可列入矿床，因此，异常：矿床 = 143:1。找矿成功概率 $\approx 0.7\%$ 。

——据Koulomzine及Dagenais（1959）及Roscoe（1971）

加拿大公司经营的4865个矿床中，148个（约3%）是盈利的。

加拿大勘探工作发现矿床的比率1951年为1%，1969年下降为0.1%。

——据Perry（1968）

在美国西南部的五个勘探队检查了352个矿点，对47个进行了物探，23个进行了钻探，只判断其中的两个有可能开拓，成功比为176:1，即约0.6%。

——据Bailly（1967）

美国主要勘探公司（Bear Creek公司）1963—1966年间对1649个可能靶区进行了认真的考虑，对其中60个进行了钻探，发现了15个新的矿化区，其中8个有某些潜在储量，而5个为显著“矿床”。成功比为330:1，即约0.3%。

——据Griffis（1971）

加拿大Cominco公司在40年间勘探了1000个以上矿点，对78个进行了详细勘探，最后18个投入生产，但仅仅7个是可盈利的，成功比为0.7%。

以上是对矿点或异常进行勘探的成功比。若不是从矿点或异常出发，而是从一个区域出发进行找矿，成功的概率又如何？

以印度为例：从1976年开始的在80000平方公里范围内的一项勘察计划，经过10个月的航空电法、磁法和放射性测量，选择出1100个异常进行地面地质普查，至1972年夏对700个异常进行了地面地质、物、化探调查，发现其中25个值得进一步工作，1972年末对12个对象进行了钻探，其中6个被认为是希望的矿点，如果其中1个真正成为有价值的矿山，则原始风险为1100:1。一般说来，面积性地质调查的成功比如下：区域评价(1000:1)，远景区普查(500:1)，详查(100:1)。

显然，找矿成功比还是时间的函数。

② 区域评价与概率法则

一个地区是否值得进一步进行普查勘探工作，一方面取决于地质评价的结果，而另一方面又必须考虑可能的经济效果。可以说，区域评价是一项在不确定条件下进行统计决策的工作。这里的不确定性主要受找矿概率的制约。

设在一个地区进行找矿，其可能找到矿床的概率和可能找到不同价值矿床的概率如图1-1所示。影响矿床价值的主要因素这里主要考虑了规模大小、品位和埋深。这些因素的变化，造成现值差异幅度很大。

由图1-1可见，影响决策的关键因素是①成本；②优劣比（找矿概率）及③报酬（期望价值）。还可以看到：理想的找矿对象，如规模大、品位高并埋藏浅的矿床尽管具有较高的价值，但一般地说这类矿床数量极少，发现这种对象的概率很低，从而最终的期望价值也是不高的。因此，对于很少机会发现的高价值矿床的探矿工作应少投资。如果找矿优劣比为1:100，则对于10万元的勘探费用，寻找的目的应该至少具有现值1000万元。当优劣比更差时，则预期目标的现值应更高，或是找矿成本必须更低才可进行。这一实例，充分

