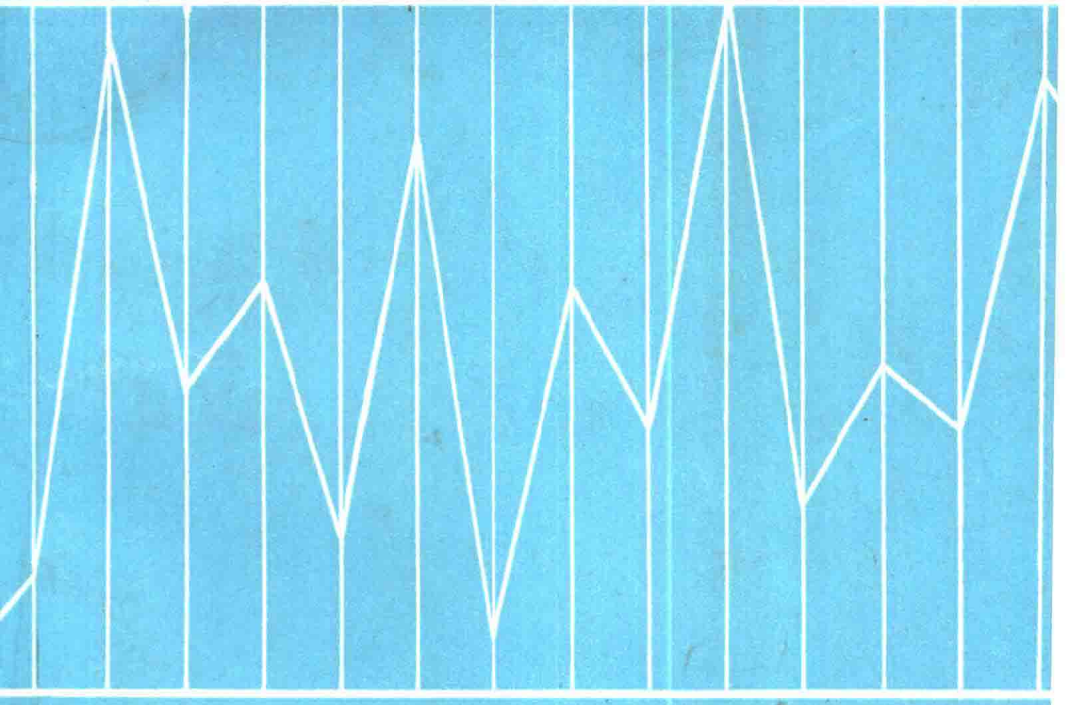
应用数理统计方法

陶 澍 编著



中国环境科学出版社

应用数理统计方法

陶 澍 编著

中国环境科学出版社
1994

(京)新登字 089 号

内 容 简 介

本书以环境科学及其相邻学科研究人员为主要对象，从应用角度出发，全面系统地介绍了经典数理统计方法。为便于读者掌握这些方法，书中还提供了大量实例。尽管这些实例多属环境科学范畴，但数理统计方法本身并无学科限制，其他学科的研究人员很容易参照这些例子，举一反三，将有关方法应用于本学科研究领域中去。

书中内容安排突破了按方法类别介绍的传统格局，完全以应用为目的，以研究者关心的随机变量的基本特征为主线，将有关方法分列于不同章节。读者很容易根据研究的需要，从有关章节中找到所需的具体方法。

全书介绍了数十种用途各异的数理统计方法。其中大部分是读者熟悉的常用手段，如比较两均值大小的 t-检验、单因子方差分析、一般的相关和回归分析等。另一些方法虽然很少被系统地介绍过，但在自然科学研究中却颇有实用价值，Box-Cox 正态变换、正态分布的 Lillifors 检验、不同模型的方差分析以及模型 I 回归分析即为其中几例。

应用数理统计方法

陶澍 编著
责任编辑 夏伟松
*
中国环境科学出版社出版发行
北京崇文区北岗子街 8 号
河北三河市宏达印刷厂 印刷
新华书店总店科技发行所发行 各地新华书店经售
*
1994 年 8 月第 一 版 开本 850×1168 1/32
1994 年 8 月第一次印刷 印张 14
印数 1—1500 字数 376 千字
ISBN 7-80093-547-7/X・811
定价：17.50 元

前　言

环境科学的研究对象是受诸多复杂因素影响的随机现象。数理统计方法正是研究这些随机现象内在规律的有效手段。充分利用数理统计方法分析各种环境现象是将环境科学研究提高到定量化水平的重要途径之一。

本书以笔者所授相关课程的内容和基本框架，补充了一些在环境科学研究中应用的实例，力图为环境科学、地球科学及其它相邻学科领域的研究人员提供一本具较强实用性的数理统计用书。

