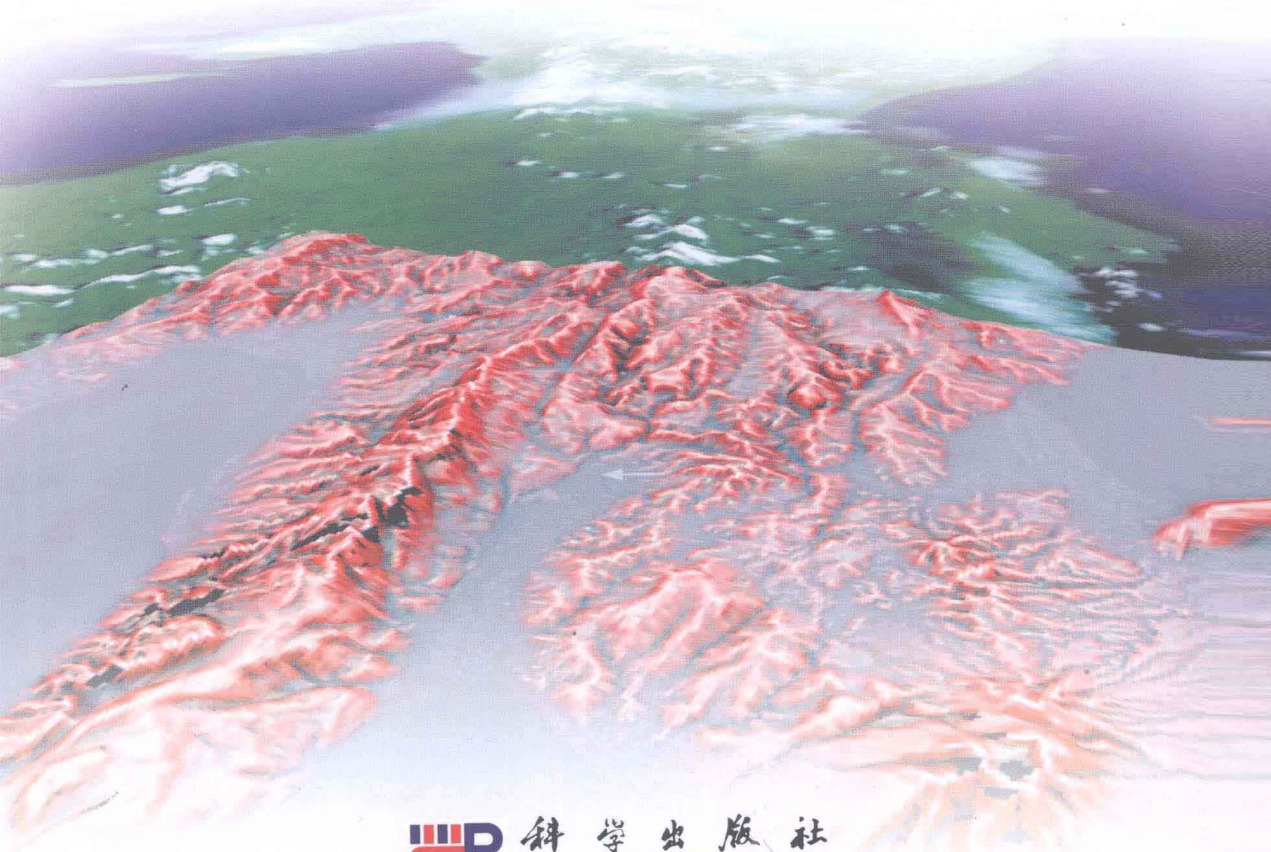
教育部高等学校地矿学科教学指导委员会推荐教材

地学数据分析教程

阳正熙　吴堑虹
彭直兴　严　冰　编著



科学出版社
www.sciencep.com

教育部高等学校地矿学科教学指导委员会推荐教材

地学数据分析教程

阳正熙　吴堑虹
彭直兴　严　冰　编著

科学出版社
北　京

内 容 简 介

本书以统计学思想为主线，分四个部分展开论述。第一部分简明扼要地阐明了如何对一元地学数据进行探索性分析和统计推理；第二部分从应用层面讨论如何对多元地学数据进行分析归纳并寻找变量之间的内在联系；第三部分深入浅出地论述了地质统计学的原理及其应用；第四部分提纲挈领地总结了一些常用地学经验图解的原理和解读。对于需要重点掌握的内容都设置了实训项目，并以光盘的形式提供。本书最显著的特点是理论与实践紧密结合，既注重突出基本概念和论述基本原理，又强调掌握基本方法和基本技能。

