工商管理优秀教材译丛
PEARSON Prentice Hall
管理学系列

Applied Multivariate Statistical Analysis Sixth Edition

实用多元统计分析

第 6 版

（美）理查德·A.约翰逊（Richard A. Johnson）
迪安·W.威克恩（Dean W. Wichern） 著

陆璇 叶俊 译

PEARSON Education
清华大学出版社

实用多元统计分析

第 6 版

（美）理查德·A.约翰逊（Richard A. Johnson）
迪安·W.威克恩（Dean W. Wichern）著

陆璇　叶俊　译

Applied Multivariate Statistical Analysis (Sixth Edition)

清华大学出版社
北京

北京市版权局著作权合同登记号　图字：01-2007-5699

图书在版编目(CIP)数据

实用多元统计分析：第6版/(美)约翰逊(Johnson, R. A.)，(美)威克恩(Wichern, D. W.)著；陆璇，叶俊译. —北京：清华大学出版社，2008. 11
(工商管理优秀教材译丛·管理学系列)
书名原文：Applied Multivariate Statistical Analysis, 6e
ISBN 978-7-302-18343-3

Ⅰ. 实…　Ⅱ. ①约…　②威…　③陆…　④叶…　Ⅲ. 多元分析：统计分析—高等学校—教材
Ⅳ. O212. 4

中国版本图书馆CIP数据核字(2008)第119261号

责任编辑：江　娅
责任校对：王凤芝
责任印制：孟凡玉
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦A座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京市清华园胶印厂
装 订 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185×260　**印　张**：38　**插　页**：2　**字　数**：897千字
版　　次：2008年11月第1版　　**印　　次**：2008年11月第1次印刷
印　　数：1～4000
定　　价：68.00元

译者序 实用多元统计分析 Applied Multivariate Statistical Analysis

多元统计分析是统计学中内容十分丰富、应用范围极为广泛的一个分支.在自然科学和社会科学的许多学科中,研究者都有可能需要分析处理有多个变量的数据的问题.从表面上看起来杂乱无章的数据中发现和提炼出规律性的结论,不仅需要对所研究的专业领域有很好的训练,而且需要掌握必要的统计分析工具.对实际领域中的研究者和高等院校的研究生来说,要学习掌握多元统计分析的各种模型和方法,手头有一本好的,有长久价值的参考书是非常必要的.这样一本书应该满足以下条件:首先,它应该是"浅入深出"的,也就是说,既可供初学者入门,又能使有较深基础的人受益.其次,它应该是既侧重于应用,又兼顾必要的推理论证,使学习者既能学到"如何"做,而且在一定程度上了解"为什么"这样做.最后,它应该是内涵丰富、全面的,不仅要基本包括各种在实际中常用的多元统计分析方法,而且还要对现代统计学的最新思想和进展有所介绍、交代.

我们认为R. A. 约翰逊(R. A. Johnson)和D. W. 威克恩(D. W. Wichern)合著的《实用多元统计分析》(第6版)(Applied Multivariate Statistical Analysis, 6^{th} Ed)就是这样一本有长久价值的多元统计分析参考书.本书的内容十分丰富,涵盖了多元统计分析的各种有广泛应用的、经典和现代的模型和方法.为便于学习者对方法的理解,本书对各种方法的计算过程介绍得很详细,从简单的数据出发,将计算过程一步一步、不厌其烦地展开,使学习者对计算方法获得感性认识.在此基础上,对实际数据用计算机软件进行分析,并给出输入方法的说明和输出结果的解释,使得学习者能够掌握统计软件的操作步骤,并读懂输出结果的含义.书中有大量来自实际问题的案例,通过对这些案例的分析,学习者可以学到如何将实际问题转化为恰当的统计问题,进而选择恰当的模型与方法来分析.

由清华大学出版社出版的本书第4版的中文译本(2000年)在我国拥有广泛的读者群,并且已经被选为许多高校本科或研究生相关课程的教科书.作为第4版的中文译者,我们又有幸翻译了本书的第6版.第6版对第4版有所修改与补充,在作者原序中已经作了较为具体的说明.我们在这里再强调一下第6版新增加的两个内容.首先是在第11章中增加了逻辑斯蒂回归(Logistic regression).平行于经典的线性回归,逻辑斯蒂回归处理对二值响应变量的预测(分类)问题.其次是在第12章中增加了一节补充材料,介绍近年来非常热门并得到广泛应用的数据分析领域——数据挖掘(data mining),以及多元统计分析方法在数据挖掘中的应用.

