中等专业学校教材

工程地质数据统计初步

武汉电力学校 袁开先 编

中等专业学校教材

工程地质数据统计初步

武汉电力学校 袁开先 编

水利电力出版社

内 容 提 要

本书为水利电力类中等专业学校教材，适用于工程地质及水文地质专业，并可供有关工程技术人员参考。

全书共5章。内容包括工程地质数据系统、参数的估计、统计假设检验、相关分析与一元回归分析、多元回归分析。书中注意了统计方法的工具性和专业的适用性。各章附有习题。书后附有BASIC程序供参考。

中等专业学校教材

工程地质数据统计初步

武汉电力学校　袁开先 编

*

水利电力出版社出版

（北京三里河路6号）

新华书店北京发行所发行·各地新华书店经售

水利电力出版社印刷厂印刷

*

787×1092毫米　16开本　9.5印张　210千字

1991年5月第一版　1991年5月北京第一次印刷

印数 0001— 1700册

ISBN 7-120-01285-1/TV·444

定价2.35元

前　言

本教材是根据原水利电力部中等专业学校水电类教学研究会审定的本门课程的教学大纲编写的。全书共分五章：第一章主要介绍工程地质数据的基本特征及其整理方法；第二章主要讲述数据特征参数的概率估计；第三章介绍常用的参数假设检验和非参数假设检验；第四章主要讲述相关分析和一元回归分析；第五章简要介绍多元回归分析的基本方法。

本书不是数学上的数理统计教材，它偏重于数理统计的基本方法在专业上的应用。教材中免去了许多数学上的证明和公式推导，适当引进了近年来出现的一些新技术、新方法，以便扩大学生视野。书后附有一些微机应用的BASIC程序供参考使用。通过本书的教学，使学生在原有专业知识的基础上，获得对工程地质数据整理和基本分析方法的技术训练，这对从事生产第一线工作的工程地质及水文地质技术人员来说，无疑是必要的。书中有“※”号的章节为选学内容。

全书由武汉电力学校高级讲师陈运志审阅，提出了许多宝贵的修改意见；武汉水利电力学院袁美月副教授对本书也提出许多修改意见。遵照他们的意见，本书进行了修改。虽然如此，难免挂一漏万，不妥和错误之处，祈望指正。

编者　袁开先

1990年6月

目　　录

绪　论

“工程地质数据统计”是用概率统计的方法研究工程岩土体及其环境的性状和行为的学科。这门课程是为了适应工程地质勘察向定量化方向发展而开设的。它的任务是研究表征工程岩土体及其环境性状和行为的工程地质测试数据的分布规律，为工程提供可信的设计参数，对岩土工程进行可靠性评价，分析工程地质数据系统中变量之间关系的密切程度和数量关系，进而对它们的时空分布和变化作出预测，以便利用有利因素，控制和防止不利因素的发展，为国民经济建设服务。

工程地质学自从20世纪30年代后期以一门新的学科从地质学中分离出来后，一直是以“工程与地质”的结合，为工程建设服务而著称。但工程地质勘察工作却是以地质学的研究方法为其主要手段，以探讨地质体的规律为其主要研究内容，以提出工程地段的工程地质条件评价为主要任务。新中国成立以来，工程地质工作确为我国国民经济建设提供了极为宝贵的基础建设资料，完成了近千个大、中、小型水库坝址的勘察工作和众多的工民建勘查工作，为社会主义建设作出了很大贡献。但随着工程建设和工程技术的发展，原有的工作方法和工程勘察体制已适应不了工程设计和运营的要求，工程地质勘察需要向定量化的方向发展。众所周知，工程规划及工程方案的优化和设计，都需要通过一系列的定量分析来实现，没有定量工作，就无从鉴别工程设计的准确性、有效性和经济合理性。在目前国家要求推行岩土工程体制的情况下，作为工程地质工作者，应当努力去学习和掌握定量化的数学方法，以适应参与工程（特别是基础工程）的设计、施工、监测和管理的需要。

在工程地质工作中，人们已认识到工程岩土体的非均质性、各向异性，以及随时间而变化的可变性。同时，也认识到对岩土体测试成果的非唯一性。但在参数的选取、稳定性分析计算、强度和变形的分析评价等方面，我们又常将岩土体视为具有某种平均性质的“均质”材料，忽视了众多“随机”因素的作用，这使我们花了巨大的代价去搜集来的资料不能有效地反映在成果中来。更重要的是，我们未注意应用概率统计这个有力的工具去指导勘察工作和成果分析，使勘察成果不能得到数量上的有说服力的论证，设计人员在应用成果进行工程设计时，当然也就缺乏足够的对工程的可靠性论证。