本书适合用作高等学校地学各专业本科生和研究生教材，也可供从事地学工作的研究人员和工程技术人员参考。

图书在版编目(CIP)数据

地学数据分析教程 / 阳正熙等编著. —北京：科学出版社，2008
教育部高等学校地矿学科教学指导委员会推荐教材
ISBN 978-7-03-022323-4

Ⅰ. 地… Ⅱ. 阳… Ⅲ. 地球科学-数据-分析-高等学校-教材 Ⅳ. P

中国版本图书馆 CIP 数据核字(2008)第 088002 号

责任编辑：郭 森 王日臣 刘希胜 / 责任校对：陈丽珠
责任印制：张克忠 / 封面设计：耕者设计工作室

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京市文林印务有限公司印刷

科学出版社发行 各地新华书店经销
*
2008 年 9 月第 一 版 开本：B5(720×1000)
2008 年 9 月第一次印刷 印张：15 3/4
印数：1—3 500 字数：293 000

定价：30.00 元(含光盘)

(如有印装质量问题，我社负责调换〈文林〉)

前　言

本书是关于如何调查、收集、整理和分析统计地学数据的一门方法论的地学专业主干课程教材，在知识体系中起到“数学知识与专业知识的继承和理解以及运用和实践”的作用，不仅能够引导学生把数学知识和计算机技术应用于地学，而且能够引发学生对地学专业课程的学习兴趣，同时鼓励学生运用所学的其他专业课程的理论和方法，分析和解决地学问题，打通学科之间的壁垒，从而开拓学生视野，为学生搭建一个既集成了地学数据分析常规方法、又能引领进一步钩深致远的学习平台。其主要教学目的是能够使学生系统地掌握地学数据处理的基础理论和基本方法，能应用所学的方法解决生产实际问题。

传统地学着重于对自然现象的描述，然而，随着计算机技术的普及，主要地学信息日益采用定量而不是定性的描述，地学工作中日益强调数据分析技术，因此，数据分析对于地学人员来说是一个必须掌握的重要技能。

本书的指导思想是力求深入浅出地为地学各专业的本科生、研究生以及地学工作者提供数据分析的思路、方法及途径，因而对于每一种方法都特别注重阐述原理、方法过程、适用条件、结果解释；尽可能淡化数学理论推导过程，并且考虑到计算机的普及，在不影响对方法理解的前提下省略一些复杂的计算过程。书中的案例经过精心挑选，尽可能结合实际，为读者创造性地应用所学知识提供范例。

全书除绪论外分为四部分进行阐述。第一部分（第 2～5 章）涉及一个变量的数据分析原理和方法。第 2 章主要介绍地学变量的类型以及取样方法；第 3 章讨论了一元数据的图形描述和数字特征归纳方法；第 4 章论述了如何利用样本数据推断总体特征的原理和方法；第 5 章着重介绍方差分析和时间序列分析方法。第二部分（第 6～9 章）涉及多变量数据分析方法。第 6 章重点阐述了相关分析和简单线性回归分析以及趋势面分析；第 7 章论述了聚类分析原理及方法；第 8 章介绍了判别分析原理及方法；第 9 章讨论了因子分析原理及方法。第三部分（第 10～12 章）着重于地质统计学的介绍。第 10 章阐明了区域化变量理论；第 11 章讨论了变差函数模型及其应用；第 12 章论述了克里金方法的原理、步骤及其应用。第四部分（第 13～16 章）比较全面地介绍了地学中常用的基于统计学原理的经验图解。第 13 章涉及主要元素的图解方法，包括二元系成分图解和三元系成分图解；第 14 章详细讨论了微量元素图解方法；第 15、16 章阐述了同位素图解方法的原理、绘制方法以及如何解读。

本书的特点是信息量大、实用性强、层次分明、基本概念清晰，并配有上机实习

指导光盘，有利于增强学生解决实际问题的能力，使学生学会利用各种统计软件处理实际数据。

本书第2～6章由中南大学吴堑虹教授编写，阳正熙教授补充修改；其余各章由成都理工大学阳正熙教授完成；所附光盘中的实习指导由成都理工大学彭直兴老师和严冰老师编写；书中图件多数由饶红娟绘制。