本书由叶俊翻译第1章至第6章,由陆璇翻译第7章至第12章并统稿;研究生冯汉杰、向红旭、廖新国和郑灿亮也参与了翻译工作。由于我们的水平有限,因此在翻译中难免有错误或不妥之处.希望各位专家和读者发现问题后及时告知我们,以便以后有机会时予以更正.

译者

2008年7月于清华园

原著序 实用多元统计分析
Applied Multivariate Statistical Analysis

读者对象

本书最初来自我们为威斯康星大学麦迪逊分校统计系和商学院开设的“实用多元分析”课程的讲稿.《实用多元统计分析》第6版的内容,介绍了描述和分析多元数据的统计方法.尽管数据分析在一个变量时就很有趣味,但当涉及几个变量时,它才真正变得具有吸引力和富于挑战性.生物学、物理学和社会科学诸领域中的研究人员经常收集几个变量的测量结果,而现代计算机程序包则能轻易地提供复杂统计分析的数值结果.本书试图为读者提供一些必要的支持性知识,使他们能对统计分析结果作出适当的解释,能选择恰当的分析方法并了解这些方法的优点和缺点.我们希望,本书内容能满足实验科学家们的需要,在广泛多样的研究课题领域内,成为一本对多元观测结果进行统计分析的入门书.

水平

我们的目标是在这样的水平上介绍多元分析的概念和方法:使那些已学过两门或更多门统计学课程的读者能毫无困难地理解这些内容.本书侧重讨论多元方法的应用,因而我们尽可能地使数学有趣味.书中避免使用微积分,另一方面,矩阵及矩阵变换的概念却十分重要.我们假定读者不熟悉矩阵代数.所以当矩阵自然地出现在讨论中时,我们会先对它进行介绍,然后告诉你它如何简化了多元模型及方法的叙述.

本书第2章介绍了矩阵代数,对矩阵代数应用于多元分析时的一些重要结论作了强调.第2章的补充部分为那些很少或从未接触过这一学科的人提供了一个矩阵代数结论汇总.补充材料不仅使本书在内容上实现自给自足,而且被用来完成各种论证.这些论证在初次阅读时可以跳过.我们希望通过这种方式使本书能为更多的读者所接受.

为了吸引从事实际工作和理论工作的广大读者学习多元分析,我们不得不在某种程度上牺牲本书内容难度的一致性:有些章节的难度要比其余部分大些.特别在第7章中,我们概括了有关回归问题的大量材料,而结论表述又相当简略,因而初次阅读时会感到很困难.希望教师们能在选择适合学生的章节时设法弥补这种不平衡性,必要时可降低要求.

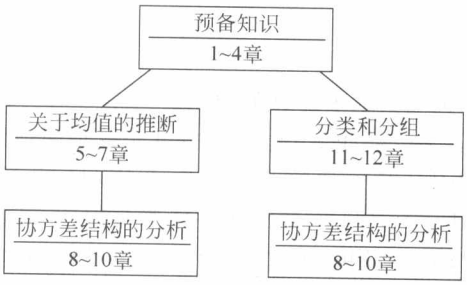
组织和方法

第5章至第12章讨论多元分析的方法论“工具”.这几章是本书的核心,不过要是没有第1至第4章的大量导论性材料,这些内容是无法理解的.即使对矩阵代数具备良好知识或愿意接受数学结论的读者,也应至少精读第3章的样本几何和第4章的多元正态分布.

在涉及方法论的各章中,我们的做法是使讨论直截了当而又有条不紊地进行.典型的情况是,我们从总体模型的表述开始,对相应的样本结果作出描述,然后用例子对每件事情作出解释.例子分两类:一类比较简单,其中的计算可用手算轻易完成;另一类则依赖实际数据和计算机软件.这些情况提供了一种机会,使我们能够:(1)重复做过的分析,(2)完成练习所要

求的分析，或(3)用我们未曾用过或未曾提出过的另一种方法来分析数据.

将叙述方法论的各章(5至12章)分成三部分，就能使教师们在将材料剪辑成适合自己需要的课程方面具有某种可塑性. 兹将一学期(两个季度)课程的可能安排图示如下：



每个教师无疑都会放弃某几章中的某几节内容，以形成比上述两种选择更宽广的课题选择.

对于大多数学生，我们建议他们迅速通过最初4章(主要是第1章，2.1节，2.2节，2.3节，2.5节，2.6节和3.6节，以及第4章中“评估正态性假定”那一节)，然后便选择方法论课题. 例如，人们也许会讨论均值向量、主成分、因子分析、判别分析和聚类等的比较. 这些讨论可以以包含在这几节中的那些“设计出来的”例子为特色. 教师可依靠图表和文字说明来讲授相应的理论推导. 如果学生们的数学水平较高而且程度均衡，本书大部分内容可在一学期内成功地授完.

