数学地质基础与方法

煤炭工业出版社

数学地质基础与方法

煤炭科学研究院地质勘探研究所
西安矿业学院数学教研室　编著

煤炭工业出版社

内 容 提 要

本书由概率统计及其在地质工作中的应用和数学地质方法两部分内容组成，主要包括：数据整理、随机变量与概率分布、多维随机变量及其分布、大数定理与中心极限定理、统计推断、方差分析。回归分析、趋势面分析、逐步回归分析、聚类分析、判别分析、因子分析、对应分析、典型相关分析等，并附有DJS-6计算机语言程序和常用数理统计表。可作高等院校地质专业教科书，也可供从事地质工作的工程技术人员学习参考。

数学地质基础与方法

煤炭科学研究院地质勘探研究所
西安矿业学院数学教研室 编著

*

煤炭工业出版社 出版
（北京安定门外和平北路16号）
煤炭工业出版社印刷厂 印刷
新华书店北京发行所 发行

*

开本850×1168 1/32 印张19 1/2
字数 517 千字 印数1—3,800
1981年12月第1版 1981年12月第1次印刷
书号15035·2418 定价2.90元

前　言

数学地质是把数学和电子计算机技术应用于地质学领域的一门新兴学科。它的出现，使地质学从定性描述阶段走向了定量分析阶段。它在生产实践中的显著效益和今后的发展前景引起了广大地质工作者的重视。

近些年来，数学地质在国内外发展很快，它已渗透到地质学的各个领域，如沉积学、矿床学、地层古生物学、构造地质学、水文及工程地质学、煤岩学等，在煤田地质工作中也开始了广泛的应用，并且收到了实效。随着这门新学科的不断发展和完善，它将在“开发矿业”的生产中起着更大的作用。

广大地质工作者迫切要求普及、推广和使用数学地质知识，一些高等院校为开设这门课程也急需教材。出于生产和教学的需要，我们在煤炭部“数学地质训练班”及近年来教学工作的基础上，编写了这本书，供煤田地质工作者及高等院校煤田地质勘探专业的教师和学生学习参考。

全书共分两篇：第一篇概率论与数理统计基础；第二篇多元统计分析及其在煤田地质中的应用。

考虑到当前多数地质工作人员受到数学基础的限制，我们在数学公式推导方面力求详细并附有习题和答案；为了便于读者掌握和使用，在多元统计分析方面附有若干实例及 DJS-6机语言程序；为了便于自学，在文字叙述方面力求通俗易懂。

本书由西安矿业学院、煤炭部地质勘探研究所合编。杜其仁同志为第一篇执笔者；门桂珍同志为第二篇执笔者。陈朝阳同志参加了第二篇的部分工作。

全书蒙西安矿业学院杨卜安副教授审阅，提出了不少宝贵意见。此外，在编写过程中参考了武汉地质学院，中国地质科学院数学地质组，中国科学院地质所、数学所、地球物理所，北京大

学地质系，云南大学数学系等单位的有关资料，在此向他们一并表示感谢。

由于我们水平所限，错误与不足之处在所难免，敬请读者批评指正。

书内注有“※”号的部分，初学者可以不读。

编　者

一九七九年十二月

目　录

第一篇　概率论与数理统计基础

第 一 篇

概率论与数理统计基础

第一章 数 据 整 理

随着地质科学定量化的发展，经常会遇到各种不同类型的实验数据（或观测数据）。这些实验数据所具有的信息，对我们认识事物的内在规律，研究事物之间的关系，预测事物今后可能的发展等一系列问题，是非常宝贵的。但是，要想从这些庞大的数据堆中找到有用的东西，得出可靠的结论，就必须对实验数据进行认真的整理和科学的分析，并应用数理统计的一些计算方法和处理技巧，进行去粗取精，去伪存真，充分揭露事物内部存在的矛盾，从而找到解决问题的线索或可能途径。

本章首先分析常用实验数据的一些基本特点；数据的统计整理方法；计算实验数据的基本特征，为推断实验数据的理论参数的数值和统计分布规律，提供合理的数学模型。

§1 总体、个体和样本

总体、个体和样本是数理统计中常用的名词，也是数理统计中的一些基本概念。它们的含义是：总体（或母体）是指研究对象的全体，它由实验数据的全部可能值组成，这些值不一定都不相同，在数目上也不一定是有限的；个体是指总体中的一个单位；样本（或子样）是指总体中随机抽取一部分个体的集合（在有些情况下，一个样本可能包括总体的全部）。

一个总体中含有个体的数目一般都很大，以至无穷，根据研究的对象不同而定。例如：某井田所有可能测得钻孔煤厚数据的全体构成一个总体；每个钻孔煤厚就是一个个体；随机抽取 n 个钻孔所测得煤厚的 n 个数据就是一个样本。如果我们研究的对象不是一个井田，而是整个煤田，此煤田又包括许多个井田，则把每个井田中钻孔煤层的平均厚度看作一个个体，所有这些个体的总

