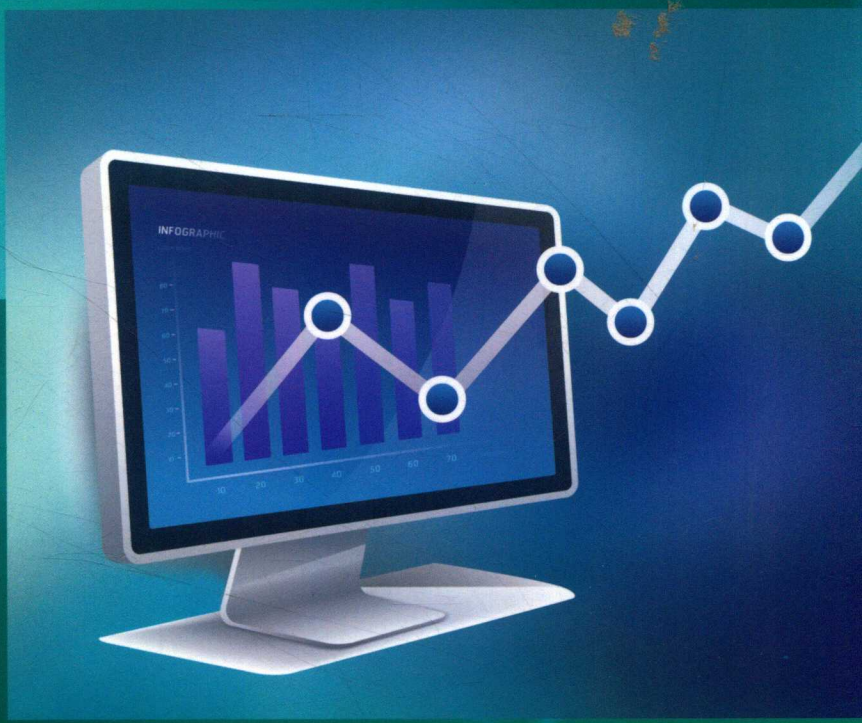
普通高等学校信息与计算科学专业系列丛书

普通高等教育“十一五”国家级规划教材

数据分析方法

（第二版）

梅长林 范金城 编



高等教育出版社

普通高等学校信息与计算科学专业系列丛书

普通高等教育“十一五”国家级规划教材

数据分析方法

（第二版）

梅长林　范金城　编

高等教育出版社·北京

内容简介

本书是为高等学校信息与计算科学专业本科生“数据分析”课程编写的教材，主要介绍常用统计数据分析的基本内容与方法，包括数据描述性分析、回归分析、方差分析、主成分分析与典型相关分析、判别分析、聚类分析、Bayes 统计分析等。另外，对 SAS 软件的基础知识以及与上述各数据分析方法有关的 SAS 过程做了简要介绍，以便于利用 SAS 软件实现各分析方法的应用。各章均配备了丰富的有广泛实际背景的习题。

本书也可作为高等学校统计专业本科生和非数学类硕士研究生教材以及数据分析工作者的参考书。

图书在版编目(CIP)数据

数据分析方法 / 梅长林，范金城编. -- 2 版. -- 北京 : 高等教育出版社，2018.10
(普通高等学校信息与计算科学专业系列丛书)
ISBN 978-7-04-050124-7

Ⅰ. ①数… Ⅱ. ①梅… ②范… Ⅲ. ①统计数据-统计分析-高等学校-教材 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2018)第 160008 号

策划编辑 李冬莉　责任编辑 李冬莉　封面设计 李小璐　版式设计 马敬茹
插图绘制 于 博　责任校对 胡美萍　责任印制 田 甜

出版发行 高等教育出版社
社　　址 北京市西城区德外大街 4 号
邮政编码 100120
印　　刷 北京宏伟双华印刷有限公司
开　　本 787mm × 960mm 1/16
印　　张 20.5
字　　数 370 千字
购书热线 010 - 58581118
咨询电话 400 - 810 - 0598

网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
版　　次 2006 年 2 月第 1 版
2018年10月第 2 版
印　　次 2018年10月第 1 次印刷
定　　价 39.30 元

