地质工作中常用的数理统计方法

甘肃省二〇七工程指挥部数理统计学习班

前　言

伟大领袖毛主席教导我们，从事一切工作都要**胸中有“数”。这是说，对情况和问题一定要注意到它们的数量方面，要有基本的数量的分析。**要多、快、好、省地进行地质工作，也应做到胸中有“数”。在地质工作中，经常要通过各种手段收集大量的地质资料和数据。对这大量的数据如果能尽快地、很好地加以整理、分析，就有可能从中提取出较多的有用信息，从而有助于地质工作者做出正确的地质判断、摸索各种地质规律，发现找矿线索等。而在地质分析的基础上，应用数理统计方法，结合使用电子计算机技术，就可以帮助人们对大量地质数据进行快速处理，和较为科学、有效地进行分析。最近十几年来，随着地质学定量化的发展和电子计算机在地质学中的推广应用而产生了一门新的边缘学科——数学地质，而数理统计则是数学地质的重要数学基础和方法之一。

无产阶级文化大革命以来，特别是伟大的批林批孔运动开展以来，我国地质战线和其他战线一样，也出现了一派大好形势，对地质工作定量化方面提出了新的更高的要求。广大地质工作者迫切希望有一本结合地质、通俗易懂的数理统计方法书籍。近几年，我教研室曾先后在甘肃省二〇七工程指挥部等单位举办的短期学习班上讲过数理统计。本讲义编写过程中，主要参考了湖北地质学院地球化学找矿教研室1972年10月编的《数理统计基本知识》和《地质与勘探》杂志1973年第2期至第8期上连载的地文同志写的“数理统计在地球化学探矿中的应用简介（一）——（七）”，并充实了放射性找矿中数理统计应用的具体经验和实例。

本讲义把重点放在数理统计方法在地质中的具体应用上，对概率论的内容以及这些方法的理论根据提得极少；尽量选用地质领域中的例题，并给出具体的计算与分析步骤；强调简算法，尽量使计算表格化，还介绍一些更简便易行的方法（如非参数方法和概率格纸的运用）；对概念的实际意义和用各种方法解决问题的基本思想尽可能给予通俗的说明，但一般都略去了理论证明和公式的推导（有一些较简单或重要的证明、推导放在（＊）或〔附注〕内），想了解这方面内容的读者可查阅有关参考书。

前六章是数理统计基本知识部分，为了便于读者加深理解和巩固所学内容，在各章后附有复习思考题和习题。第七章趋势面分析和第八章判别分析，虽然也是数学地质的一部分，但我们侧重讲其数理统计方法和可以不用电子计算机就能解决的较简单的例子。书后附录表中给出一些数理统计常用表，以便读者查用。最后，列出主要参考书和其他参考书目录。

本书是普及性的，主要对象是广大地质工作人员（包括找矿员、技术员等）。凡具有高中数学程度的读者就可以较容易地阅读。

由于我们水平很低，又缺乏实践经验，讲义中一定有许多错误和缺点，欢迎广大读者批评指正。

武汉地质学院数学教研室编

1974. 12

目　　录

第一章　数理统计概述

恩格斯说："**纯数学是以现实世界的空间的形式和数量的关系——这是非常现实的资料——为对象的。**"概率论和数理统计也不例外，它们都是从数量关系上研究大量同类现象中的统计规律性的数学分支。

为要真正了解概率论和数理统计的研究对象，还须首先了解偶然性和必然性以及统计规律性的概念。

§1.1　客观世界中的偶然性与必然性

毛主席教导我们："**事物的矛盾法则，即对立统一的法则，是唯物辩证法的最根本的法则。**"客观世界中各种事物或现象，都是充满着矛盾，并在对立面的斗争和统一中发展的。其中有一对矛盾就是偶然性和必然性的矛盾。客观世界中各种现象都是既有偶然性方面，又有必然性方面，是这二者的对立统一物。例如，在地质工作中，任何一批实测数据都既包含有偶然因素的作用，又包含着带有倾向性的系统变化。于是，初看起来这些数据似乎是没有规律的、偶然的，但经过整理、列表、制图后，往往就能初步看出数据中有一定的规律性。再看在其它领域中，这种结论是否也正确呢？比如说，在某工厂中生产一定长短的螺丝。一般认为所生产出的螺丝长短是固定的，是必然性的了。但是，只要经过仔细测量，还会发现每个螺丝的长短是不一样的。有的比标准长些，有的比标准短些，当然也有的与标准一样长，这就是偶然性了。造成这种偶然性的因素很多，有机床、材料、加工方法、测量仪器等因素，它们的影响是不能完全消除掉的。即使在精密仪器(如钟、表)零件的生产中，也还是避免不了这种偶然性的。因此，人们才要对产品的规格进行检查，并规定各种产品的公差范围。另一方面，有些现象看起来偶然性很大的，也都存在着其必然性的方面。例如，气象方面变化万千，但人们通过长期研究也逐渐把握住其中的规律性，从而做出了长、中、短期的天气预报；又如那年会发生洪水或干旱，也是偶然性很大的，但人们通过长期观测、研究，也可掌握其中某些规律，知道多大的洪水是200年不遇的(即平均来说，200年才遇到一次)，多大的干旱是100年不遇的(即平均来说，100年才遇到一次)。