在工程地质勘察中，不仅要科学地进行野外调查工作，还要进行大量的勘探、试验和观测工作。然而，无论是野外调查还是勘探试验，都不能无限度地进行下去，常常需要根据部分资料去推断全体的情况。那么，怎样去进行这种合理性的推断呢？在岩土试验中，怎样取样才能使样品不受较大扰动，从而保证试验指标的正确性和代表性呢？取样数量如何确定？如何估计试验中的误差？对试验值怎样判断异常、怎样检验和插补等等，这些问题的解决都要借助概率统计这门数学工具。再如岩土体的成因分析，工程地质性质合理分类，不同地区岩土体性质类比，各种工程地质因素（或作用）及其与环境因素（或作用）之间的相关分析、数量关系，岩土工程结构物的稳定性分析和预测等等，也都离不开概

率统计法。

现在，应用数学方法，以电子计算机为手段定量地研究各种地质问题已形成一门新的学科——《数学地质》。“工程地质数据统计”是《数学地质》的一个基本组成部分，也是工程地质学的一个分支。70年代以来，在工程地质、水文地质工作中，应用《数学地质》的理论和方法解决实际问题已日益广泛。特别是引进电脑技术以后，用数学模型去模拟工程地质作用过程，预测它的发展，合理推断因素之间的相关性，数据库的建立，数据采集与自动处理，岩土工程的可行性、可靠性论证诸方面的工作已逐渐广泛展开，并取得了较好的效果。

还必须指出，数学和电子计算机技术确能很好地帮助我们定量地分析各种工程地质问题，确实推动了工程地质、水文地质向更深更广阔的方向发展。但我们切不可丢掉工程地质、水文地质学原理中定性分析的基本方法，否则我们的应用就成了空中楼阁。正确的作法是使定性方法和定量方法两者结合，相互补充、协同发展。例如前面提到关于岩土体可变性问题，这种可变性包含有两个分量，即趋势分量和随机分量。趋势分量表明趋势性变化，受控于工程地质学原理，即受确定性法则的支配；随机分量表明随机因素的影响，受概率法则的支配。描述这种可变性的数据，通常都是这两个法则综合支配的结果，在整理、分析这些数据时，决不能脱离工程地质原理的基本分析。

考虑到中专的特点和基本要求，这门课程定名为《工程地质数据统计初步》。因此，课程内容不可能完全回答前面提出的所有问题，而把侧重点放在对观察、测试数据的整理和基本分析上，以使学生获得统计学中的最基本的知识和基本技能训练，为今后工作打下一定基础。本课程安排在高等数学、概率论和专业理论基础等课程讲完之后开设，书中涉及上述有关内容时，将不再展开论述。

第一章　工程地质数据系统

在水文地质、工程地质工作中，经常会遇到各种不同类型的观测数据或试验数据。这些数据所具有的信息，对认识事物的内在规律，研究事物之间的关系，预测它们的可能发展，进而利用有利因素，控制与防治不利作用的产生与发展，都是非常重要的。但要想从这些数据堆中找到有用的东西，得出可靠的结论，还必须对它们进行科学的整理加工和分析，应用数学的方法和处理技巧，去粗取精、去伪存真，充分揭露事物内部存在的矛盾，才能求得问题的解决。

本章主要介绍工程地质数据系统的基本概念，数据的基本特征，误差的产生，数据整理方法，经验分布和理论分布的概念等。上述这些内容是今后数据分析的基础。

第一节　工程地质数据系统的基本概念

一、统计单元的概念

大家知道，我们研究的对象——工程岩土体，它不仅经历过漫长而复杂的地质历史变迁和地质营力的作用，而且在现代内外动力的作用下（包括人类活动），正在以前所未有的速度变化着。因此，地质体与周围环境是一个相互影响、相互制约、相互联系的有机整体，形成一个地质——工程——环境紧密相关的复杂系统。这样，我们研究的中心——工程岩土体的物质成分、结构构造、性质状态、活动特征等诸方面，就会在不同的空间、不同的时间上表现迥异。当我们用数据去描述这些特征的观察试验结果时，数据的不一致性是显然的。然而，统计规律告诉我们，在一定条件下，一定范围、一定时间域内，数据的分布仍有规律可循。

