观测数据的数学处理

林纪曾 编

地震出版社

1981

前　言

数据是人们用以了解和研究客观事物的性质、运动规律，以及事物间相互关系的最基本素材，科学技术和生产实践中的任何领域都离不开它。但是，由于客观世界的复杂性，观测工具的不够准确和观测者感官的不够灵敏等等原因，往往使通过观测得到的数据不能直接而有效地加以利用。为了更好地发挥数据的效能，例如正确估计所测结果的可靠程度，在数据中提取更多需要的信息，对结果提出更为合理和准确的解释等，就要对数据进行一系列数学上的加工处理，这叫做观测数据的数学处理，简称数据处理。

可以这样说，对于原始的观测数据，如果不作任何数学处理的话，几乎是无法应用的。因此，一个从事观测或实验的科技工作者，如果不具有基本的数据处理知识，就不仅不能充分利用所得到的观测资料，还可能被复杂多变的数据所迷惑而得出非常错误的结论，那么也就无法胜任自己的工作。

本书主要是根据地震预报学科的需要，为地震预报的观测分析人员编写的。近年来，随着地震预报事业的迅速发展，吸收了大量的新生力量，他们中有的从其他部门转移过来，有的则是刚刚参加工作的年青人。他们对地震预报事业抱有极高的热情，埋头工作于第一线，整天和观测数据打交道，但往往由于缺乏数据处理的基本知识，而使工作的进一步开展遇到严重的障碍。尽快地向这些同志普及数据处理的知识，提高观测分析水平，已成了提高整个地震预报水平的当务之急。笔者希望所编写的这本书能够在这方面起到一些有益的作用。

本书涉及的内容较为广泛，从观测数据的基本性质及需要用

到的数学基础知识开始，进而阐述误差理论的基本内容，然后叙述数据处理中最基本和最常用的一些方法，这些都是作为一个观测分析人员必须具备的知识。而且这些内容在其他的科学观测和实验中也基本适用，因而也可供在其他领域进行观测或实验的科技人员参考。

考虑到本书的主要对象是刚开始从事科技工作或缺乏这方面基础知识的同志，因而在写法上不重视严格和较高深的数学推导，而着重于通俗地阐明原理和应用途径，以使读者易于理解和使用。为此，往往不厌其烦地在文字上多作深入浅出的叙述，因而这也是一本有关数据处理的入门书。书中每章后面附有少量习题，供读者通过练习来加深内容的理解和掌握，并附有习题解答以资核对。书后所附的一些数表既是本书内容的需要，也是在一般数据处理中经常用到的。

由于笔者水平有限，经验不足，书中的缺点和错误在所难免，亟望读者批评指正。

林　纪　曾

1979 年 9 月，广州

目　　录

第一章 绪 论

§1. 数据的作用

对客观事物的描述，可以分为两种方法，一种是定性的描述方法，它采用普通的语言或类似的方式来进行描绘；另一种是定量的方法，它采用数字来进行描绘。所谓数据，就是指人们用来描述客观事物的定量的表示。例如：说“地球很大”、“风刮得很猛”，这是定性的描述；说“地球的平均半径是6371公里”、“风力是12级”，这是定量的描述。显然，一旦引入了“量”这个概念，就使人们的认识大为改观，变得清晰和深刻了。

事实上，人类对自然界的认识总是由浅入深的，这种由浅入深的变化，经常是在人们对客观事物的认识由“定性”过渡为“定量”时，表现得最为明显。这是因为定性的东西往往只能给人以一种笼统的、模糊的印象；而定量的东西却能使人有一个明确的、具体的概念。

不仅如此，在科学技术领域，数据对于利用自然、改造自然，预测客观事物的运动变化，更是不可缺少。例如：地质学家只有掌握了某一矿床的储量、品位、埋藏深度等等数据时，才能正确判断该矿床是否有开采的价值；天文学家只有掌握了地球、月亮运动的各种数据时才能准确预测日食、月食等天文现象的出现时间；这样的例子是不胜枚举的。可见，数据在人类认识客观世界的过程中，占据着极其重要的地位。