事物的偶然性和事物的必然性一样，都是有原因的，只不过必然性的原因往往是少数几个起主要决定作用的因素，而偶然性的原因则往往是为数较多的、起着次要和辅助作用而又交互影响的那些因素而已。如考察炮弹射击现象时，炮弹的初速度、发射角和炮弹的弹道系数等，就基本上决定了炮弹的理论弹道。但是，每次实际射击时，炮弹命中点对目标总有或大或小的偏离，这是其偶然性的一面。这偶然性也是有原因的，如风力的影响、弹壳制造上的误差、弹药重量的偏差、弹药成分的不均匀性、炮身位置的偏

差、炮筒的磨损程度等等，故偶然性中也是有因果联系的。所以，我们说：世界上根本没有“无因果性的随机现象”，也没有“无偶然性的因果现象”。许多资产阶级学者，把概率论看成是研究无因果性现象（即随机现象）的数学理论，硬把客观现象截然划分为确定性现象和随机现象两大类，是违背辩证唯物论的。恩格斯在《自然辩证法》中早就批判过这种把偶然性和必然性对立起来的形而上学倾向，他说：**“偶然的东西是必然的，而必然的东西又是偶然的”**。并且指出，这种倾向必然导致唯心主义和不可知论。

客观现象中的偶然性和必然性不仅同时存在、互相依存，必然性往往通过偶然性表现出来，偶然性中就包含有必然性，而且二者还总是在不断斗争着的，在一定的条件下，二者还会互相转化。例如，在一定温度下考察一个容器内气体对器壁的压力时，发现器壁各处所受的压力相等，这是必然性。但是，当我们把气体从容器中抽出，使容器内接近真空时，就会看到这些少量气体分子对器壁各处产生的压力出现了偶然性的波动，这就转化为偶然性了。如果再把气体逐渐回到容器中，则压力又趋向于稳定状态，又转化为必然性了。在这里条件是很重要的，这条件就是容器中气体分子的数量。因为单个气体分子的运动方向、速度偶然性很大，但分子一多，运动时分子间互相碰撞、互相抵消的机会就增多了，于是对器壁碰撞产生的压力就变成一种集体的平均效果，从而呈现稳定状态。又如在地质工作中，若在大范围内考察某一地质变量，可能看出有增高（或降低）的趋势，但在小范围内观察时，这种变量的偶然性就突出了，这里的条件就是考察的范围大小。

当客观现象中的必然性成为矛盾的主要方面时，就可用微分方程等数学工具单个地找出各个变量间的函数关系，从而掌握这种现象中的规律性。如在引力作用下的行星运动，由于行星间的距离比行星本身大小大得多，使得我们几乎永远可以把行星理想化为质点；又由于行星所经历的空间中气体极为稀薄，使阻力小到几乎为零；……；于是，在行星运动中必然性就成为矛盾的主要方面，用微分方程求出的行星运行轨道与实际观察结果很好地相符，偶然因素的影响很小。但是，当偶然性成为客观现象的矛盾的主要方面时，就不能也不必要用微分方程等工具一个个地找出各因素间的函数关系，而要用概率论和数理统计这种数学工具，来研究大量同类现象的这种偶然性侧面中的统计规律性（或概率规律性）了。

§1.2 统计规律性

我们先看两个例子。掷一硬币，若单看掷一次的结果时，则究竟出现正面还是出现反面，不能准确预言，偶然性较大。但是，如果多次重复掷下去，并记下每次结果，那么就会看到有一种明显的规律性：平均地说，掷的次数越多，则出现正面和出现反面的次数就越接近于相等，或说出现正面的频数(次数)与总频数之比越接近于1/2。又例如观测一定数量放射性物质的原子核蜕变（或衰变）。若单看某一个原子核时，它究竟何时发生蜕变，偶然性较大。但若作为集体来观察全部原子核（数量极大）的蜕变时，就会发现明显的规律性：其半衰期（放射性物质衰变掉原有数量的一半所经历的时间）对该种物质来说，是一个常数。于是，得出放射性物质的衰变规律：$N=N_0e^{-\lambda t}$，（其中，N_0为放射性元素在开始时原子核的总数，λ为该放射性元素的衰变常数，N为放射性元素经