例如，测定土层的密度，设它是流水作用形成的砂层。砂层的密度值在垂直和水平方向上都不会是同一数值。其原因除了测试过程中产生的某些误差外，还在于形成砂层时，水流对砂的搬运、沉积作用并不是时时处处都相同的。众所周知，水流动能取决于挟砂力（与流速有关）和输砂量（与流量大小和河床特性有关）的比值，只要输砂量大于挟砂力就会产生沉积作用，而挟砂力和输砂量又可能因许多偶然因素的作用，处处表现不一。也就是说，沉积作用既受规律性因素的支配，又受随机因素的影响。因此，在砂的沉积作用中，在相似的环境下，将会形成以某一粒径为主，伴以或粗或细的粒度成分的沉积物，砂层的密度值也就会大体限制在一定范围内，不会偏离很大。对于坚硬的基岩，这样一种不均一的各向异性的非连续介质，其性质（包括物理、水理、力学等性质），也同样具有某种地域的相似性。这种差异中存在着共性，即矛盾对立统一规律，就是我们进行数据统计分析的基础。为了寻找上述统计规律，必须首先划分具有上述特性相似的区域，即划分统计单元。在一个单元内可以认为随机因素起主导作用，从而能够利用数理统计的知识进行

资料的整理分析。

什么是统计单元呢？当我们在研究与工程活动（或人类其它活动）有关的某一地区地质体时，把其中成因、岩性、表现特性（例如岩体的完整性、结构面的发育特征、岩石的强度、含水性和透水性等）基本一致的地段（垂直或水平）划为一个工程地质单元，对该单元内岩土体的工程地质特性及环境因素（常用某些指标表示）进行观察、试验、量测，获取指标数据进行统计分析，就称这种地质单元为统计单元。

二、总体、样本的概念

总体和样本是今后学习中经常遇到的两个重要术语。一个统计单元内，研究对象的全体称为总体（或称母体），组成总体的每一个基本单位称为个体。因此，在一个统计单元内，随研究对象不同，就有不同含义的总体。例如研究某软弱夹层的抗剪强度，它有两个参数 c、φ 值，c、φ 值的全体就构成两个不同含义的总体，每一个可能的 c 值或 φ 值就构成该总体中的个体。由于我们在研究时，常常关心的是研究对象的指标值及其大小，因此，把总体定义为研究对象一切可能的观测值。其中，每一个可能的观测值就是个体。组成总体的个体可能是有限的，也可能是无限的。要对总体作出某种推断，得出合理的结论，通常都不可能、也无必要对每一个个体都进行观测。因为即使是有限总体，也要考虑其经济性和可能由试验带来的破坏性。统计学的任务之一就是研究如何合理地从总体中获取有代表性的一部分个体（称为抽样技术，这是一个专门课题，限于篇幅，本书不作介绍），通过这一部分个体的分析达到对总体的推断。由这一部分个体构成的集合，称为样本。对样本进行观测，得到一组数值（x_1，x_2，…，x_n），称为样本值，n 称为样本容量。

需要指出，这里指的总体、个体和样本，是就统计单元中某一特性指标而言，它与地质上的“样品”、“标本”含义不同。样品、标本是工程地质单元中的实体。一个样品可以包含不只一个同类（即同一特性指标值）个体，也可能包含若干不同类的个体。例如对某一岩土样品要求测定它的 c、φ 值，c 值、φ 值是隶属于不同类的两个总体，而且它们就不只有一个值，作统计时，自然不能将 c 值和 φ 值混合一起。

为了获得具有代表性的样本，抽样必须具有随机性。如果一个容量为 n 的样本，样本值相互独立，且与总体有相同的分布，这样的样本称为简单随机样本。以后不特加申明，所指的样本就是指简单随机样本。通常将样本容量 $n \geqslant 30$ 的样本称为大样本，$n < 30$ 的样本称为小样本。所谓随机抽样，就是要求在抽样时，使总体中的个体具有同等的机会被抽取，且被抽取的个体并不影响总体的组成成分。

顺便指出，这里讲的总体，是从统计学的概念出发，所以称它为统计总体。对于工程地质单元来说，因为是研究的目标，所以称为目标总体。一个目标总体往往含有多个统计总体。在今后叙述中，凡讲总体都是指统计总体。