因此，在一切科学技术领域中，为了研究某个事物或某种现象，都采用了一个共同的研究方法，就是利用一定的方式方法和一定的仪器设备来进行观测或实验，从而得到我们所需要的资料，而这些资料几乎都是用数值来表示的。我们把最初所得到的数值

资料叫做原始数据。取得这些数据的过程叫做数据的测定。

§2. 数据的测定方法

利用仪器设备对某一事物进行观测或实验，在测量仪器中得到的读数，叫做观测数据，也叫测定值。

在现代的科学研究工作中，为了测定研究对象的某一个量，一般都经过三个步骤：仪器安装、观测和读数。为了尽可能取得正确和精密的结果，对其中任一步骤中任何细节都不能疏忽，否则，就会影响所观测的数据的质量。因此，观测者在仪器的安装、观测程序的设计、技术要求的制订和读数准确性方面都必须严肃、认真、细致地对待，这样才能取得尽可能满意的结果。

对于测定的方法，若从测定的对象、过程和所得结果的性质来看，可以分成如下两大类：

(1) 直接测定法：

把需要测定的某一未知物理量与已知的该物理量的标准量相比较，从而确定出未知量的大小，这叫做直接测定法，由于测定中一定包含了相对比较的过程，故也叫相对测定法。

例如：用一根米尺来测量一张桌子的长度，实际上是将桌子与作为标准量的米尺进行比较，看桌子的长度是米尺的多少倍，从而确定桌子长度的数值，若比较后得到桌子是米尺的 1.5 倍，则桌子长度是 1.5 米。又如：用天平测定某一重物的重量，砝码是标准量，根据天平平衡时的砝码数值即测得重物的重量。

(2) 间接测定法：

已知需要测定的未知量与另一些完全不同的量之间存在确定的函数关系，通过直接测定这些量，并利用已知的函数关系，来间接算得未知量的大小，这叫间接测定法。由于最后测定结果是由若干个与未知量完全不同的量根据确定的函数关系换算出来的，而在测定过程中并没有采用相对比较的方法，所以，人们又把这种测定方法称为绝对测定法。

例如：采用直接测定法测出圆柱体的半径 r 和高 h 后，由公式：

$$V=\pi r^2 h$$

计算出体积 V，故体积是间接测定的。

又如：在测震学中，利用体波测定近震震级 M_L 的经验公式是：

$$M_L=\lg\frac{1}{2}\left(\frac{y_E}{V_E}+\frac{y_N}{V_N}\right)+R(\Delta),$$

其中 y_E、y_N 分别为地震记录图中东西向和南北向的体波最大半振幅，即最大振幅处波峰和波谷的垂直距离的一半，可以用尺子直接测定；V_E、V_N 是相应的地震仪放大倍数，是已知常数，故 $\frac{y_E}{V_E}$、$\frac{y_N}{V_N}$ 是东西向和南北向的最大地动位移；$R(\Delta)$ 叫起算函数，与震中距 Δ 有关，已列成数值表；Δ 则与 P 波和 S 波的到时差 t_S-t_P 有关，由于时间的长短在记录图中表现为记录点移动的距离，故 t_S-t_P 也可以根据 S 波和 P 波到达点之间的距离由尺子直接测定出来，再除以走纸的速度而得到，它与 Δ 之间的函数关系也列成了数值表。于是，为了测定震级 M_L，首先要从地震纪录图中直接测定出 y_E 和 y_N，由此计算出公式右边的第一项；然后测定出 t_S-t_P，再通过两次查表——相当于两次运算，得到 $R(\Delta)$值，最后将上述二项合并得到 M_L 值。可见，M_L 是通过直接测定与 M_L 本身完全不相同的三个量 y_E、y_N、t_S-t_P，再经过比较复杂的运算后而得到的，故 M_L 是由比较复杂的间接测定法测定的。