决策枝状图

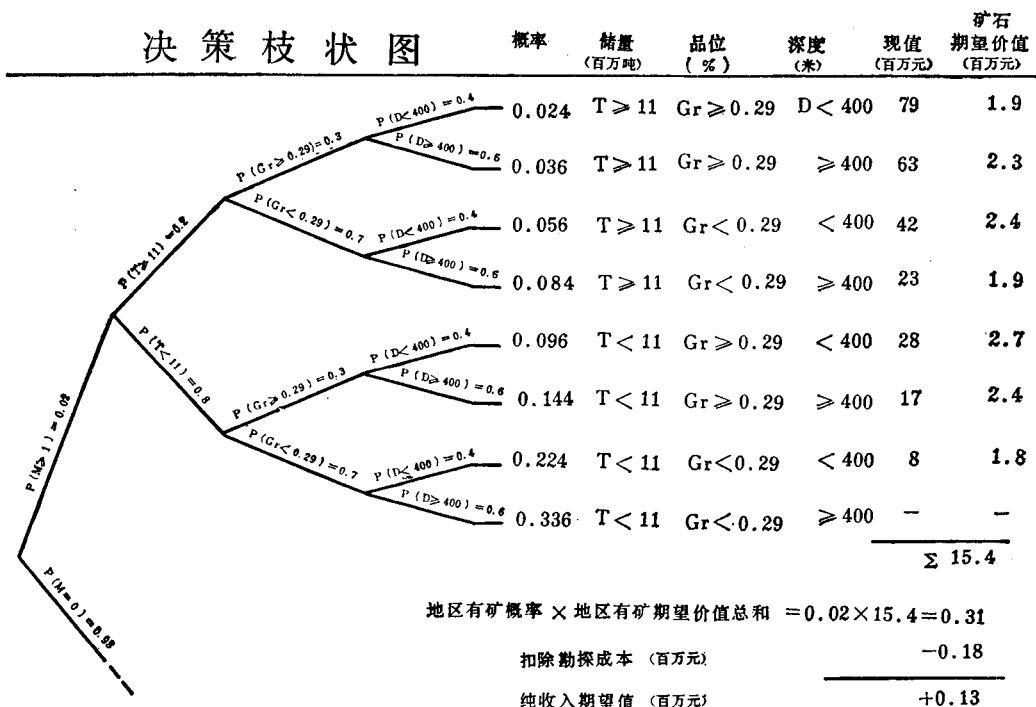


图 1-1 找矿决策枝状图及找矿纯收入期望值的计算

说明对一个地区进行评价，决策普查勘探工作部署时概率法则的明显支配作用。

3. 地质研究中的抽样观察是最基本的工作方法

从获取原始资料的最基本方法来说，地质观测是在一定间距的点或线上进行的，通过点与线上资料的内插和外推，或通过确定某种平均特征来表示地质体的空间展布或代表地质体的某种属性。换句话说，地质工作的基本方法是通过抽样观察来推断总体，即使是现在利用航空或卫星遥感技术，开始对地壳进行面的观测，但对于深部的情况来说，它仍然是一个局部。显然，任何地质观测结果都不可避免地存在有“代表性误差”。只要保证抽样的随机性，那么这种误差也带有随机性质，从而具有一定的概率分布。因此，各种观测结果也是带有不确定性的随机变量。我们在大多数情况下只能以一定的概率作出具有一定精确程度的结论。这就是数理统计中所常用的“区间估计”。区间越窄，估计的精度越高，但落入此区间的概率就越小，或者说结论的可靠程度越差。我们知道，区间大小取决于

$t \frac{S}{\sqrt{n}}$ 。在观测值数量 n 不变，对可靠程度要求一定 (t 一定) 的情况下，标志的变化程度 S 越小，则所作估计的精度越高。另一方面，若变化性 S 及 t 为固定的，则提高精度的途径只取决于增加观测数量 n ，所以我们对各种观测结果或研究结果都要进行一种精度评价和可靠性评价。当置信区间不太宽，可靠程度不很低时可求出最佳解。

迄今为止，人们往往忽视对观测结果或各种估计作出精度和可靠程度的分析，这可能导致错误的结论和推断。

例如，在一交代蚀变花岗岩体内取样，样品数 $n = 12$ 。根据此样品计算 Th 及 Cu 二元素的平均含量 ($\bar{x}$)、标准差 (S)、变异系数 (V) 以及概率系数 $t = 2$ (可靠程度为 95%) 条

件下估计平均数的绝对误差(m)和相对误差(p)示于表1-2。

表 1-2

元 素	$\bar{x}$ (ppm)	S (ppm)	V (%)	$m_{0.05} = 2 \times \frac{S}{\sqrt{n}}$ (ppm)	$p = \frac{m}{\bar{x}} \cdot 100(\%)$
Th	26	4.05	16	2.3	8.8
Cu	52	72.6	140	42.0	80.7

由上表资料可见, 根据12个样品所计算的两元素平均含量的误差差别很大, 对于Th, 平均数精度较高, 其相对误差 $<10\%$, 而Cu元素平均数误差很大, 达80%以上, 这是由于Cu元素在该花岗岩体内较之Th有更大的变异性。为了提高确定平均数的精度, 则必须增加取样数量。因此, 如何能够获得正确的观测结果以及如何评价已经取得的观测结果是十分重要的问题。

由上可见, 数学地质的产生和发展乃是生产发展和地质科学进步的客观需要和必然结果。

第三节 概率统计中的某些基本概念

下面结合一些实例, 回顾若干最基本的名词和概念:

1. 总体与样本

假设我们从一个矿体上采下一组刻槽样品, 或从一个花岗岩体上采下一组岩石标本, 并测定这些样品或标本中的某个(或某些个)指标值, 得一系列测定值。这个矿体或岩体上这种指标可能取得的一切值的集合在数理统计中称为总体, 由样品或标本上所得到的测定值称为样本; 显然, 样本是总体的一个子集合。为了方便, 我们也常把整个矿体或岩体称为总体, 把所取的一组样品或标本称为样本。但是有时需要根据研究目的对所研究对象进行更具体的划分。例如一个矿床, 或者一个矿体, 可能由多种成因或多次成矿阶段产物叠加而成。如果把整个矿体作为一个勘探对象, 而且不同成矿阶段的矿石不可能或无必要进行分选开采, 则可把整个矿体某种组分的一切可能的品位值看成一个统计总体。但是, 如果目的是要研究矿体成因或评价不同矿化阶段的矿石特征, 那么就必须以成因为基础来确定统计总体。不能不加分析地把所有样品化验值归为一个总体。这种因工作目的而确定的总体称为“目的总体”。样本中的单个样本值称为个体。一个总体中所含有个体的数称总体大小或总体容量。它可以包含无限多个体, 如矿石品位值; 也可能只有有限个数的个体, 如某个矿床已施工的钻孔数。前者称“无限总体”, 后者称“有限总体”。总体是相对的, 与我们研究工作的目的、对象和给定的条件有关。

样本中所包含个体的数称为“样本大小”或“样本容量”, 如从矿体中采了5个矿样, 得到5个化验值, 则样本大小为5。因此要注意, 不能将统计学中的样本大小与地质上的“样品大小”相混。另外, 在地质学中, 从地质体上采集样品的过程或手续称为采样或泛称取样, 而统计学中从总体中抽取样本的过程则通称“抽样”。

2. 变量与随机变量

我们在这里不准备给出随机变量的正式定义,而只加以描述。例如在矿体不同部位取样,我们将获得一组不能在事前加以预测的高低不同的品位值。可见品位是一个其值不能预测的变量,它所描述的是,矿体内各点所含某元素的量由于多种偶然原因而不同,这样一种所谓随机现象。这种变量称为随机变量。从矿床勘探实践中可以得知,某些变化复杂的贵金属、稀有或有色金属矿床,即使两个紧密相邻样品的品位值也可以差别很大。因此很多地质变量都可视为随机变量。但另一方面,这些变量的值有时确实显示一定的相依性或地区性。前者表现为当变量在某一点的数值较高时,则其邻近点的数值也趋于偏高,反之亦然。后者则表现为变量数值的大小与空间位置有关,即变量在空间的每一点上有其一定的数值。因此,地质变量,其中包括品位、厚度等是否为随机变量是一个长期有争论的问题。有人将这类受空间位置制约的变量称为“区域化变量”(如马特隆、布赖斯等),并发展了一整套被称为“地质统计学”的理论与方法。随机变量可分为两类:只能取有限多个或可列无穷多个值的随机变量称为离散型随机变量,或简称离散型变量,如某地区单位面积内的矿点数,各钻孔中的特高品位样品数等。另一类随机变量则可以取一切实数值或某一区间中的一切值。如矿体埋藏深度可取一切值,又如岩心采取率则可取区间 $(0, 100)$ 中的一切值,这种变量称为连续型随机变量,或简称为连续型变量。

3. 频率与概率

某一事件 A 在 N 次试验中出现的次数 m 称为频数(f^*),而出现次数与试验次数的比 $\frac{m}{N}$ 为频率或相对频数(f)。如有一批外检化学分析样品共30个,其中为正误差者12个,则出现正误差的频率为 $12/30 = 0.40$ 。根据定义可知,事件 A 出现次数 m 不可能为负数,也不可能多于试验次数 N ,即 $0 \leq m \leq N$,所以某事件出现的频率应满足 $0 \leq f \leq 1$ 。如果在相同条件下多次重复试验,当试验次数 N 足够大时,可发现某随机事件出现的频率是趋于稳定的,它围绕某一固定数值作微小地波动,这一固定数值体现了随机事件的统计规律性。它反映某一事件在某种条件下出现的客观可能性。这个作为某事件出现客观可能性的度量的固定数值(常数),即称为该事件出现的概率,通常用 P 表示。事件 A 出现的概率记为 $P\{A\}$ 。不可能事件的概率为0,必然事件的概率为1,而随机事件的概率介于0与1之间。著名的“大数定律”的最简单形式,即是当试验次数无限增大时,事件{频率无限接近于其概率}的概率趋于1,即接近于必然事件,这又称贝努里定理,可用下式表达:

$$P\left\{\left|\frac{m}{N} - P\right| < \epsilon\right\} > 1 - \delta \quad (1-1)$$