由上述可知，要正确地进行统计分析，必须首先确定研究目标和划分工程地质单元。工程地质单元的划分，主要依靠所掌握的专业知识，以野外工程地质研究的成果为依据，并与工程部位、受力方式、荷载大小等工程条件结合进行划分。此外，单元划分的粗细程度还与勘测设计的精度要求有关，不同勘测阶段，有不同的要求，这方面的知识在有关的

专业课中讲述。

第二节　工程地质数据的类型及数据的基本特征

一、工程地质数据类型

在工程地质工作中，数据可能来自不受控制的自然过程，也可能来自受人为控制的试验，前者如地震震级、岩层产状、厚度、风化、物质组成、岩体变形破坏等等；后者如实验室或野外现场对岩土的物理、水理、力学性质试验和各种物理模拟试验等。这些数据可以通过观察和试验仪器获得，根据数据性质不同，可将它们分为定性数据和定量数据两类。

定性数据是表示事物性质差异的数（或差异程度的数）。在野外的观察中，常用到许多定性术语，如岩体是风化的新鲜的，是透水的不透水的，是强度不高的强度高的，土层是软土或坚硬的土，是膨胀土或是非膨胀土，是湿陷性黄土或是非湿陷性黄土等等。这些术语中只有性质上的不同，没有数量的变化，但是通过数量化方法，可以给它们赋值。若按普通集合中二值逻辑量化，属于某种性质赋值为1，不属于该性质就赋值零，量化取值只有{0，1}两个值。这种基于二值逻辑基础上的量化方法，反映了人类对界限分明事物的精确思维方式；但是，人们在思维活动中除了这种非此即彼的思维方式外，大量是进行模糊思维，这种思维方式反映了人们在观察和处理性质复杂事物时的高度灵活性和合理性。本来人类在日常交往和社会活动中，在描述性质复杂的客观事物时，就大量地使用着模糊语言，例如人的个子高矮胖瘦、年青年老、聪明、漂亮等。在地质语言中，岩石风化严重轻微、岩体强度高低、岩体质量好坏、裂隙发育的强弱、工程稳定程度的高低等等。这语言反映了客观事物差异的中间过渡中的不分明性。怎样量化这些语言，并使之能进行数学运算，得出有意义的论断呢？建立在多值逻辑基础上的模糊数学就是研究解决这类模糊现象的学科。模糊数学把研究对象具有某种性质或情况可能性程度的大小，用一个在[0，1]闭区间的实数来描述，其值称为隶属度，研究对象性质或情况的差异，表示为隶属函数，一般记为$\mu(x)$，x称为基础变量。隶属函数的确定方法很多，如可用统计方法获取基础变量后，用频率表示；可以根据经验给出隶属函数的型式；还可请专家直接评分法确定。下面举出一个简单例子来说明隶属度的确定方法。

对岩石强度的性质，常描述为坚硬的、软弱的，其间有许多中间的过渡状态，人们常用“很”、“较”、“一般”等来形容。怎样量化它们呢？设基础变量为岩石的干抗压强度，且认为干抗压强度≥100MPa者为坚硬的，并赋值为1；干抗压强度≤20MPa者为软弱的，并赋值为零。中间过渡状态设用一个简单的线性函数表示，即

$$\mu(x)=\frac{x}{80}-0.25 \quad (20\leqslant x\leqslant 100)$$

为比较三个不同建筑地段的建基面上岩石的强度，各地段分别取三组样进行干抗压强度试验。其试验值及按上式计算的隶属度$\mu(x)$值一并列于表1-1，表中显示了Ⅱ地段岩石的强度较其它两地段都好。

表 1-1 某工程三个地段上岩石坚硬程度隶属度表

组	项目 \ 地段	I	II	III
1	强度(MPa)	20.96	78.4	80.00
	$\mu(x)$	0.012	0.73	0.75
2	强度(MPa)	4.16	92.56	85.6
	$\mu(x)$	0.27	0.907	0.82
3	强度(MPa)	72.0	60.00	53.6
	$\mu(x)$	0.65	0.50	0.42

在求得隶属度后，根据模糊数学的若干定理和运算规则，把模糊集合化为普通集合问题，得到问题的求解。应用模糊数学解决工程地质分析中的问题，已日益为人们所重视，并取得了很多有实际意义的结果。