过时间 t 后原子核的数量）。

上述两个例子中的规律性，都是统计规律性。它与人们较为熟悉的函数规律性相比，有两个不同的特点：

①在一次试验，或观察单个个体时，偶然性较大，而当多次重复试验，或观察集体现象时（也可说是观察大量同类现象时，或说是大量统计时）才表现出来的规律性，是统计规律性；

②统计规律性给出的规律性结论是统计平均性的，不象函数性规律那样只要每次给定了自变量X，则函数y就被完全确定了。如放射性元素的衰变律，就是一种统计性规律。它的平均衰变率D次/分虽是一个常数，但是在某个具体时间的衰变率并非正好是D，而是有出入的，这种出入，我们称为统计性涨落。

尽管统计规律性有它的特点，而且当人们开始接触它时可能很不习惯，然而它确实是一种不依人的意志为转移的客观规律性，则是不容怀疑的，正如恩格斯所指出的那样：**“凡表面上看去是偶然性起作用的地方，其实这种偶然性本身始终是服从于内部隐蔽着的规律的。”**

§1.3 概率论与数理统计

概率论与数理统计是研究大量同类现象中统计规律性的数学分支，它们从数量方面研究大量同类现象中所特有的一种矛盾——某一事件在大量重复试验（或观察集体现象）中出现的偶然性与必然性。概率论与数理统计是密切不可分的，但也稍有区别。概率论侧重于从建立数学模型和进行数学理论推导上来研究大量同类现象的客观基本规律，而数理统计则在概率论理论的基础上着重于从整理各种观测数据中进行分析、归纳和研究。但它们所得出的结论都是要通过客观实践的检验来决定弃取的。

自从十七世纪中叶，在阶级斗争、生产斗争和科学实验这三项社会实践的基础上产生了概率论以后，经过几个世纪的发展，概率论和数理统计已逐渐在农业、工业、医学、射击学、生物学、气象学、水文学、地质学等部门以及各种尖端技术方面获得了越来越广泛的应用；产生了许多概率论的新分支，如随机过程论、信息论、对策论、排队论等。数理统计在很早以前就是概率论发展的源泉，也是它的应用对象。数理统计当它和具体的研究对象结合时，就形成种种不同的统计学：如纺织工业统计学、农业和生物学中的统计方法、森林统计学、天气预报中的统计方法、水文学中的数理统计方法等。今天，概率论和数理统计已经成为人们认识世界和改造世界的一种有力的数学工具，成为应用数学的重要部门之一。特别是1958年大跃进以来，数理统计在我国工农业生产各部门中得到广泛应用。在我国地质工作各方面的应用也日益增多。当前，我国地质勘探工作已发展到深部找矿阶段，迫切需要更加科学地、迅速地处理大量数据；另一方面，在无产阶级文化大革命、批林整风和批林批孔运动的推动下，我国电子计算机技术迅速发展，又为更复杂的概率论、数理统计方法的应用提供了新的极大的可能性；因此，完全可以预料：在不久的将来，数理统计在我国地质工作中所起的作用会越来越大；同时，数理统计本身在结合地质的过程中也将得到新的发展。

§1.4 数理统计的基本概念——母体（总体）、个体（样本点）与子样（样本）

例如，我们要对某地区进行分散流找矿。从该地区水系沉积物中取了5000个样品，分析了其中的铜含量，得到5000个数据。我们把在一次统计分析中所要研究的对象单元全体（在这例子中，就是该地区水系沉积物中所有可能的铜含量数据的全体）叫做“母体”（或“总体”）；把组成母体的每个对象单元（在这例子中，就是该地区水系沉积物中每一个铜含量数据）叫做“个体”（或“样本点”）；把从母体中取出的一部分个体的集合（在这例子中，就是从所取出的5000个样品中分析出的5000个铜含量数据）叫做“子样”（或“样本”）。子样中包含个体的数目（在这例子中为5000），称为子样的“大小”（或“容量”）。母体中包含有限（无限）个个体时，称为有限（无限）母体。