作者在编写本书的过程中参考了国内外许多相关的优秀教材，从中得到巨大的启迪和帮助；本书的出版得到教育部高等学校地矿学科教学指导委员会、成都理工大学教务处和中南大学地学与资源环境工程学院的大力支持，谨借此机会一并致以最诚挚的谢意！

由于我们学术水平有限，书中不足和疏漏之处在所难免，恳请读者批评指正。

阳正熙　吴堑虹

2008年5月10日

目　录

第二部分　多元地学数据分析

第三部分　地质统计学

第四部分 岩石地球化学常用的图解方法

第1章 绪 论

1.1 本课程的理念

1.1.1 本课程的目的

地学工作者掌握数据分析的技巧是非常重要的，尤为重要的是需要掌握下述三个方面的内容：

(1) 如何获得数据；

(2) 如何利用好现有数据；

(3) 如何对数据处理结果进行科学合理的解释。

本课程的目的包括：

(1) 获取数据并根据数据得出结论，或者针对问题获取数据并解释数据。我们的目的是要强调数据分析的过程及其应用而不是数学理论的推导、注重数据分析的思维而不是数据计算、强调现有软件的应用而不是计算机编程。

(2) 每个地学工作者(尤其是矿产勘查者)都有大量的数据资料，在准备报告或交流研究成果时都需要对这些数据资料进行概括。本课程的另一个目的是通过介绍实用的数据处理方法培养学生分析问题和解决问题的能力，避免把未经解释或解释错误的数据附在论文或报告上作为装饰。

(3) 把分散在各学科或课程中的数据处理方法集成为一个方法系统进行教学，培养学生使用专业知识有效地解决实际问题的能力。

(4) 使学生能够更有效地利用计算机技术进行数据处理，充分发挥现有软件的强大功能。

1.1.2 统计学思想

统计学是收集数据、分析数据并且由数据得出结论的方法学。统计学思想是本书的灵魂，一元地学数据处理是利用统计学研究一个地学变量的数据处理方法；多元地学数据处理则涉及多个地质变量的数据处理；地质统计学方法是研究数据空间分布规律的统计方法；地学中许多经验图解也是基于统计学思想构建起来的。因此，对于本课程的学习来说，最主要的任务是掌握统计学思想，理解相关的方法原理，能够根据实际情况提出解决问题的一个或几个合适的方案，并懂得选择最优方案。

我们以矿产勘查中估算矿床平均品位为例来说明统计学思想：

矿床拥有的矿石资源储量及其平均品位是矿产勘查获得的最重要成果。然而，影响矿床平均品位的因素非常多(如取样方法、样品的数量和质量、化学分析等方面)，想要准确地预见一个矿床的平均品位，需要把所有的影响因素都控制住，这无疑是做不到的，因而，就不可能确切地预见矿床(体)的平均品位，也就是说，矿床平均品位取何值呈现出不确定性。不过，这种不确定性是有一定的规律可循的。这种规律就叫做统计规律。假若我们得到了某矿床的 n 个品位数据，统计学把这些数据叫做样本，而把上述统计规律叫做总体的统计规律。统计学认为，样本数据荷载着总体的信息，可以用样本数据去推断总体的统计规律，这就是统计学思想。不难想到，当我们用样本数据去推断总体的统计规律的时候，还需要解决如何保证科学性、合理性的问题。于是，就派生出统计思想的更多内容。实际上，我们所得到的矿石品位数据是矿体实际取样分析的状况。我们把实际发生的状况叫做事物的现实状态，可以用数据直接描述事物的现实状态。

统计学思想的精髓：

(1) 从随机性中归纳出规律性。当我们不能预测某一种现象可能的结果时，随机性就和这种现象联系起来了。通过对看起来随机的现象进行统计分析，我们开始认识这个世界。统计学思想从如何观察事物和事物本身如何真正发生两个方面，帮助我们理解随机性和规律性的重要性，告诉我们如何把随机性归纳于可能的规律性中。