许多物理量既可以用直接测定法来测定，也可用间接测定法测定。例如，电路中的电流值，可以通过在电路中串连电流表来直接测定。这时，电流表的刻度可看成是经过“标定”的标准量，整个电流表则是测量的装置。测量时，根据指针偏转的位置与作为标准量的刻度相比较，即得电流的数值，故属直接测定。另外

也可根据公式：

$$I=\frac{V}{R},$$

用其他方法测定出电压V和电阻R后计算出电流I值，这样就是间接测定法。

值得注意的是，在科学技术中的所有测定，不管它是直接的还是间接的，基本上都属于长度的测定，或者说，不论要测定的是什么物理量，最终总是转化为对长度的量度。日常中用尺子测量物体的长度不用说是属于长度的量度；天平测量重物要看指针是否在标尺上指零，电流表测量电流要看指针在标度上的偏转度数，也就是要读出指针偏转的距离，都是长度的量度；前面提到震级的测定中，最后也是转化为对图纸中最大半振幅和S波与P波到达点之间的距离的测定，显然也是长度的量度。即使在现代化的测量技术中，最后测定结果可以有直接的数字显示或图形显示，实际上所有测定仍然是转化为长度的量度，只不过配备了一整套自动读数、自动计算、自动显示的装置而已。

另外，若从测定方法的性质来说，则可分成下面三类方法：

(1) 偏位测定法：

即测量时根据测量仪器的零件所产生的偏转或位移来定出未知量的大小。如前所述电流表测定电流的大小，是根据电流表的指针偏转位置的大小来确定的。

(2) 零位测定法：

是用已知标准量去抵消未知量的作用，使仪器上的指示器指在平衡的零位，这时未知量即等于标准量。天平测定重量，电桥测定电阻，都是零位测定法。

(3) 逐次比较法：

它介于偏位法与零位法之间，必须通过逐次比较才能决定未知量的数值。例如，利用秤测定重物时，要逐次移动秤砣的位置，直至两边平衡时才能读出重物的重量。

§3. 观测数据的不确定性

任何测定过程都不可能得到与实际情况完全相符的测定值。在测量过程中，由于测量者的主观因素，测量仪器的精度限制和周围环境条件的影响等等，总会在测量中不可避免地引进这种或那种偏差。事实上，如果用同一种测量仪器重复多次测量同一种量时，每次测量结果不可能是完全相同的。举一个日常生活中的简单例子，用一把尺子来测量一张桌子的长度，重复测量若干次时，不管测量者如何细心，测量的结果不会每次都一样。一般说来，测量出的数值会大体一致，却又参差不齐。造成这种情况的原因很多。例如，在各次测量中，尺子刻线起点与桌子起始边界的相对位置不完全一致；人的眼睛、尺子刻度与桌子终点边界所构成的角度也会有些微小变化，在估读最后一位数时主观判断上亦会有一些差异；此外，尺子精度的限制，在不同时间里，不同的气温、气压、湿度等对桌子和尺子的长度的影响，以及其他种种未知因素的影响等等，都会使得每次测量的数值不相同。由于上述种种原因，使得任何一种测定中，对于多次测定所得到的数值结果，都不可避免地在一定范围内出现数值上的波动，这种现象叫做数据的不确定性。

尤其要指出的是，在实验室进行某一物理量的测定时，由于实验条件可加以控制，实验对象容易约束，往往能较好地控制数据的波动，特别是环境因素的影响所造成的波动。但是在直接对大自然进行观测的一些领域，如气象、水文、天文等，环境因素的影响往往是难于控制的，由此而带来的观测数据的波动就要大大加强，有时这种波动甚至超过观测量本身的许多倍。同时，由于具有影响力的因素多而复杂，数值波动情况也相应变得非常复杂，观测数据的不确定性就更加明显。