如果对每个样品还分析了其中的银含量，于是又得到另外5000个数据，它是银含量的子样，相应地也有银含量的母体与个体。因此，地质中常说的“样品”（或“标本”）与“子样”（或“样本”）是不同的，前者是一个地质实体，后者是n个数据。“样品”与“个体”也有区别：首先，“个体”是指从“样品”测得的某方面特性的数值；其次，如果对一个“样品”分析了铜和银两种金属的含量，则一个“样品”就对应有两个不同类的“个体”了。

统计母体必须是由那些从内部来说有一定内在联系（即在某种意义上是同类的），而从外部来说又有差异的诸个体所组成。例如上述地区内的水系沉积物，它们形成的地质条件大体上是相同的，故可作为一个母体。一般来说，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母体来统计，因为它们在地质上是不同类的。只有当不同岩性对我们所考察的某一地质变量不起显著影响时，才可以把这不同岩性的地区视为同一母体。如果上述5000个分散流样品的铜含量包含一些异常值，就可认为这子样来自复合母体（即若干个母体混合在一起）。区分复合母体，圈定异常区，是化探和放射性勘探中经常遇到的统计问题之一。

§1.5 数理统计的内容和工作步骤

数理统计的中心课题，就是根据子样（一批实际观测数据）来分析、推断母体（某一特定的地质研究对象）的各种特性、特征。

数理统计的内容是很广泛的，而且不断地有新的内容补充进来，但常遇到的基本内容大致包含以下五个方面：

①找寻描述一个母体某种特性的综合指标（统计参数），以及此种特性指标的概率分布规律（包括统计特征数、概率分布、参数估计、适线理论等）；

②判定二母体间有无显著差异，或一特征数前后有无显著变化（统计假设检验）；

③分析使母体发生显著变化的因素（方差分析）；

④分析一母体的各种特征之间的相互关系，或环境对该母体各种特征的影响（相关分析与回归分析）；

⑤研究科学的抽样（取样）与试验的方法（抽样理论、试验方法）。

此外，还有工业产品质量控制、非参数性统计推断、序贯分析、实验设计、统计决策函数等。由于电子数字计算机的采用，使多元统计分析方法能实际地应用于地质工作中来，因此，近一二十年又发展起一门叫做“数学地质”的边缘学科。本材料中，除了介绍数理统计的基本方法外，还简单介绍“数学地质”中的趋势面分析和判别分析部分。

数理统计的工作步骤，大体可分为三步：

①收集原始数据。收集时，要了解较详细的地质背景；并注意数据的可靠性（包括观测条件要一致）；此外，还要保证数据具有一定的数量。如果没有一定的数量，统计结果就会没有多大的实际意义。当然，也不是说数据越多越好，要根据问题的具体情况来决定取样观测的合理数量。如果原始数据的收集方案是按照一定的理论（抽样理论）取得的，那么统计工作就会更有目的性，更有效果了。

②数据整理。将所得数据进行统计处理，包括分组、统计、列表、制图，以及按照一定公式计算一些特征值等。计算时，要注意公式的应用条件，并尽量采用简算法和计算工具及表格等，还要设法经常核对，避免最后出错时再全部返工。

③推断和解释。根据计算结果，进行统计推断和预测，并结合具体地质条件作出合理的地质解释，再回到实践中加以检验。

§1.6 在地质工作中应用数理统计的一些问题

1. 在地质工作中应用数理统计的必要性

马克思曾指出：“**科学仅当它成功地利用数学时，才达到完善程度。**”任何一门自然科学的发展，当其由局部资料积累进入大量资料积累，由宏观观察进入微观研究，由定性描述进入定量分析，由观察推断进入模拟实验的情况下，总是要求借助数学这一强有力的工具的。地质科学现在则正处于这种情况之中。

一方面，由于地质现象是作为经过漫长的地质年代，多种多样的地质作用叠加在一起的结果而出现的，其中既有区域性因素和局部性因素的影响，也有许多偶然性因素的影响；另一方面，又由于地质对象不可能全部取来进行研究，只能通过抽取子样来研究母体；这两个特点正好与数理统计的特点相符合。因此，在地质工作所借助的各种数学工具中，数理统计就占有很重要的位置。

在地质工作中应用数理统计，可以充分发掘和利用已有数据中所包含的各种有用信息；可以通过计算一些统计参数（特征数），为做出地质判断、分类、解释、评价等提供定量的依据；还可以帮助探索元素分布和矿床成因等地质规律；在此基础上，帮助指导找矿和储量计算等，以取得地质勘探上多、快、好、省的效果。