式中 ϵ, δ 为任意小的正数。

因此,当观测或试验次数足够大时,可取事件的频率做为其概率的近似值。在矿床勘探实践中,当样品数量足够大时,可以用样品平均品位估计矿体真实平均品位,其理也正在于此。

4. 频率分布与概率分布

在 N 次试验中,可以有 N 个试验结果(或事件)发生,但这些事件发生的可能性大小不一定相等,这就表现为每种事件发生的频率不一定相同。如果我们用等间距的区间将试验结果或观测值分组,则各组观测值可能具有不同的频率。在横轴上划出各区间,依次以

它们为底并以各组的频率为高作矩形，它们称为频率分布的直方图。如果从左边开始用直线段依次连结上述直方图顶边中点，得一条折线，它称为频率分布的多角形图。这种以样本观测值为基础得出的频率分布又称为观测值的统计分布或经验分布。假设样本数量无限增大并将各分组区间的长无限减小，则可得出一条反映事件出现可能性大小的光滑曲线。实际上，我们当然不可能用无限增大样本数量的办法来得到这种光滑曲线。通过理论概括，可以得出反映总体分布的理论频率曲线，也即概率曲线 $y = f(x)$ 。函数 $f(x)$ 称为概率分布密度，给定 $f(x)$ ，可求出随机变量 X 在区间 (a, b) 取值的概率。

$$P \{ a \leq X < b \} = \int_a^b f(x) dx \quad (1-2)$$

全部事件的总概率为1：

$$P \{ -\infty < X < +\infty \} = \int_{-\infty}^{+\infty} f(x) dx = 1 \quad (1-3)$$

上述理论分布密度函数 $f(x)$ 称为随机变量的分布律，也是理论分布的数学模型。在实际工作中，往往通过抽样先得出样本的频率分布，然后选择与之相适应的理论分布函数。

一切地质数据，都离不开某种类型的分布律。查明各地质变量的分布律是统计分析研究的最基础性工作。

5. 参数与统计量

在各种类型的分布密度函数的表达式中，例如正态分布密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

其中 x 表示变量， μ ， σ 是可变的常数，称为参数，分别代表总体的平均数和标准差。它们对于一个特定的总体为固定值，但对不同的总体又是可变的。当 μ ， σ 变化时，正态分布曲线也随着发生变化。为了唯一地确定分布形态，必须求出 μ 及 σ 。上式中 $e = 2.71828 \dots$ ， $\pi = 3.14159 \dots$ 为固定常数。参数为 μ 和 σ 的正态分布常记为 $N(\mu, \sigma)$ 。

作为总体特征值的参数，事先是未知的，它们只能分别由样本的相应特征值来代替。这种由样本导出的数显然随样本而定，因此是样本的函数，称之为统计量。换句话说，总体参数是由相应的统计量估计的，这就是一种通常所说的“参数估计”。一般教科书和文献中，常用希腊字母表示总体参数，如 μ ， σ ， ρ 而用相应英文字母表示统计量，如 $\bar{x}$ ， s ， r 等。还应注意，随机变量的函数仍是随机变量。样本本身就是一个（多维）随机变量，故样本的函数—统计量也是随机变量。例如样本平均数、方差、标准差等统计量都是随机变量。这从大量的矿床勘探实践中可以证明。例如从一个矿体上按同样间距取同样多样品，但前后几次起始样品的位置稍有不同时，则几组样品的平均品位各不相同，其它统计量也各不相同。因此，样本平均数、方差等随机变量也有自己的分布和相应的参数。

6. 中心极限定理

不论原始观测值的分布如何，当样本大小增加时，样本平均数的分布趋向于接近正态分布。样本平均数的分布的两个参数（平均数及方差）有如下特征：1) 样本平均数的平均数等于总体平均数，即 $\mu_{\bar{x}} = \mu$ ，2) 样本平均数的方差等于总体方差除以样本大小，即 $s_{\bar{x}}^2 = \frac{\sigma^2}{n}$ ，或 $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ 。上述定理只有当 n 足够大时才成立。对于地质变量，一般认为，当样本大小在50以上时，不论其原始分布如何，样本平均数的分布均趋于正态。