和就是一个总体。这里所说的总体包含的个体是有限的。数理统计就是解决如何从样本的统计特征来研究总体的特征。要从样本的统计特征来作出关于总体特征的推断，这个样本必须是总体的代表。要取得一个真正有代表性的样本，抽样的方式必须具有随机性。

通常要求样本：X_1，X_2…，X_n是相互独立的，而且与总体有相同的分布，这种样本叫做“简单随机样本”。

总体、个体和样本是就某一数量指标而言（如钻孔煤厚），和地质中的“样品”或“标本”不同，前者是若干个同类的实验数据，后者则是一个地质实体。例如：对一个“煤样”（样品）分析了碳、氢、氧三种不同元素的含量，则此“煤样”就对应有三个不同类的个体。在数理统计中不能将此三者混杂在一起，应分别作出统计。

统计总体必须是由那些从内部来说有一定的内在联系，从外部来说又有差别的许多个体组成。一般地说，在岩浆岩地区所采的样品，与在沉积岩地区所采的样品，不应放在一起作为一个总体来统计。因为它们在地质上是不同类的。

总体的性质由其中各个体的性质来确定。所以要了解总体的性质，就必须测定各个个体的性质。但是，要想测定所有个体的性质是极困难的。一方面总体中所含个体甚多，以至无穷，无法逐个测定；另一方面有的总体虽然所含个体不甚多，但是对每个个体的某种数量测定是一种具有破坏性的测定，仍不能逐个一一测定。通常采用通过了解样本的性质来推断总体的性质，也就是通过局部了解总体，构成了数理统计的主要内容和基本课题。

§2 实验数据的属性

为了要对实验数据进行有效的整理和分析，首先就要对实验数据的基本特点有所了解。这些特点是蕴含在一批数据的内部，具有一般的规律性，通过观察、分析、比较，是容易发现的。

下面以某井田六号煤层的94个钻孔煤厚的数据为例，说明常

见实验数据的一些基本特点。今测得煤厚数据如下(单位：米)：

1.03	2.29	1.44	1.43	1.92	1.40	1.51	0.25	0.72	2.16
0.70	1.50	1.46	1.02	0.85	0.52	1.23	1.31	0.62	1.19
2.87	0.54	2.19	1.31	0.95	1.19	0.76	1.15	1.43	0.68
1.57	1.11	1.43	2.47	1.05	1.61	1.63	1.35	1.05	1.23
1.27	0.37	1.25	1.04	1.49	0.93	1.07	1.44	1.69	2.52
1.95	0.85	1.20	1.19	0.95	1.67	1.01	2.48	1.80	1.47
1.22	1.64	1.13	1.77	1.25	0.89	1.53	1.25	1.37	1.42
1.56	1.23	1.31	1.26	1.27	1.12	1.11	0.74	2.33	1.72
1.95	2.03	2.28	1.80	1.92	0.55	1.96	1.88	2.13	2.05
2.22	1.70	0.81	1.75						

经过简单分析，容易看出，这些数据有如下特性，

一、波动性

设实验数据以$\{x_i\}$表示：

$$x_1,\ x_2,\ x_3,\ \cdots,\ x_n。$$

以有限次数n（这里n＝94）由离散化形式给出。这些数据虽然是在同一煤层的条件下随机测得的，但是数值并不完全一样，表现出一定的波动，这种波动反映了在不同钻孔位置煤厚的起伏变化。

二、规律性

数据虽有波动，但不是杂乱无章的，仔细观察会发现数据呈现出一定的规律。都在0.25到2.87之间；在1.10到1.70之间的机会多一些；其他范围机会少一些，只有个别数据在0.25和2.87附近。在对数据进行了一定的分析整理和统计检验之后，可以看出，实验数据大都具有一定的统计规律性。找出实验数据的统计规律，估计实验数据的基本统计参数，是数理统计计算要解决的基本问题之一。

三、实验数据的误差

由于种种因素的影响，实验数据与其真值之间存在着一定的误差。这些误差，按其性质可以分为三类。

1）随机误差：也叫试验误差，是由一系列偶然因素引起的。在多次反复实验过程中，随机误差取值可大可小，可正可负，具

有一定的统计规律。随机误差在实验过程中是难免的，随着实验次数的增加，随机误差的算术平均值将愈来愈小，逐渐接近于零。

2）系统误差：把实验过程中服从确定性规律的误差称为系统误差（或条件误差）。在多数情况下，系统误差是个常数。在实验观测过程中或数据分析整理过程中，可以通过一定的方法来识别和消除这类性质的误差。