2. 在地质工作中应用数理统计的可能性

数理统计在地质勘探各个领域中都有广泛应用的可能性。例如，就我们目前所知，在放射性测量工作中，就至少有以下几方面的可能应用：

①利用正态分布和对数正态分布曲线求出底数（背景值）和标准差（均方差），以确定偏高值、高值和异常值下限，从而作出偏高值、高值和异常值图；还可用来统计不同岩性的γ强度分布，以确定地质界线，辅助地质填图；

②利用相关分析找出各元素含量间的相关关系，以便间接找矿，或计算伴生元素储量；还可用来分析在元素含量的物理分析和化学分析间是否存在系统误差，以及从化探全分析数据中研究各种元素间的共生组合关系；

③利用趋势面分析对测区内的大量数据进行筛选，来评价远景地段及找寻打钻位置；

④利用统计假设检验判断两个地段可否算做一个母体；

⑤利用统计假设检验来检验各元素是否服从正态分布或对数正态分布；

⑥利用判别分析判别含矿异常和不含矿异常，等等。

3．在地质工作中应用数理统计的条件

数理统计在地质工作中有广泛而有效的应用这点已为越来越多的事实所证明；另一方面，也决不应夸大数理统计的作用，把它说成是万能的。实际上，数理统计只是起着从数量方面协助地质工作的作用。因此，决不应脱离开地质对象的本质和其自身所固有的规律性来孤立地应用数理统计方法。相反地，只有在密切结合地质理论和实际、充分利用各种地质资料的条件下，数理统计的应用才可取得有效的成果。不仅在数据的收集、整理、计算时，必须考虑地质背景，而且对统计计算的结果也需要尽可能地说明其地质意义。这就要求地质工作者和数学工作者紧密结合，共同协作。

其次，数理统计方法的基础是建筑在观测数据的质量上的。观测数量够不够多、子样的代表性强不强、观测或分析的精度够不够要求、不同母体的数据是否分开了等等，都影响到数据的质量。因此，必须重视研究抽样方法、实验设计和实验室误差等基础工作，不断提高观测数据的质量，才可使统计推断建筑在可靠的基础上。

第一章复习思考题

1．什么是统计规律性？它有何特点？试举出一两个服从统计规律性的实例。

2．地质中常说的“样品”与数理统计中的“子样”、“个体”有何联系和区别？

*　　*　　*

本章主要参考书为参考书目录中的〔6〕、〔1〕、〔4〕、〔7〕等。

第二章 数据整理

在这章中，介绍对收集到的地质数据如何进行统计整理的一些常用方法，并为下一章讲随机变量及其概率分布打下感性认识的基础。

在讲具体方法之前，有必要先分清连续变量和离散变量这两个概念。只能取间断数值的变量叫离散变量。例如，重砂样品中锡石的粒数，岩石样品中古生物化石的个数等，它们只能取0或正整数，故为离散变量。按其本质来说，可以取一个或几个区间中，或整个数轴上一切数值的变量，叫连续变量。例如，化探样品中的锡含量、铜含量，以及放射性测量中的γ辐射强度等，则是连续变量。虽然分析报告中的数字是离散的，但那是由于受分析精度限制之故，该变量本质上还是连续变化的。离散变量和连续变量在数据的整理方法上是有差别的。由于在物化探及放射性测量中主要碰到的是连续变量数据，故下面只重点介绍连续变量地质数据的整理方法。

§2.1 分组、列表、制图

对地质数据进行统计整理，最常用的一套方法，就是分组、统计、列表、制图。其目的是为了突出地质数据中的统计规律性，或具体地说，就是突出其中频率(频数/总频数）分布的稳定性。例如，未经整理的铜含量数据，乍一看去，比较杂乱，但如果统计了落于各区间内的铜含量数据个数占全部数据个数的百分比，就可看出其频率分布的某种稳定性来。以下我们结合安徽月山闪长岩体中Cu含量的比色分析结果之例，具体说明分组、列表、制图的方法。原始数据及其常用对数值列于表2—1中。

1. 分组

将实测数据划分成几组是没有一成不变之规，要根据数据的性质、数据的多少、数据变化范围的大小、计算的精度要求和测量（或分析）质量的高低适当划分之。一般的经验是，分组数不宜少于5组，但也不必多于15（或20）组；每组平均至少要能分摊到5个以上的数据，譬如有70个数据时，至多分成14个组。分组过少或过多都不好。分组过少，则偏差太大，难以显示数据中的规律性；分组过多，不仅工作量很大，而且可能出现有些组中没有数据的现象，反而不能清晰地反映频率分布的规律性。