§4. 数据中的信息和噪音

在不大严格的情况下，可以把观测数据中由于测量仪器的精度限制、不稳定性和观测者的主观因素等造成的数据的波动成分称为观测误差，把周围环境条件的变化对仪器和观测对象的影响所造成的波动成分叫做干扰，并把它们总称为噪音。在地震预报的观测研究中，环境干扰经常比观测误差大得多。

在观测某一个量时，噪音总是伴随着观测过程而叠加进来。换句话说，不要认为我们所观测到的量就是我们要观测的量的真实数值，实际上，它是这个量本身以及噪音的混合物。这一点是绝不能模糊的。

在研究工作中经常碰到这样的情况，我们需要研究某一事物发展变化的规律，但由于某种原因，无法对这一事物本身进行直接的和有效的观测。可是，它往往与另一些现象存在有机的联系。这就为我们提供了一条间接研究的途径，即通过不断观测与其有联系的现象来研究该事物发展变化的规律。由此所得到的一系列观测数据，叫做数据的序列。在这个数据序列中，与所研究的事物存在着直接或间接关系的部分，称为信息，或叫做信号；另一部分就是噪音。例如，人们了解到一个大地震的孕育和发生的过程会引起震中附近的地壳变动，便采用倾斜仪来观测这种变动，从而达到研究地震的孕育和发生规律的目的。在通过倾斜仪观测到的数据的时间序列中，与地震有关的部分就是地震的信息，这在地震预报中称为异常；由于气温、气压、雨量、太阳和月亮的引力等环境因素而产生的地壳变动以及仪器、观测者本身引起的误差等则属于噪音。

可见，任何观测数据都是由两部分组成的，一部分是信息，一部分是噪音。信息部分与噪音部分的大小比值叫做“信噪比”。

信噪比是一个很重要的概念。在所观测到的数据序列中，如果信息的成分大于噪音的成分，才能把有用的信息识别或区分出

来；反之，则信息就会湮没于噪音之中，使我们毫无所得。例如，收音机发出的全部音响相当于观测数据，其中播送的音乐是信息，交流声及其他原因引起的杂音就是噪音；当噪音小于音乐的声响时，音乐就能分辨出来，相比之下，噪音越小，音乐越清晰，噪音越大则音乐越模糊，当噪音超过音乐的声响时，音乐就被噪音所掩盖而难于分辨出它的旋律了。

§5. 数据处理的目的

在取得了原始的观测数据之后，我们的任务并没有完结，相反，从某种意义上说，才仅仅是开始。因为最终的目的是应用这些数据来得到我们所需要的科学结论。为此，还经常需要对观测数据进行一系列的分析和处理。这种分析处理的工作往往比取得数据的过程更加复杂、更加困难，所采用的方法基本上是数学的分析处理方法，所以，我们把这一过程叫做观测数据的数学处理，或简称数据处理。

数据处理大体上有三方面的任务：

(1) 压制噪音，突出信号，提高信噪比。例如，用回归分析方法确定干扰因素，排除干扰成分；用数字滤波方法来"过滤"无规则的噪音等等。

(2) 对数据资料作出科学的评价。如根据误差理论定量地描述数据的可靠性和精确性等等。

(3) 对客观事物本身的发展规律和客观事物之间的相互关系作出定量的描述。例如在地震预报的研究中，通过对某一物理量的观测数据进行整理、归纳、分析，判断其中是否存在地震发生的信息，即有没有地震的前兆异常，如果有的话，与地震发生的三要素——时间、地点、震级之间存在着怎样的数量关系等等。

如前所述，由于原始的观测数据中混杂了许多所谓"噪音"的无效成分，直接用它来揭露事物的本质和规律是困难的。对观测数据进行数学处理的过程，也就是对这些资料进行"去粗取精，去