物 料 号 50124 - 00

数据分析方法

（第二版）

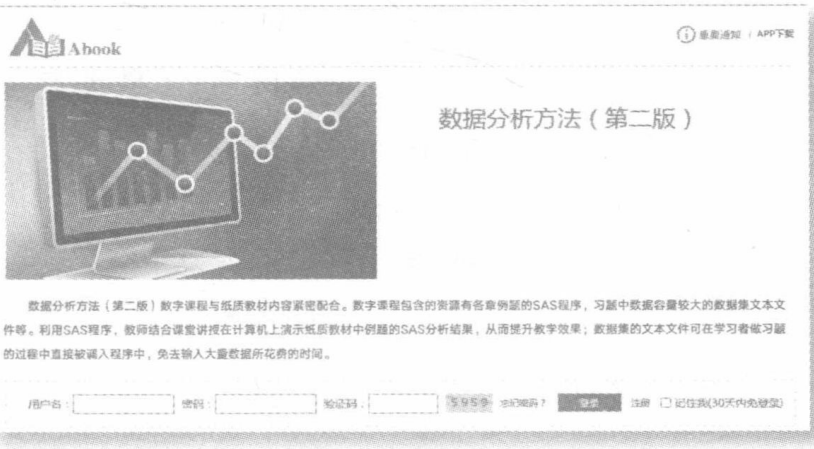
梅长林　范金城

1　计算机访问http://abook.hep.com.cn/122390904，或手机扫描二维码、下载并安装Abook应用。

2　注册并登录，进入“我的课程”。

3　输入封底数字课程账号（20位密码，刮开涂层可见），或通过Abook应用扫描封底数字课程账号二维码，完成课程绑定。

4　单击“进入课程”按钮，开始本数字课程的学习。



课程绑定后一年为数字课程使用有效期。受硬件限制，部分内容无法在手机端显示，请按提示通过计算机访问学习。

如有使用问题，请发邮件至abook@hep.com.cn。

http://abook.hep.com.cn/122390904

信息与计算科学专业系列教材编委会

总　序

根据教育部1998年颁布的普通高等学校专业目录,“信息与计算科学”专业被列为数学类下的一个新专业(它覆盖原有的计算数学及其应用软件、信息科学与运筹控制等专业)。这一新专业的设置很好地适应了新世纪以信息技术为核心的全球经济发展格局下的数学人才培养与专业发展的需要。然而,作为一个新专业,对其专业内涵、专业规范、教学内容与课程体系等有一个自然的认识与探索过程。教育部数学与统计学教学指导委员会数学类专业教学指导分委员会(下称教指委)经过过去两年艰苦细致的工作,对这些问题现在已有了比较明确的指导意见,发表了《关于信息与计算科学专业办学现状与专业建设相关问题的调查报告》及《信息与计算科学专业教学规范》(讨论稿)(见《大学数学》第19卷1期(2003))。为此,全国高等学校教学研究中心在承担全国教育科学“十五”国家级规划课题——“21世纪中国高等教育人才培养体系的创新与实践”研究工作的基础上,根据教指委所颁布的新的教学规范,组织国内各高校的专家教授,进行其子项目课题“21世纪中国高等学校信息与计算科学专业教学内容与课程体系的创新与实践”的研究与探索。为推动本专业的教材建设,该项目课题小组与高等教育出版社联合成立了“信息与计算科学专业系列教材编委会”,邀请有多年教学和科研经验的教师编写系列教材,由高等教育出版社独家出版,并冠以教育科学“十五”国家级规划课题研究成果。

按照新的《信息与计算科学专业教学规范》(讨论稿),信息与计算科学专业是以信息技术和计算技术的数学基础为研究对象的理科类专业。其目标是培养具有良好的数学基础和数学思维能力,掌握信息与计算科学基础理论、方法与技能,受到科学研究的训练,能解决信息技术和科学与工程计算中的实际问题的高级专门人才。毕业生能在科技、教育、信息产业、经济与金融等部门从事研究、教学、应用开发和管理工作,能继续攻读研究生学位。根据这一专业目标定位和落实“强基础、宽口径、重实际、有侧重、创特色”的办学指导思想,我们认为,本专业在数学基础、计算机基础、专业基础方面应该得到加强,各学校在这三个基础方面可大体一致,但专业课(含选修课)允许各校自主选择、体现各自特点。考虑到已有大量比较成熟的数学基础与计算机基础课程教材,本次教材编写主要侧重于专业基础课与专业课(含选修课)方面。