3）过失误差：一般把明显歪曲实验结果的误差称为过失误差。把含有过失误差的实验数据称为异常点。过失误差一般是由于实验观测系统测错、传错、记错等不正常原因造成的。在数据整理过程中，必须消除这类过失误差，舍去异常点。否则会严重影响计算结果的准确性，得出不正确的结论。

在一组实验数据中，实验误差总是综合性的，即随机误差、系统误差、过失误差同时错综复杂的存在于实验数据中。在实验观测、数据收集和数据整理过程中，必须认真对待。

§3 实验数据的统计图表

我们现在的地质报告中，对于有关地质数据的使用，常以

$$\frac{\text{最大值}-\text{最小值}}{\text{平均值}}$$

的形式表示。它在一定程度上反映了这些数据的统计特征，比较直观，也容易计算。但是，仅仅采用这种统计方法对待所取得的大量数据，是很不够的，它没有能够充分利用这批数据所提供的信息，来揭露数据之间的内在联系，可能要丧失一些有价值的东西。

在数理统计中对一批实验数据的归纳整理，应充分利用数据所提供的信息，揭露数据之间的内在联系，最常用的方法，首先是对数据进行分组、列统计表和制直方图。其目的就是为了突出实验数据的统计规律。

一、分组

实验数据的数目（样本容量）较大时要进行分组，分组数目要依据容量的大小来确定。组数不宜过少，至少5组，最多也不要超过20组，以平均每组要有5个数据以上为宜。进行分组时，首先将数据按数值大小的顺序排列，然后划分数据为K个互不相交的分组区间。组距一般要相等，各组间隔的中点叫组中值，用它来代替组内各数据的平均值。分组的方法是：（1）确定实验数据的上界和下界；上界的确定可以比数据最大值稍大些、下界可以比数据的最小值稍小些，即允许向外伸延些。（2）确定分组数K和分组点数值，定出组距 $l = \frac{\text{上界} - \text{下界}}{K}$。（3）以唱票方式统计落入各组内的实验数据的个数，称为频数。对于微量元素，实验数据的数值变化常达好几个级次，可以先取对数，然后将对数值按等组距分组。

划分实验数据的分组区间，是数据整理过程中比较麻烦的一步，但是，它是非常重要的一步，要认真耐心地进行。确定数据应当分几组为宜和组距的大小时，要以能突出数据比较集中处的频数为准。分组点要尽量避免与数据值相重叠，同时要避免出现有的组中不含数据的现象发生。

二、列统计表

实验数据经过分组后，列出统计表。它可以把大量的原始数据压缩成较为精练的形式表示出来，并为制作直方图提供一个依据。统计表的项目主要包括：组段、组中值、频数、频率，累积频率等。频数μ_i是指实验数据落入第i组内的个数。各组频数之和等于实验数据的总个数n（子样容量），即

$$n = \sum_{i=1}^{K} \mu_i$$

式中　K为组数。

频率f_i是指第i组内的频数μ_i与实验数据的个数n之比，各组频率之和应等于1，即

$$f_i = \frac{\mu_i}{n} \quad (i = 1, 2, \cdots K)$$

$$\sum_{i=1}^{K} f_i = \sum_{i=1}^{K} \frac{\mu_i}{n} = \frac{1}{n} \sum_{i=1}^{K} \mu_i = 1$$

累积频率 F_i 是指各组内频率依次相加之和，总的累积频率为1。

$$F = \sum_{i=1}^{K} f_i = 1 \qquad (1\text{-}1\text{-}1)$$

或者 $$F_K = 1$$

利用这些关系式可以检查统计表中各项计算有无错误。它的理论含义指出：任何一个统计整体的总频率恒等于1。

例如：某井田六号煤层94个钻孔煤厚数据，最大值为2.87，最小值为0.25，按数据大小共分9组。上界取2.90；下界取0.20；组距的大小为$l = \frac{2.90 - 0.20}{9} = 0.30$。见表1-1-1。

表 1-1-1

组号	组段	组中值	频数 μ_i	频率 f_i	累积频率 F_i
1	0.20—0.50	0.35	2	0.021	0.021
2	0.50—0.80	0.65	9	0.095	0.116
3	0.80—1.10	0.95	14	0.149	0.265
4	1.10—1.40	1.25	25	0.266	0.531
5	1.40—1.70	1.55	19	0.202	0.733
6	1.70—2.00	1.85	12	0.128	0.861
7	2.00—2.30	2.15	8	0.086	0.947
8	2.30—2.60	2.45	4	0.043	0.990
9	2.60—2.90	2.75	1	0.010	1.000
Σ			94	1.000	

三、制图

主要是制等距离频数（频率）分布直方图和累积频率多角形图。分别介绍如下。