伪存真，由此及彼，由表及里”的加工制作过程，其目的是为了最充分地发挥资料的效能。

有关数据处理方面的知识，自从十八世纪末高斯创立误差理论以来得到了很大的发展和广泛的应用。近代科学技术的飞跃前进，常常要求用复杂的方法处理大量的数据，而电子计算机的问世和日臻完善，又使人力无法进行的复杂的数据处理过程变成现实，从而大大地促进了数据处理技术的发展。目前它已成了一门内容极其丰富、用途十分广泛的学科，所有科学技术领域，甚至许多社会科学部门都要用到这方面的知识。

本书仅根据我们的实际需要，主要对地震科学领域里经常要用到的一些基本方法和内容作简要的介绍。当然，对其他一些部门也有一定的参考价值。

第二章　有关的数学知识

在数据处理的内容中，要应用到一些高等数学和概率论方面的知识。为了帮助解决没有学过这些知识的读者在阅读本书时所产生的困难，这里把高等数学中的微积分和概率论的一些知识作概念性的介绍。读者需要作进一步的了解时，希望能参阅有关的书籍。

§1. 函数和函数曲线

能够变化的数值叫做变数，不能变化的数值叫做常数。例如，地球上任一地点的温度的数值是变数，它不仅在一年四季中随着季节的变化而变化，就是在一天中也随着时间的推移而不断的变化。而真空中光的传播速度 c 则是常数，它不因时间、空间的不同而变化。

如果一个变数 y 的变化情况是由另一个变数 x 的变化情况所决定，则叫变数 y 与变数 x 存在着函数关系。x 叫做自变量，y 叫做因变量，变数 y 又称为变数 x 的函数。这时，数学上常常用下面的形式来表示：

$$y=f(x). \tag{2-1}$$

这种函数关系可以很复杂，也可以很简单。如果能用一种具体的数学形式来表示 y 与 x 之间的数值关系，则这种关系式叫做函数公式或叫函数方程式。

最简单的一种函数关系叫线性关系，又叫线性方程。其公式表示如下：

$$y=ax+b, \tag{2-2}$$

其中 a 和 b 是常数。例如：假定汽车等速度前进，则汽车所走的

总路程 s 与时间 t 的关系为：

$$s=vt+s_0, \tag{2-3}$$

其中 v、s_0 为常数，v 是速度，s_0 是零时以前已经走的路程。

物体从空中向地面作自由落体运动时，降落行程 s 与降落时间 t 的函数关系就稍为复杂，其关系式为：

$$s=\frac{1}{2}gt^2, \tag{2-4}$$

其中 g 是重力加速度，是常数，则 $\frac{1}{2}g$ 也是常数。函数关系式的右边是包括变量 t^2 的较复杂的形式，叫抛物线方程。

一个自变量值只对应一个因变量值的函数，称为单值函数，对应多个因变量值的函数则称为多值函数。上面所举的例子都是单值函数，而大家熟悉的反三角函数则是多值函数。另外，上述都是只有一个自变量的情况，这种函数关系也叫单元函数关系。

为了更直观地表现自变量和因变量之间的变化关系，经常把函数关系化为图象。这里只讲单元函数的图示。

取二条互相垂直的轴线，叫坐标轴，其中横轴一般又叫 x 轴，代表自变量，纵轴一般叫 y 轴，代表因变量。若两轴的交点取为零，则在横轴和纵轴上分别按与零点的距离标出刻度，各自代表自变量和因变量的数值。一般说来，规定从零点向右和向上为正值，向左和向下为负值。这就是所谓的笛卡儿坐标系，也叫做直角坐标系。在实际使用中，常常把两轴的交点不取零，而按需要取分别对应于横轴和纵轴中的两个数，这时，相当于把横轴和纵轴作了一定距离的平移。

根据已知的函数关系，任何一个自变量的取值，都可以求出因变量的一个或几个数值，同时能够在坐标轴所在的坐标平面上相应地找到一个或几个点，每一个点表示由自变量和因变量构成的某一对数值。这些点要求它在 y 轴上的垂直投影点对应因变量的数值；在 x 轴上的投影点对应自变量的数值。

下面以单值函数为例，看看函数的图示情况：假定当自变量