信息与计算科学，就其范畴与研究内容而言，是数学、计算机科学和信息工程等学科的交叉，已远远超出数学学科的范畴。但作为数学学科下的一个理科专业，信息与计算科学专业则主要研究信息技术的核心基础与运用现代计算工具高效求解科学与工程问题的数学理论与方法（或更简明地说，研究定向于信息技术与计算技术的数学基础），这一专业定位明显地与计算机科学与信息工程专业构成区别。基于这一定位，信息与计算科学专业可包括信息科学与科学计算（计算数学）两个大的方向。我国在科学计算方向已有长期的办学经验，科学计算方向通常被划分为偏微分方程数值解、最优化理论与方法、数值逼近与数值代数、计算基础等学科子方向。然而，对于信息科学，它到底应该怎样划分学科子方向？应该怎样设置专业与专业基础课？所有这些都仍是正在探索的问题。

任何技术都可以认为是延伸与扩展人的某种功能的方式与方法，所以信息技术可以认为是扩展人的信息器官功能的技术。人的信息器官主要包括感觉器官、传导器官（传导神经网络）、思维器官和效应器官四大类型，其功能则主要是信息获取、信息传输、信息处理和信息应用（控制），因而感测技术、通信技术、智能技术与控制技术通常被认为是最基本的信息技术（常称之为信息技术的四基元），其他信息技术可认为是这四种基本技术的高阶逻辑综合或分解衍生。所以可以把信息科学理解为是“有关信息获取、信息传输、信息处理与信息控制基础的科学”。从这个意义上，我们认为：信息处理（包括图像处理、信号分析等）、信息编码与信息安全、计算智能（人工智能、模式识别等）、自动控制等可构成信息科学的主要学科子方向。这一认识也是教指委设置信息与计算科学专业信息科学方向课程的基本依据。

本系列教材正是基于以上认识，为落实新的《信息与计算科学专业教学规范》（讨论稿）而组织编写的。我们相信，该系列教材的出版对缓解本专业教材的紧缺局面，对推动信息与计算科学专业的快速与健康发展会大有裨益。

从长远的角度看，为适应不同类型院校和不同层次要求的课程需求，本系列教材编委会还将不断组织教材的修订和编写新的教材，从而使本专业的教学用书做到逐步充实、完善和多样化。我们诚恳地希望使用本系列教材的教师、同学们及广大读者对书中存在的问题及时指正并提出修改意见和建议。

信息与计算科学专业系列教材编委会

2003 年 8 月 31 日

第二版前言

本书是为全国普通高等学校信息与计算科学专业数据分析课程所编写的教材，自2006年出版以来，承蒙广大读者及各院校的支持，已连续印刷10余次，总发行量近3万册。但在我校及兄弟院校多年使用本教材的过程中，也发现一些不妥和纰漏之处，另外，部分内容也需要更新和补充。为此，在保持原教材特色的基础上，主要针对如下几个方面做了修订。

首先，考虑到回归分析是数据分析的重要方法，其内容十分丰富且应用背景极其广泛，我们在原线性回归分析的基础上补充了广义线性模型中最常用的logistic回归模型，并在章末增添了从参数回归模型到非参数回归模型的概括性综述，以期使读者对回归分析模型的多样性及不同特点有初步了解，为进一步学习近代回归分析方法并将其应用于不同类型的数据分析开启窗口。

其次，考虑到目前所使用的SAS软件大多都是SAS V.8及以上版本，将原教材中以SAS V.6版本为主介绍SAS系统的界面及其功能的相关内容修订为以高版本为主予以介绍，同时增加了logistic回归分析的SAS过程的相关内容。

最后，对全书进行了全面的修订，改正了原书稿中的一些不妥和纰漏之处，使叙述和分析更为清晰和严谨。各章例题的SAS程序连同各章习题中容量较大的数据集的文本文件，可通过网址 http://abook.hep.com.cn/122390904 下载，例题的SAS程序可供教师教学时参考，习题的数据文件可供学生上机完成课后习题时直接调入，免去输入大量数据所花费的时间。

新增内容大约可在4学时内完成。根据我们的教学实践，一种可供参考的教学方法是与多媒体教学相结合，首先讲授第8章中的SAS基本内容简介，使学生初步了解SAS软件并掌握必要的操作和简单的编程技能，然后结合各章具体内容进行讲授。将相关SAS过程的介绍分散到各章，通过对例题分析的多媒体演示与对输出结果的解释，使学生掌握各SAS过程的应用。这样，全书内容估计可在56学时内讲授完毕，再配以8学时左右的集中上机实习，共需要约64学时。对于该课程不足64学时的院校，可略去第7章Bayes统计分析的内容。