(2) 通过变量估计常量。统计学不是堆放全部数据的仓库，而是抽象出数字特征，用以概括表达数据的内在规律性，利用随机变量估计常量。由于数据本身的随机性，势必会引起估算误差，在这种意义下，统计学思想必然要突出对数据中的偏差问题的研究。

(3) 借助于样本的研究推断总体的特征。统计学最朴素的思想就是借助于样本推测总体的情况。由于这种推理的方法有一定的局限性，容易产生错误，因而还要求统计学采取相应的技术措施设法尽量降低在这种推论过程中发生错误的可能性。

1.2 地学数据分析的基本概念

1.2.1 数据分析的概念

1. 数据和信息

数据(data)是事实、概念或指令的形式化表现，是用于表示客观事物的未经加工的原始素材，如图形符号、数字、字母等。信息(information)是人们根据数据表现形式中所用的约定而赋予数据的意义。两者关系表现在：

(1) 数据是现实世界的信息载体，是信息的具体表现形式。信息是数据的内

涵，是形与质的关系。

(2) 数据只有对实体行为产生影响时才成为信息；数据只有经过解释才有意义，成为信息。例如，对于"1"和"0"而言，独立的1或0没有意义，而当它表示某个实体在某个地域内存在与否时，它就提供了"有"或"无"的信息；如果用它来标识某种实体的类别，那么，它就提供了特征码信息。

数据是一种宝贵的科学资源，要了解它的价值就必须确切地知道它是如何产生的。数据的价值在于我们希望利用它来解决什么问题以及希望从数据中获得何种重要的结论。

2. 地学中常见的数据类型

地学中数据种类繁多，难以枚举，在此仅以部分数据种类为例：

(1) 地形数据：地理坐标和高程；

(2) 地质构造数据：地质构造要素的定向数据(方位角、走向和倾向等)；

(3) 地球化学数据：地球物质的主要元素含量、微量元素含量、同位素成分等；

(4) 矿床数据：品位、吨位、矿产品价格等；

(5) 地球物理数据：磁法、重力、地震、电法测量数据；

(6) 岩石学数据：岩石密度、孔隙度及沉积物粒度大小、形状等；

(7) 地层古生物数据：地层颜色、厚度、时代、化石种类和数量等。

3. 数据集

有限个观测值的集合称为数据集(data set)。数据集一般是从更大的数据总体中通过取样观测获得的。因此，数据集也可以看作为样本。

(1) 一元数据集：从总体中采集的每个样品只对一个独立变量进行观测获得的一组观测值。

(2) 多元数据集：相对于位置独立变量观测一个变量值构成的数据集(例如，以 x，y，z 坐标为参照的金品位数据；参照时间"位置"的水平面变化数据；地形高程数据等)，或者在一个样品中同时观测两个或多个变量获得的数据集(例如，同一个地球化学样品中同时分析多个元素构成的数据集，有时也可以没有相应的位置信息)。

4. 数据分析

数据分析(data analysis)是指分析数据的理论和方法。数据分析的目的是把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律、验证有关的理论或假设。在实际工作中，利用数据分析结果可以有效地帮助人们作出判断，以便采取适当行动。

数据分析一般包含三个主要步骤：

(1) 探索性数据分析：原始数据往往是杂乱无章的，看不出规律。探索性数据分析是应用各种技术(主要是图形方法，也包括一些定量技术)对数据进行分析的

途径。探索性数据分析有利于对数据集的深入了解、揭示数据集中隐含的结构、提取重要的变量、确定异元值和异常值等，从而为更深层次的研究奠定基础。

探索性数据分析的图形技术都非常简单易行，可以直接采用原始数据绘图，如散点图（相关图）、直方图、概率图等；也可以采用简单的统计量进行投图，如利用平均值或标准差投图、盒须图等。

探索性数据分析过程中，需要对缺失值进行处理。所谓缺失值是指在数据采集与整理过程中丢失的内容。缺失值的处理一般有两种方式。一是删除对应的记录，如在微量元素地球化学分析中，部分样品的某个元素含量低于仪器的检测限而出现缺失，则将该元素信息全部从数据集中删掉。这种方式在数据缺失非常少的情况下是可行的，但如果各个元素的分析项目中都有少数的数据缺失，对所有缺失的记录都进行删除可能会使总样本量变得非常小，从而损失许多有用信息。缺失值处理的第二种方式是进行插值处理，所谓插值，就是指人为地用一个数值去替代缺失的数值。