在环境科学研究中，错误使用数理统计方法的例子并不罕见。其原因之一是许多数理统计方面的书籍偏重于介绍方法本身，忽略了向读者交代建立这些方法的基本假设以及应当在什么条件下、如何使用有关方法。本书不仅详尽说明了每种具体方法的适用对象、数据要求和使用条件，还系统比较了相似统计方法的内在差别，使读者能够针对自己面临的问题正确地选择和使用书中提供的具体方法。

承蒙北大数学系程乾生教授对本书初稿进行审阅，在此表示诚挚的谢意。

一九九四年三月

于北京大学城市与环境学系

目　录

图表目录

绪　论

随着自然科学的不断发展，在任何学科领域中，仅仅依靠定性描述方法已远不能满足研究需要，定量研究方法日趋重要。然而象生物科学、大气科学、地球科学和环境科学等学科，其研究对象往往是一些受到多重复杂自然因素和人为因素影响的随机现象。一般的数学方法在这些现象的研究方面无能为力。正因为如此，这些学科研究的定量水平始终低于物理学等传统学科，而数理统计方法的发展和应用在提高此类学科研究的定量水平方面无疑具有举足轻重的地位。

数理统计方法是研究随机现象内在规律的科学。所谓随机现象是指那些在相同条件下，可能得到不同观测结果的现象。譬如用某一装置去除废水中的有机物，尽管废水流量和废水成份保持恒定，设备运转条件也可以维持不变，但处理后出水的有机物含量将在一特定范围内波动，而不可能是一个常数。此处所述出水有机物含量即为典型的随机量。虽然随机现象在个体上表现出随机性，但对它们进行大量观测的结果中却蕴含着内在的规律性。在上述水处理设施的例子中，这样的规律性就表现为出水有机物含量总是围绕着某一固定值上下波动，越接近该值，测定结果出现的可能就越大。揭示随机现象的内在规律正是运用数理统计方法的根本目的。

虽然应用数理统计方法在不同学科领域的应用过程中已发展成为不同的分支，如生物统计学、医药统计学、工程统计学和经济统计学等，但这些分支学科除了在具体研究对象方面有所不同以外，其方法学本身并没有根本区别。一般认为近代统计学的中心课

题是所谓统计推断(statistical inference),即在概率论的基础上,根据少量样本的观测结果对所研究对象的某些特征进行判断。举例说明,如果研究者对某条河流河水中多氯联苯(PCB)含量感兴趣,他只能从中采集数量有限的样品加以测定。一方面,这样的直接观测只是针对少数样本进行的,另一方面,又要求就全部河水的PCB含量特征得出某些结论,这就是典型的统计推断问题。一般认为,统计推断包括以下两个主要方面:

1) 参数估计(parameter estimation);

2) 假设检验(hypothesis test)。

有人甚至据此将统计学划分为描述统计学和推论统计学两个分支。前者通过对样本特征的统计描述反映总体的某些特点,即用样本统计量对相应的总体统计量进行估计。后者则根据一些专门设计的统计方法对研究者提出的各类假设的正确与否加以检验。仍以河水PCB含量为例,如果研究的最终目的仅限于了解河水PCB含量的高低或波动幅度等具体特征,那么研究者可以根据实际测定结果计算有关样本统计量(如均值和方差),并将其视为全部河水PCB含量相应统计量的估计值,这就是所谓参数估计。若研究者的兴趣不限于对河水PCB含量一般特征的了解,而且关心另一些更复杂、更深入的问题,那么单纯的参数估计方法就不能满足要求了。当问题为该河流的两个或多个断面上河水的PCB含量是否相同?不同季节的河水PCB含量值有没有差别?河水PCB含量的频率分布是否服从正态形式抑或呈对数正态分布?河水PCB含量与其它参数如悬浮物含量、pH值和腐殖酸含量等有没有共同消长关系?能否利用一些常用的水化学参数观测值对PCB含量进行预测等。对上述问题,研究者可以首先提出自己的假设,然后再用假设检验方法对这些假设的正确性作出统计判断。

尽管参数估计与假设检验是应用数理统计方法的主要课题,但毕竟不是它的全部内容。另一种比较全面的分类方法将数理统计的应用领域概括为以下五个方面:

1）参数估计(parameter estimation)；

2）假设检验(hypothesis test)；

3）回归问题(regression problem)；

4）多重决策(multiple decision)；

5）其它问题(如采样方法、试验设计和预测方法等)。

这种分类方式将回归问题单列出来。一般的回归分析实际上包含参数估计和假设检验两个方面。回归分析中对两个或多个变量间数量关系特征的描述(如回归系数的计算)实质上是一种参数估计，而对有关估计值的一系列显著性检验则属于假设检验问题。多重决策是一般假设检验的推广。假设检验只能在是与否，即接受或拒绝某种假设之间进行判断，而多重决策允许在两种以上(有限种)可能性之间作出抉择。准确地说，是将这些可能性按概率大小排序。由于对发生在自然界中许多问题的解答远非是或否那样简单，目前尚不成熟的多重决策方法具有十分诱人的发展前景。