衷心感谢高等教育出版社编辑对修订本书的鼓励、指导与帮助。另外，本书的修订得到西安交通大学教务处教材建设专项基金的资助，编者的博士研究生张晓承担了新增内容的录入工作，谨表示感谢。

编　者

2017 年 12 月于西安交通大学

第一版前言

数据作为信息的主要载体在当今信息化社会中扮演着重要的角色。各行各业的各个领域无处不有数据的存在，数据为我们提供了丰富的信息。然而，如何从大量的看似杂乱无章的数据中揭示其中隐含的内在规律、发掘有用的信息以指导人们进行科学的推断与决策，还需要对这些纷繁复杂的数据进行分析。

顾名思义，数据分析就是分析和处理数据的理论与方法，从中获得有用的信息。从这个意义上讲，数据分析不存在固定的解决方法，分析的目的和分析的方法不同，会从同一数据中发掘出各种有用信息。因此，数据分析内容丰富、方法众多，尤其借助计算机的强大计算能力，各种数据分析方法层出不穷，并得到空前的发展。然而，作为一门学科，数据分析的性质定位问题目前尚未得到很好的解决，还未见到对“数据分析”共同认可的确切定义，但以数据为主要研究对象的统计方法无疑是最重要的数据分析工具之一。按照全国普通高等院校信息与计算科学专业委员会的课程设置思路，数据分析教材应以介绍数据分析的常用统计方法为主。

我们注意到，数据分析和统计学在内涵上还是有差异的。就编者的理解，数据分析的基本命题是从数据中挖掘尽可能多的有用信息，面对数据，应强调可解决什么样的问题，如何解决。可以说数据分析是统计学理论与方法的综合应用，更注重解决实际问题的全过程。在本教材内容选取与编写中，也力图体现这一点。

本书在数据的描述性统计分析的基础上，重点介绍了一些应用十分广泛的多元数据分析的统计方法，包括线性回归分析、方差分析、主成分分析、典型相关分析、判别分析、聚类分析等。另外，鉴于 Bayes 统计分析在信息科学中得到越来越广泛的应用，本书对 Bayes 统计的基本理论与方法也作了适当介绍。

数据分析需要处理大量的数据，进行复杂的运算，因此计算机和现代统计软件的使用似乎是必不可少的。本书的编写与当今国际上著名的数据分析软件系统之一的 SAS 软件紧密结合，并在最后一章对 SAS 软件的基本知识以及与本书内容有关的 SAS 过程作了简要介绍。书中大部分例题都结合 SAS 软件予以分析，各章习题大多也需要借助计算机软件来完成。为便于教学，本书各章例题的 SAS 程序和习题中数据容量较大的数据集文本文件可从网上下载，希望能结合

课堂讲授在计算机上演示书中例题的SAS分析结果,提高教学效果;另外,数据集的文本文件可在学生上机完成习题时直接调入程序中,免去输入大量数据所花费的时间。

本书在写作上一方面与SAS软件的应用紧密结合,另一方面也注重严格的理论推导和方法步骤的详细介绍以及对各方法统计思想的阐述和对分析结果的解释,以提高学生分析问题和解决问题的能力,避免只盲目地使用计算机软件得到结果,而对其信息内涵理解不深。例题和习题的选取一般均是有实际背景的涉及众多领域的实际观测数据,其中不乏我国国民经济等领域内的实际数据。希望通过这些例题,让学生看到统计数据分析方法的具体应用,体会数据分析的全过程。

学习本教材需有数学分析、线性代数、概率论与数理统计的基础知识。根据我们的教学实践,建议各章的教学时数如下:

第1章:数据描述性分析,6学时;

第2章:线性回归分析,8学时;

第3章:方差分析,6学时;

第4章:主成分分析与典型相关分析,6学时;

第5章:判别分析,6学时;

第6章:聚类分析,4学时;

第7章:Bayes统计分析,8学时;

第8章:SAS软件及有关数据分析过程简介,8学时。

再配备12学时左右的上机实习,全书内容大约可在64学时内完成。第8章中的§8.1内容建议首先讲授,§8.2内容可分散到前七章介绍,再通过上机实习和完成各章习题使学生加以巩固。鉴于本书各章的模块化结构,各校可根据自身情况予以适当取舍。

本书第2,3,4,8章由梅长林执笔,第1,5,6,7章由范金城执笔,最后由梅长林统一修改定稿。衷心感谢审稿者对本书初稿提出的宝贵修改意见和对数据分析课程定位等方面的讨论。感谢全国信息与计算科学专业指导委员会对编写本教材的大力支持以及高等教育出版社李蕊、王瑜编辑的辛勤工作。限于编者的水平和对数据分析的理解,书中难免会有不妥之处,恳望读者不吝指教,以期提高本教材的质量。

编 者

2005年11月于西安交通大学

目　录

第 I 章

数据描述性分析

数据的描述性分析即是从数据出发概括数据特征，主要包括数据的位置特性、分散性、关联性等数字特征和反映数据整体结构的分布特征，它是数据分析的第一步，也是对数据进行更进一步分析的基础.