模糊数学是一门新兴的数学分支学科，它诞生于1965年，创始人是美国自动控制专家扎德（L. A. Zaded）教授。这门学科涉及的内容极为广泛，有志于学习这方面知识的同学可以参阅文献[5]及其它有关书籍，本书就不展开叙述了。

定量数据是用仪器对研究事物进行量测所获得的数据，它是我们今后讨论的重点。由于这些数据是通过野外现场测试和实验室量测得到，所以把它们统称为测试数据，简称数据。

二、数据的基本特征

为了有效地对采集到的数据进行整理分析，必须首先对数据的分布特征有一个明确的了解。下面结合一个例子来说明这种分布特征。

表1-2是为了研究某工程地质单元内某土层的抗剪强度，从勘探钻孔中取土样，在0.2MPa垂直荷载作用下进行抗剪试验所获得的数据。这40个抗剪强度值组成一个大样本，样本容量 $n=40$，每一个试验值就是一个个体，该土层在0.2MPa垂直荷载作用下的一切可能的试验值，就是它的抗剪强度总体。细心地观察这个样本，就会发现数据具有以下基本特征：

（1）数据具有波动性　数据的波动特性表现在它的非唯一性和具有一定的波动范围上。数据的最大值 $x_{max}=1.46$，最小值 $x_{min}=0.71$，极差 $R=x_{max}-x_{min}=0.75$。也就是说，数据是在[0.71, 1.46]区间内，以变幅 R 上下波动。

（2）数据波动的规律性　数据虽然有波动，但并不是杂乱无章的紊动，多数数据集中在一个较小的范围内（0.9～1.2之间），在此区间范围之外的数据不多，特别大和特别

表 1-2　　0.2MPa垂直荷载作用某土层的抗剪强度数据表　　单位：10^{-1}MPa

序号	试验值	序号	试验值	序号	试验值	序号	试验值	序号	试验值
1	0.77	9	0.83	17	1.14	25	1.13	33	1.22
2	0.71	10	1.26	18	1.02	26	0.92	34	1.08
3	0.74	11	1.24	19	1.07	27	1.04	35	1.17
4	1.31	12	0.85	20	1.12	28	1.18	36	1.03
5	1.46	13	0.86	21	0.91	29	0.93	37	1.18
6	0.72	14	1.23	22	0.96	30	0.98	38	1.21
7	1.35	15	0.97	23	1.06	31	1.06	39	1.06
8	0.82	16	1.04	24	1.11	32	0.94	40	0.88

小的数据更少。也就是说数据的分布具有相对集中的趋势和分布的某种程度的对称性。

当然，凭直觉观察数据，得出上述认识是困难的。以后我们通过数据的整理，这些基本特征就会明显地表现出来，并且还可以用相应的数学模型来描述。根据统计规律，如果再从同一总体中在相同的条件下获得另一个样本，其数据的波动特征将与前一个样本相似。

第三节　误差的来源及其分类

上述数据的非唯一性，说明试验值x_i与研究对象的本质特性（即对象的物理量或真值）a之间产生了偏差e_i，即

$$e_i = x_i - a$$

那么，误差是怎样产生的呢？产生误差的原因很多，它可以是在广阔的空间和漫长的地质历史时间中，由复杂的地球动力学规律支配而产生，表现为地质体的非均质性、各向异性，甚至非连续性特征。可以在采样、运输、保管及测试过程中产生。还可以是在测试时由于气象要素（如温度、压力、湿度等）的变化而产生。归纳起来有以下三类性质不同的误差。

1.条件误差（系统误差）

服从于确定性规律所产生的误差称为条件误差，又称系统误差。例如，断层破碎带的岩石和非断层破碎带的同类岩石，在地质历史时期中受动力作用的强度是大相径庭的，这两类岩石的物理、力学特性自然差异就大；又如试验仪器的某些缺陷，岩土样在采集、运输、保管过程中条件的重大改变（如取样扰动、失水、震动、某些细小沉积物颗粒流失，试验时岩土样的环境条件——温度、压力等的重大差异等）也会造成数据大小的差异。通常条件误差不会使数据呈现有规律的波动，要么偏大，要么偏小。这类误差可以通过合理划分统计单元、选择和校准仪器、严格取样、试验的操作规程来消除；如果样本值中只有个别数据是由条件误差引起，还可通过数据整理分析来识别。如果样本是来自两个不同的总体，可以通过统计分析把它们筛分出来。