按照上述分类方式，本书的内容主要包括参数估计、假设检验和回归分析三个方面。参数估计的具体方法将在第一章中详述；第二、第三和第四章分别介绍不同目的的假设检验方法；第五章则集中讨论回归分析及与之有关的相关分析。此外，将在第一章中用少量篇幅简述采样问题。

为了更好地掌握和运用数理统计方法，对概率论和统计学原理的深入钻研固然具有重要意义，然而对大多数自然科学工作者而言，他们更迫切需要的不是理论上的探讨，而是一本"菜谱"式的手册，以便能够参照其中提供的方法和例子，直接使用有关方法解决本专业领域中的特殊问题。笔者的目的正是为了提供这样一本纯粹以应用为目的的书籍。为此，书中略去了所有公式推导。对绝大多数读者来说，他们完全可以在不了解特定计算公式推导过程的前提下正确使用有关方法。限于篇幅，本书也没有包括诸如大数定律和中心极限定理等统计学的基本定理，读者很容易从其它书籍中找到这方面的内容。

出于同一目的，书中各类统计方法完全根据研究目的安排。最典型的例子是大多数数理统计书籍将非参数检验方法单独列为一章，而本书则根据检验方法的基本目的分别将它们与相应的参数方法一起介绍。运用统计方法的最终目的可以按照总体大小、离散程度和分布形式三大特征加以区分（详见第 1.2 节），本书中所有方法正是围绕着这三类特征编写的。此外，笔者选择的都是常用的数理统计方法。有些所谓快速检验，象正态分布的 D'Agostino 检验和作图检验以及非参数相关的 Olmstead-Tukey 偶角检验等在计算机技术得以普及的今天，已失去了其仅有的计算简便的优点，故不再介绍。此外，对那些仅仅在计算方法上有所区别，而其它方面完全等价的检验方法一般只择其一种详细介绍（如略去了与 Mann Whitney U-检验很相似的 Wilcoxon 秩和检验）。与此同时，本书侧重介绍了一些反映应用数理统计发展方向的方法，如正态分布的 Lillifors 检验、模型 II 回归分析和方差分析的不同模型等。为了帮助读者理解和掌握有关方法的计算和使用，在介绍每种方法时都提供了相应的例题。限于笔者的专业知识，所选例题多属环境科学范畴，但熟悉其它学科的读者应很容易将方法本身"移植"到不同学科领域。

几乎对每一种具体的研究问题，应用数理统计方法均提供了若干可供选择的手段。究竟采用哪种方法更为恰当常常是一个令人困惑的问题。不十分熟悉应用数理统计的研究者常常会面对数种相似的选择而无所适从，错误地使用统计方法的例子也并不罕见。为此，本书用了大量篇幅去说明每类以至每种方法的适用范围，并尽可能对相似方法的内在差别及优缺点加以比较，目的在于为读者提供一些方法选择方面的基本依据。

计算机的普及为数理统计方法的推广使用提供了重要条件。一些复杂的统计计算在没有计算机帮助的情况下几乎是无法实现的。目前流行的应用数理统计计算机软件包括两种类型。其中大量的是为特定目的编写的专用软件，如用于方差分析的软件、用于

回归分析的软件等。另一类软件具有多种功能，通常被称为数理统计软件包(statistical software package)。不同数理统计软件包的性能相差悬殊，但一般具有以下几方面基本功能：

1）数据输入；

2）数据管理(如再次采样、加权以及获得新的衍生变量等)；

3）报表和图形生成；

4）各类统计分析。

随着微型计算机的迅速发展和普及，很多重要的数理统计软件包已经移植到微机上。其中最著名的包括 SAS(Statistical Analysis System）和 SPSS（Statistical Package for Social Sciences)。前者几乎是最复杂也是功能最强的微机统计软件包，后者虽然是专门为社会科学研究设计的，但也颇受自然科学研究者的欢迎。除此之外，象为生物学统计和医学统计编制的 BMDP、强调人机对话和预测功能的 IDA、通用性和移植性都很好的 STAPAK、图形能力较强且直观性好的 STATGRAPH 以及专供不熟悉计算机操作的研究者使用的 MINITAB 等目前都很流行。由于这类软件包都经过系统检测和长期试用，与研究者手工计算或用自编软件计算相比，它们的使用在保证统计结果的可靠性和可比性方面具有明显的优越性。这也是越来越多研究者在发表论文的同时具体指明他所采用的数理统计软件包名称的原因。

没有合适的数理统计用表，正确运用数理统计方法几乎是不可想象的。本书的附录提供了本书介绍的各种方法用到的统计用表。为避免译名不同引起的混乱，书中介绍的检验方法均使用英文原名，如用 Grubbs 检验而不是格拉布斯检验。使用符号也尽可能统一，并全部列在附录中。