一组中的最小值叫做组下限，一组中的最大值叫做组上限；分组点既是前一个组的组上限，又是后一个组的组下限；组上限与组下限之差叫组距（即组的间隔长度）。分组有等组距分组和不等组距分组两种，但一般都采用等组距分组，这样便于计算。等组距分组时，组距记为l。各组间隔（或组段）的中点的数值（或说组上限与组下限之和

表 2－1　月山闪长岩中Cu含量数据及其常用对数值

样品号	Cu含量(γ/g)	对数值	样品号	Cu含量(γ/g)	对数值	样品号	Cu含量(γ/g)	对数值
1	28	1.45	21	14	1.15	41	22	1.34
2	20	1.30	22	10	1.00	42	45	1.65
3	4	0.60	23	48	1.68	43	13	1.11
4	20	1.30	24	40	1.60	44	13	1.11
5	16	1.20	25	8	0.90	45	5	0.70
6	32	1.51	26	96	1.98	46	15	1.18
7	10	1.00	27	10	1.00	47	18	1.26
8	20	1.30	28	20	1.30	48	11	1.04
9	16	1.20	29	6	0.78	49	17	1.23
10	20	1.30	30	20	1.30	50	20	1.30
11	8	0.90	31	32	1.51	51	13	1.11
12	10	1.00	32	10	1.00	52	14	1.15
13	12	1.08	33	20	1.30	53	8	0.90
14	10	1.00	34	11	1.04	54	13	1.11
15	16	1.20	35	28	1.45	55	24	1.38
16	6	0.78	36	8	0.90	56	25	1.40
17	16	1.20	37	13	1.11	57	8	0.90
18	12	1.08	38	13	1.11	58	18	1.26
19	16	1.20	39	10	1.00	59	6	0.78
20	16	1.20	40	20	1.30	60	8	0.90

的一半）叫组中值，我们正是用组中值来代替落于该组间隔内的各个数据，从而初步简化原始数据的。

对于大多数微量元素（包括放射性元素）来说，其样品中元素含量的变化常达好几个级次，如果按等差的方法来划分间隔，则作图结果将产生极大的不对称性。为使作出图形有对称性，通常在分组前，先将该元素含量数值取常用对数，然后再把其对数值按等差的方法来划分间隔（实际上，就是把含量数值本身等比地划分间隔）。如果直接在对数格纸上作图，则可省去查对数表这一工序〔注〕。对于半定量分析结果，最好把含量等级取作组中值，以免统计中某些间隔的频数为零。

实际分组步骤为：

①确定数据的上界和下界（上界可比最大值稍大一点，下界可比最小值稍小一点。

〔注〕：以后要讲到，取对数的根本原因在于：这些元素含量值服从对数正态分布。

因是连续变量，故可以允许这样做）；

②适当确定分组数 n 和组距 l。如果先确定了n，则组距 $l=\frac{上界-下界}{n}$；如果先确定了组距l，则分组数 $n=\frac{上界-下界}{l}$；

③决定分组点；

④统计出频数。

例如，月山闪长岩中铜含量的数据，因为铜是微量元素，其含量变化达好几个级次，故须先取常用对数值，然后对其对数值进行等组距分组。其对数值中最小者为0.60，最大者为1.98。可稍向外延拓，取下界为0.47，上界为2.07，于是上界－下界＝2.07－0.47＝1.60。数据总个数为60，考虑分 8 组较好，于是 n＝8，组距 $l=\frac{1.60}{8}=0.20$。这样，分组点就是0.67，0.87，1.07，1.27，1.47，1.67，1.87。在统计频数时，如果有的数据恰好落在分组点上，那么究竟它应划归哪个组呢？解决这个问题有三种办法：

①统一规定：凡落在分组点上的数据，都划归分组点左面（或右面）的那个组；

②取分组点比原数据的测量精度高一位（如上例中，取分组点为0.675，0.875，……），就可避免出现上述麻烦（但这样做会给计算上带来不便）；

③另行分组，改变上界、下界和分组数n，使分组点避开原数据（这样做，有时要反复分几次组，比较费时间）。

根据一些同志的实践经验，认为在分组前先作出原始数据的散点图，再用直尺（有标度的）在该图下方沿横轴左、右移动，寻找最合适的上界、下界及确定合适的分组点，是有方便之处的。对表 2－1 中的对数值，用方格纸作出散点图如下（图 2－1）：