2.随机误差（偶然误差）

随机误差又称偶然误差，它是由许多偶然的因素变化引起的误差。这些因素是大量的、时隐时现的、瞬息即逝的、变化多端的、不确定的和不能完全预测的。在学习概率论时，曾接触过这方面的例子。在岩土试验中也有类似的情况，例如对某岩石进行抗压强度试验，虽然根据专业知识知道它的大概试验值，但在试验结束前，它的确切数值是不知道的。随机误差是不可避免的，是引起数据有规律波动的主要原因。随机误差可正可负，并随着试验次数的增大，其算术平均值愈趋近于零。由于随机误差是后面学习的重点，故把它简称为误差。一切试验值中误差总是存在，试验获得的数据仅仅是真值的近似值，真值常常只能通过试验值来作出估计和推断。

3.过失误差

一般把明显歪曲试验结果的误差称为过失误差。它是由测试系统测错、传错、记错等不正常的原因造成。在数据整理过程中，必须消除这些误差，否则影响结果的正确性。这

些异常值或是特别大或是特别小，通过检查数据往往会被发现。

在一个样本数据中，上述三种误差可能同时存在，在整理数据时，应仔细察看分析，把明显的异常值检出备查。

第四节　数据整理方法及数据分布特征

一、数据整理方法

数据整理工作大致可分为三步：即检查数据、统计分组、绘制图表。

1.检查数据

数据检查的目的在于保证它的正确性、完整性和系统性。检查项目包括数据是否齐全，各项记录是否完整，有无错记、重复或遗漏，有无相互矛盾或不合理的数据等。经详细检查后，对数据进行必要的补充、更正或删除，然后再进行下一步工作。

2.统计分组

这一步是数据整理的中心内容，目的是找出数据的分布规律。分组的方法有单项式和组距式两种。

单项式法就是不分组的统计方法，它是直接将数据代入计算公式计算数据特征参数的方法。

组距式法是将数据按一定规则分成若干组以获得数据分布规律和特征参数的统计方法。各组的数据是一个区间数，区间与区间之间的数首尾相接，各区间内首尾两端数之差称为组距。分组时一般按等组距划分，有时根据需要，也可按不等距分组。下面以上节表1-2数据为例，具体介绍等组距分组法的步骤。

1）将数据排成有大小顺序的数列，如从大到小排列为 $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n$。也可从小到大排列。并求极差 R；$R=0.75$。

2）确定组数（K）。分组组数的多少视样本容量的大小而定，至少分5组，最多分20组，以能代表数据的分布特征为原则，自行分组。也可参照以下经验公式或按表1-3的数字确定。

$$K=1+3.32\lg n \qquad (1\text{-}1)$$

式中　n——样本容量。

表1-3　　分组组数与样本容量关系表

样本容量	32	64	128	256	512	1024	2048	4096
组　数	6	7	8	9	10	11	12	13

表1-2样本容量 $n=40$，按式（1-1）计算：$K=1+3.32\lg 40=6.32$，四舍五入，取整数，令 $K=6$。

3）计算组距和确定组限。组距 d 按下式计算：

$$d=\frac{R}{K} \tag{1-2}$$

本例：$d=\frac{0.75}{6}=0.125$。各组区间值按d值划分；各区间按d进行累减（以x_{max}为准）或累加（以x_{min}为准），形成组序。例如累减时的组序是：$[x_{max}-Kd,\ x_{max}-(K-1)d]$，$[x_{max}-(K-1)d,\ x_{max}-(K-2)d]$，…，$[x_{max}-2d,\ x_{max}-d]$，$[x_{max}-d,\ x_{min}]$；累加法的组序是：$[x_{min},\ x_{min}+d]$，$[x_{min}+d,\ x_{min}+2d]$，…，$[x_{min}+(K-1)d,\ x_{min}+Kd]$。这样就可把数据分为$K$组了。这时要检查有无数据落在上述各区间的界限值上，若有，就应适当扩大组距位数。本例按累减法划分区间后发现数据1.21落在界限点上，决定将组距定为：$d=0.1255$，以使所有数据都落在各区间内。组序确定以后，按表1-4将其填入“分组”栏内。