（2）模型分析：在探索性分析的基础上提出一类或几类可能的模型，然后通过进一步的分析从中挑选一定的模型。

（3）推断：通常使用数理统计方法对所定模型或估值的可靠程度或精度做出推断。

5. 数据的精度和准确度

（1）精度（precision）：指观测误差。精度越低，需要越大的样本容量才能抵消数据中的噪声。精度随统计量而变，例如，如果同一总体多个样本的平均值非常相近，我们就说精度很高，因此可以把精度作为参数的估计范围。

（2）准确度（accuracy）：指估值与真值的接近程度。

6. 数据集的几个重要特征

（1）形状（shape）：数据集的形状是确定采用哪一组统计量才能够对该数据集进行最佳总结的重要因素，因而，形状应该是需要首先关注的特征。数据集形状一般分为对称的、左偏斜的或者右偏斜的以及单峰、双峰或者多峰。

（2）位置（location）：在经典统计学中，数据集的“位置”和“中心趋势的度量”这两个术语可以相互替换，但“位置”一词更加简单明了。平均值、中位数以及众数等是常用于度量数据集位置的参数（统计量）。

（3）离散性（spread）：离散性是数据集变化性的度量，表示离散性的参数（统计量）主要包括方差、标准差、四分位数间距以及极差等。

上述三个特征对于理解和解释数据集或总体的分布都是很重要的。

（4）异元值（outlier）：异元值是远离数据集群的值。地学数据集一般都含有异元值。异元值不应当简单地删除，而应当对其进行研究，因为它们可能含有研究区重要的地质信息。例如，在环境地球化学中，异元值可能指示污染区；在勘查地球化学中，异元值可能指示矿化区。因此，对于每个异元值都需要认真考察，以确

定其是否代表所研究总体的可能值。如果异元值属于所研究的总体，则应予以保留；如果不属于，则应予以剔除。

盒须图是确定异元值的最好途径之一(参见第3.3节)。

(5) 群聚性(clustering)：群聚性意味着数据趋向于围绕某个数据中心集聚。

7. 地学数据分析资料

地学数据分析资料指可用以推导出某项结论的资料数据。地学资料数据一般应包括地学数据以及研究区的地学背景资料。

地学数据分析离不开地学的专门知识和敏锐的判断力。形式化的数据分析方法只是一种辅助手段，借助于计算机技术可以帮助人们进行判断或推理。然而，地学数据分析所需要解决的问题和目的、数据及其结构、分析结果的解释等，都需要结合研究区的地学背景资料综合考虑和评价。显然，只有在相关地学理论的指导下，地学数据分析的潜力才能够得到充分发挥。而且，地学专门知识与地学数据分析方法以及计算机技术的有机结合，可能形成新的学科生长点或衍生出许多研究课题，这方面杰出的实例有法国巴黎矿业学院马特龙教授及其同事在20世纪60年代初期发展起来的地质统计学以及中国地质大学赵鹏大院士及其同事在70年代初期发展起来的地质统计预测。

1.2.2 变量及其变化性

1. 随机性和规律性

在地学领域中，大量的地质现象很难预先确定，难以用确定的公式或论述来描述，这种不确定性称为随机性(randomness)。但是，从总体上来说，这些现象又表现出一定的趋同性或稳定性，这就是随机性中的规律性(regularity)，而且，这种规律是统计规律。

在研究实际问题时，要对客观事物进行观察。观察的过程称为试验，试验的结果称为事件。事件可以分为确定性事件和随机性事件两种。在给定条件下，一定发生或一定不发生的事件分别称为必然事件和不可能事件，它们是确定性事件；在给定条件下的每一次试验中可能发生也可能不发生，而在大量试验中具有某种规律性的事件称为随机事件。

2. 变量和变异

变量(variable)是可变的数量或属性标志，变量的具体表现称为变量值，也就是数据。根据变量的变化性质可分为随机变量和确定性变量。

(1) 随机变量(random variable)：为了研究随机事件的数量规律性，把表征随机事件的变量称为随机变量；根据国际标准化组织(ISO)的定义，随机变量是指取自一组特定值中的任意值的变量，该变量与概率(可能性)分布有关。如果随机变量的取值是有穷尽的或者是可数无穷尽的，这种随机变量称为离散随机变量；如果