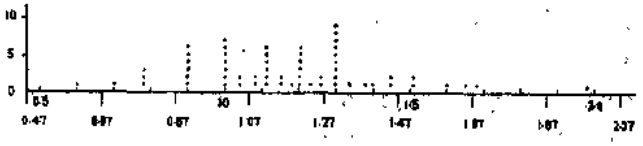

图 2－1　月山闪长岩中Cu含量对数值的散点图

在散点图上左右移动直尺确定上、下界及分组点的方便之处在于：①比较直观，对一种分组方案的优缺点一目了然；②调整好后，可以一次分组成功，避免反复分组，节省了时间；③便于比较各种分组方案，全面考虑各方面要求，选择比较最优的分组方案；如考虑组距时，首先必须使组距大于观测精度（如放射性测量中不应低于仪器的灵敏阈），否则，可能使某些测量值不能正确落入相应的间隔内；其次，二倍组距长最好大于相离最远的两个数据的距离，以避免出现有的组中不含数据的现象；第三，组距的大小要以能突出数据比较集中处的频数为准。此外，在确定分组点时，要尽量避免与数据相重叠。

在综合考虑了上述要求后，在散点图上确定的下界为0.47，上界为2.07，分组点为

0.67，0.87，1.07，1.27，1.47，1.67，1.87。然后，再统计频数。

2. 列表

主要是列频数分布表、频率分布表和累积频率分布表，有时这三者在一个表中给出。对上例我们可列出其频数、频率及累积频率分布表如下（表2－2）：

表2－2　月山闪长岩中Cu含量对数值的频数、频率和累积频率分布表

对数间隔	组中值X_i	统计计数	频数f_i^*	频率f_i(%)	累积频率F_i(%)
0.47—0.67	0.57	一	1	1.7	1.7
0.67—0.87	0.77	正	4	6.7	8.4
0.87—1.07	0.97	正正正	15	25.0	33.4
1.07—1.27	1.17	正正正正	20	33.3	66.7
1.27—1.47	1.37	正正正	14	23.3	90.0
1.47—1.67	1.57	正	4	6.7	96.7
1.67—1.87	1.77	一	1	1.7	98.4
1.87—2.07	1.97	一	1	1.6	100.0

表中第一列，是用下界、上界和分组点构成8个对数间隔；第二列是各组的组中值，记为X_i（i=1,2，…，8）；第三、四列是根据表2－1中的对数值，用唱票的办法统计落入各间隔中的数据个数(即频数)，记为f_i^*（i=1,2,…,8）；第五列是把各组的频数分别被总频数（即数据总个数N，此例中N=60）除，得到频率值f_i(%)（i=1,2,…,8）（有的书中称“频率”为“相对频数”）；第六列是把第五列的频率值依次累加起来得到的累积频率值F_i(%)(i=1,2,…,8)，其方法是：以第一组频率f_1就作为第一组上限处的累积频率F_1，再把F_1加第二组频率f_2作为第二组上限处的累积频率F_2，再把F_2加上第三组频率f_3作为第三组上限处的累积频率F_3，……余类推，最后第八组上限（即上界）处的累积频率F_8必为100(%)，否则就说明计算中有错误。

同样，为了检验列表中有无错误，也可利用频数和频率的基本性质：

①设分组数为n，则n个频数之和应等于总频数N，即$\sum_{i=1}^{n} f_i^* = N$（“Σ”号是“总和”之意，读为“西格马”$\sum_{i=1}^{n} f_i^*$就是$f_1^*+f_2^*+\cdots\cdots+f_n^*$的简写，读作$f_i^*$当i从1到n时的总和）。在上例中，n=8，N=60，故应有$\sum_{i=1}^{8} f_i^* = 60$，如此式不满足，便说明统计中有错误。

② n个频率之和应等于1，即 $\sum_{i=1}^{n} f_i = 1$（或100%）。这是因为频率 $= \frac{\text{频数}}{\text{总频数}}$，

即 $$f_i = \frac{f_i^*}{N},$$

所以 $$\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} \frac{f_i^*}{N} = \frac{\sum_{i=1}^{n} f_i^*}{N} = \frac{N}{N} = 1$$

这个式子不但可以用来检验统计计算有无错误，同时它还包含一个深刻的理论含义，即任何一个统计整体（包括子样或母体）的总频率恒等于1。

3. 制图

主要是制频数（或频率）分布直方图和累积频率多角形图。这两种图在制法和特点上有些不同，下面分别介绍之：

①制频数（频率）分布直方图：

以横轴为X轴，以横座标表示变量X的值，在X轴上标出上、下界及各分组点；预先选定一个以组距1为底边的小长方形作为单位面积；然后，以各间隔的线段为底边，在其上分别立一个面积等于该组频数（频率）值的长方形，这样作出的图形就叫频数（频率）分布直方图。显然，在同一单位面积下频数分布直方图比频率分布直方图高N倍。因此，如果以频数分布直方图的单位面积作为频率分布直方图单位面积的1/N的话，则此二直方图将重合在一起。根据表2—1数据作出的频数（频率）分布直方图见图2—2。