表1-4　　数　据　统　计　表

组　序	分　组	$\bar{x}_i$	m_i	w_i	W_i(%)
1	0.707～0.8325	0.7698	6	0.150	15
2	0.8325～0.9580	0.8953	7	0.175	32.5
3	0.9580～1.0835	1.0208	12	0.300	62.5
4	1.0835～1.2090	1.1463	7	0.175	80.0
5	1.2090～1.3345	1.2718	6	0.150	95.0
6	1.3345～1.4600	1.3973	2	0.050	100.0
Σ			40	1.000	

4）计算组间中值x_i、落入各组内数据的个数——频数m_i、各组数据出现的频率w_i及累积频率W_i。

$$\bar{x}_i=\frac{\text{组上限值}+\text{组下限值}}{2} \tag{1-3}$$

m_i可用点名唱票办法求得，频率w_i和累积频率W_i按下两式求得：

$$w_i=\frac{m_i}{n} \tag{1-4}$$

$$W_i=\sum_{j=1}^{i}w_j \tag{1-5}$$

将计算结果填入表1-4相应栏内。填表时，无论是用累减法或累加法，组序都按从小到大排列。

5）绘制数据分布图。一般要求绘制两个图，一个是频率密度直方图，一个是累积频率直方图。

频率密度直方图（又称经验频率密度分布图），该图是在直角坐标系中，用长方形的长度表示频率密度（其值等于频率除以组距）、宽度表示各组组距的图件。根据表1-4绘制的频率密度直方图如图1-1。该图上展现出在[0.958，1.0835]区间内数据最多，呈单峰状向两侧近似对称降低，直方图形面积总和为1。

累积频率直方图也是用长方形表示的，图的横坐标表示数据区间值，纵坐标表示累积

频率百分数值。根据表1-4绘制的累积频率直方图如图1-2，在该图上与50%累积频率所对应的横坐标上的A点，为中位数数值。

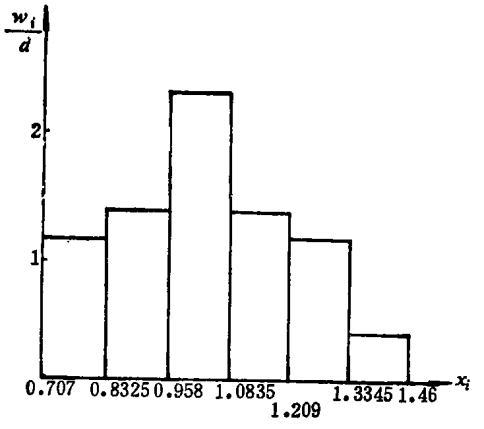

图 1-1　频率密度直方图

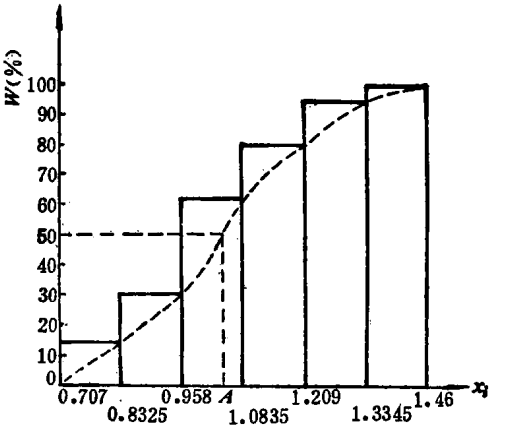

图 1-2　累积频率直方图

上述两个图虽然可以大体了解数据的分布情况和集中趋势，但还不能确切表现出数据的分布性质和数量特征，还需要用下面的计算方法把数据分布的特征值表示出来。

二、数据分布特征的定量计算

1.数据集中趋势的代表值

数据集中趋势的代表值常用的有平均值、中位数、众数等。

（1）平均值　平均值是代表数据整体特征的数。工程中常用算术平均值来表示。此外还有几何平均值、大值平均值、小值平均值等。

1）算术平均值：在表示频率分布的量当中，最常用的是算术平均值。它能反映出总体的平均水平和数据的集中趋势，一般将其简称为样本均值。计算方法有二：

①简单算术平均值。设有n个数据x_1，x_2，…，x_n，其算术平均值记为$\bar{x}$：

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (1\text{-}6)$$

表1-2中40个数据求得$\bar{x}=1.039$。

②加权平均值。简单算术平均值计算时，不考虑各数值在计算中起作用的大小。事实上在考察某些具体问题时，样本中某些数据在平均值中起作用大，应该获得较大的权，起的作用小，权就应该小。例如在工程地质问题中，关于岩石的强度指标，就可能考虑到该岩石在某一岩体平面上所占面积的大小；在求层状含水层平均渗透系数时，要考虑在渗透水流场内不同透水性岩石的厚度等。这时就应该用加权平均值。上述实际问题中的面积、厚度的大小可用f_i表示，权定义为：$\frac{f_i}{\sum_{i=1}^{n}f_i}$。