本章重点介绍一维和多维数据描述性分析的基本内容，包括数据的数字特征与分布特征的描述与分析.另外，基于后续各章的需要，对多维正态分布的定义和性质也作了简单介绍.

§1.1 一维数据的数字特征

设有 n 个一维数据

$$x_1, x_2, \cdots, x_n,$$

它们是从所研究的对象(即总体) X 中观测得到的，这 n 个值称为样本观测值，n 称为样本容量.数据分析的任务就是要对该样本观测值进行分析，提取数据中所包含的有用的信息，进一步对总体的信息做出推断.对于作为信息载体的数据，我们首先要用某些简单的量概括其中包含的主要信息或特征，这些量称之为数据的数字特征，包括数据的集中位置、分散程度、数据分布的形状特征等等.

1.1.1 表示位置的数字特征

1. 均值

均值即是 $x_1, x_2, \cdots, x_n$ 的平均数

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.1}$$

它描述了数据取值的平均位置.

在通常情况下，均值有许多优良的统计性质，这些在必修课数理统计基础部分已得到广泛讨论.然而，当数据中存在异常值时，它则缺乏抗扰性或稳健性，即易受异常值的影响而使其值有较大变化.因此，在数据分析中，还要考虑其他一些描述位置的数字特征.

设 $x_1, x_2, \cdots, x_n$ 是 n 个观测值，将它们从小到大记为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$，即

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)},$$

称它们为次序统计量值，其中第 i 个次序统计量值是 $x_{(i)}$. 特别，最小次序统计量值 $x_{(1)}$ 与最大次序统计量值 $x_{(n)}$ 分别为

$$x_{(1)} = \min_{1 \leqslant i \leqslant n} x_i, \qquad x_{(n)} = \max_{1 \leqslant i \leqslant n} x_i.$$

2. 中位数

中位数的计算公式是

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ 为奇数}, \\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ 为偶数}. \end{cases} \tag{1.2}$$

中位数是描述数据的中心位置的数字特征，比中位数大或小的数据个数大体上为整批数据个数的一半. 对于对称分布的数据，均值与中位数比较接近，对于偏态分布的数据，均值与中位数差异会较大. 中位数的一个显著特点是受异常值的影响较小，具有较好的稳健性或抗扰性，是数据分析中相当重要的一个统计量.

3. 分位数

对 $0 < p < 1$，数据 $x_1, x_2, \cdots, x_n$ 的 p **分位数**是

$$M_p = \begin{cases} x_{([np]+1)}, & np \text{ 不是整数}, \\ \dfrac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & np \text{ 是整数}, \end{cases} \tag{1.3}$$

其中 $[np]$ 为 np 的整数部分. 当 $p = 0$ 时，定义 $M_0 = x_{(1)}$；当 $p = 1$ 时，定义 $M_1 = x_{(n)}$.

大体上整批数据的 $100p\%$ 的观测值不超过 p 分位数，第 0.5 分位数就是中位数 M. 在实际应用中，0.75 分位数与 0.25 分位数比较重要，它们分别称为上、下四分位数，并分别简记为

$$Q_3 = M_{0.75}, \qquad Q_1 = M_{0.25}.$$

均值 $\bar{x}$ 与中位数 M 皆是描述数据位置的数字特征，在正常情况下，$\bar{x}$ 比 M 有更优良的性质，能更充分反映数据的信息. 而当数据中有异常值时，M 具有很强的稳健性. 考虑到既要充分利用样本信息，又要有较强的稳健性，可用如下的三均值 $\hat{M}$ 作为概括数据位置的数字特征.

4. 三均值

三均值 $\hat{M}$ 的计算公式是

$$\hat{M} = \frac{1}{4}Q_1 + \frac{1}{2}M + \frac{1}{4}Q_3, \tag{1.4}$$

即 $\hat{M}$ 是 Q_1, M, Q_3 的加权平均，权重分别是 $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$.