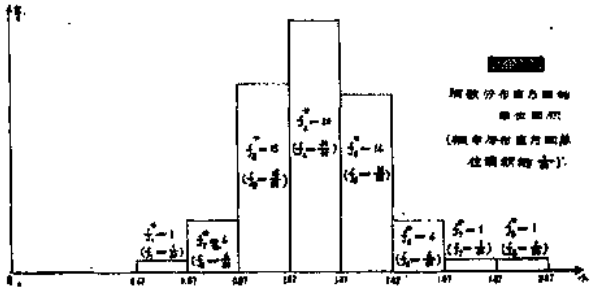

图2—2 月山闪长岩中Cu含量对数值的
频数（频率）分布直方图

直方图能比较客观而明显地反映出统计对象的频数(频率)分布特征，如从图2—2

中容易看出X值落在0.87—1.47内的频数（频率）很大，往左、往右都较小，而且图形基本上是对称的。

这种直方图的特点，是用长方形面积大小，而不是用线段高度，来表示频数（频率）的。由于总频率恒等于1，故在频率分布直方图中，各长方形面积总和等于1（即单位面积）。那么在频率分布直方图中，长方形的高度（以纵座标y表示）具有什么意义呢？因为

$$\text{长方形高度}y=\frac{\text{长方形面积}}{\text{长方形底边长}}=\frac{\text{频率}}{\text{组距}}$$

表示在X轴单位长度上平均分布了多大的频率，故Y的意义可以说是频率分布密度。

在实际工作中，往往把直方图中相邻长方形顶边的中点两两连结起来，构成频数（频率）分布多角形图。这种图与直方图作用基本上一样，只是在表达分布的形状上比直方图更直观些，虽然也能保持了总频率等于1的性质，但在各组间隔上方的多角形面积则不一定等于该组的频数（频率）。

②制累积频率多角形图

仍以横轴为X轴，而以纵座标表示累积频率F(%)的值，在X轴上标出上、下界及各分组点，再在各组组上限（上界为最后一组的组上限）处立一高为对应的累积频率值F(%)的虚线段，依次连结各虚线段的顶点（下界点与第一组组上限处虚线段顶点连结），就构成了累积频率多角形图，见图2—3。

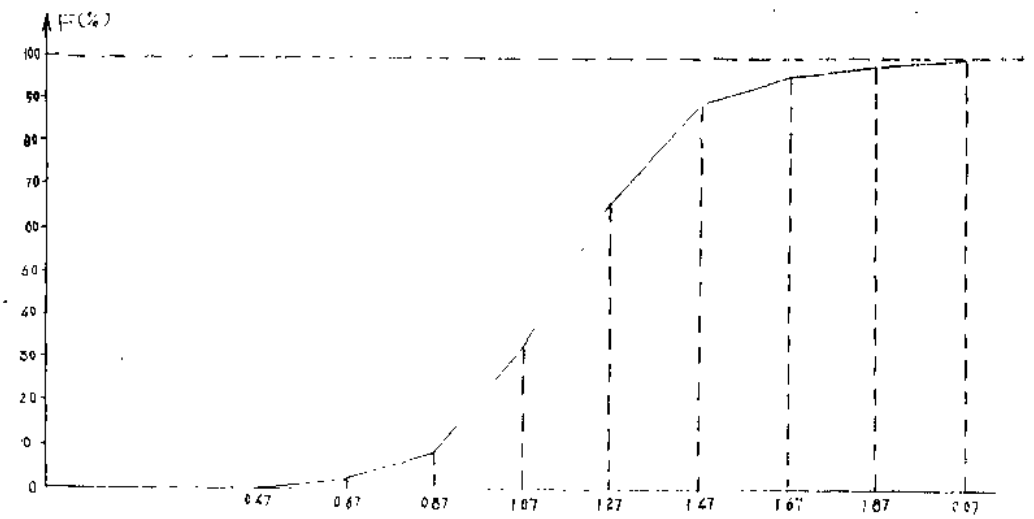

图2—3　月山闪长岩中Cu含量对数值的累积频率多角形图