在分组统计时，各组的组中值x_i是该组m_i个数据的代表值，一共有K组，按权的概念，$f_i=m_i$，$\sum_{i=1}^{K}f_i=\sum_{i=1}^{K}m_i=n$，所以权就是频率$w_i=\frac{m_i}{n}$，其算术平均值为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{K}m_i\bar{x}_i=\sum_{i=1}^{K}w_i\bar{x}_i \qquad (1\text{-}7)$$

即样本的算术平均值是各组组中值与其对应频率乘积的总和。

上述算术平均值从计算方法上要说明以下几点：

第一，样本数据与算术平均值的离差为零，其离差平方和最小。这可以从以下推演中得到证明。

分组统计时，每组有m_i个离差$\bar{x}_i-\bar{x}$，故

$$\sum_{i=1}^{K}m_i(\bar{x}_i-\bar{x})=\sum_{i=1}^{K}m_i\bar{x}_i-\bar{x}\sum_{i=1}^{K}m_i=n\sum_{i=1}^{K}w_i\bar{x}_i-n\bar{x}=0$$

同理，可推出$\sum_{i=1}^{n}(x_i-\bar{x})=0$。

离差平方和为$\sum_{i=1}^{n}(x_i-\bar{x})^2$，设$x_0$为非$\bar{x}$的任意实数，则有$\bar{x}=x_0-c$，其中$c$为常数。于是

$$\sum_{i=1}^{n}(x_i-x_0)^2=\sum_{i=1}^{n}[x_i-(\bar{x}+c)]^2=\sum_{i=1}^{n}[(x_i-\bar{x})-c]^2=\sum_{i=1}^{n}(x_i-\bar{x})^2-2c\sum_{i=1}^{n}(x_i-\bar{x})+nc^2$$

因为$\sum_{i=1}^{n}(x_i-\bar{x})=0$，所以

$$\sum_{i=1}^{n}(x_i-x_0)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+nc^2$$

因为$c^2>0$，所以

$$\sum_{i=1}^{n}(x_i-x_0)^2>\sum_{i=1}^{n}(x_i-\bar{x})^2$$

因此，样本数据与其算术平均值的离差平方和最小。

第二，算术平均值是样本中所有数据的代表，因而它受极端值的影响，数据中的最大值和最小值会影响它的代表性。

第三，在分组统计中，计算的加权平均值，是假定各区间内的实际数据用m_i个$\bar{x}_i$来表示的。但实际情况并非完全如此。故有时这种加权平均值与简单算术平均值结果不相等，分组过粗，这种差异就明显。例如，按表1-4计算的加权平均值就等于1.04。

2）几何平均值：在工程地质数据统计中，有时要研究诸如边坡变形速率、地基在外荷载作用下的变形速率，对比各发展阶段平均速率的变化等有关速度发展情况，这时要用到几何平均值$\bar{x}_g$，它定义为n个数据连乘积的n次方根，即

$$\bar{x}_g=\sqrt[n]{x_1\cdot x_2\cdot x_3\cdot\cdots\cdot x_n} \qquad (1\text{-}8)$$

为便于计算，常将式（1-8）作对数变换，再用反对数求$\bar{x}_g$，即

$$\lg\bar{x}_g=\frac{1}{n}\sum_{i=1}^{n}\lg x_i \qquad (1\text{-}9)$$

在分组统计时：

$$\bar{x}_g=\sqrt[n]{\bar{x}_1^{m_1}\cdot\bar{x}_2^{m_2}\cdot\cdots\cdot\bar{x}_K^{m_K}} \qquad (1\text{-}10)$$

$$\lg\bar{x}_g=\frac{1}{n}\sum_{i=1}^{K}m_i\lg\bar{x}_i=\sum_{i=1}^{K}w_i\lg\bar{x}_i \qquad (1\text{-}11)$$