我们发现个别数据分析方案有助于将来自方法论各章的材料组成一个整体. 所以我们对多元方差分析、回归分析、因子分析、典型相关、判别分析等内容进行相当完善的论述是有益的，尽管这些内容在讲课时也许不会专门提到.

第6版的内容更动

新材料 本书前几版的读者会注意到第6版有以下几个改动.

- 12个新的数据集，主要有：各国男子和女子的径赛记录，心理评分，汽车车身装置测量结果，手机塔故障，纸浆和纸张的性能测试，马里家庭农场数据，股票收益率以及布拉索斯河的水蛇数据.
- 添加了37道新习题和修正了20道原来的习题，而且其中很多习题都是基于新的数据集.
- 添加了4个基于新数据的习题并修正了15个习题.
- 6个新添的或扩充的节

1. 6.6节的协方差矩阵相等性的检验
2. 11.7节的逻辑斯蒂(Logistic)回归和分类
3. 12.5节的基于统计模型的聚类
4. 扩充了6.3节，新增了“样本量不大时，对正态总体的 T^2 分布的逼近”
5. 扩充了7.6节和7.7节，新增了赤池(Akaike)信息准则
6. 将以前关于两组别的判别式分析的11.3节和11.5节合并成11.3一节

网址 为了突出多变量分析的方法，我们删除了结论7.2，7.4，7.10和10.1冗长的证明，但是将证明放在了以下网站：www.prenhall.com/statistics. 点击“Multivariate Statistics”，然

后点击本书可获得这些证明. 此外,本书所用的所有数据集的 ASCII 文件可以在该网站下载.

教师使用手册 我们在网站 www. prenhall. com/statistics 上提供了教师使用手册. 如果想知道更多与本书相关的资料或更多感兴趣的主题等信息,读者可以访问 Prentice Hall 的网站: www. prenhall. com.

致谢

首先感谢我的同事们,他们贡献了自己的数据作为本书的例题和练习,完善了本书的应用. 在本书的修订过程中,很多人给了我很大帮助,也非常感谢他们,他们是 Minnesota 大学的 Christopher Bingham, Michigan 大学的 Steve Coad, Florida 大学的 Richard Kiltie, George Mason 大学的 Sam Kotz, Michigan 州立大学的 Him Koul, Drexel 大学的 Bruce McCullough, Virginia 大学的 Shyamal Peddada, Illinois 大学 Chicago 分校的 K. Sivakumar, Virginia 理工大学的 Eric Smith 以及 Illinois 大学 Urbana-Champaign 分校的 Stanley Wasserman. 我们还要感谢过去 35 年来参加多元分析课程学习的同学们,他们的反馈意见很有意义,他们的评论和建议对本书的成稿有很大帮助. 此外,特别感谢 Wai Kwong Cheang, Shanghong Guan, Jialiang Li 和 Zhiguo Xiao,他们对本书许多例题的计算给予了很大帮助.

我们还应感谢 Dianne Hall 帮忙完成了教师使用手册, Steve Verrill 帮助完成了很多计算,感谢 Alison Pollack 完善了切尔诺夫(Chernoff)脸的编程工作. 感谢 Cliff Gilman 在第 12 章中讨论多维标度例子中提供的帮助. Jacquelyn Forer 负责了本书的大部分打字工作,在此我们感谢她的专业与耐心. 最后,感谢 Petra Recter, Debbie Ryan, Michael Bell, Linda Behrens, Joanne Wendelken 以及 Prentice Hall 的相关工作人员对本项目的支持.

R. A. Johnson
rich@stat. wisc. edu
D. W. Wichern
d-wichern@tamu. edu

目 录

实用多元统计分析

Applied Multivariate Statistical Analysis

第1章 多元分析概述

1.1 引言

科学研究是一个反复学习的过程. 首先必须指定一些与某种社会现象或自然现象有关的解释作为目标,然后通过收集数据和分析数据对这些目标进行检验. 对通过实验或观察收集来的数据进行分析之后,人们通常会对现象提出一个改进的解释. 在这个反复学习的全过程中,往往有些变量会被增添到研究中去,有些则会被剔除. 因此大多数现象的复杂性要求研究人员去收集许多不同变量的观测值. 本书讨论能从这几类数据集中获取信息的各种统计方法. 由于这些数据包含许多变量的同时测量值,所以这一类方法称为**多元分析**.

人们需要了解许多变量之间的关系,这就使多元分析必然成为一个困难问题. 因为一方面人的头脑常常被一大堆数据弄得不知所措; 另一方面,供推断用的多元统计方法的推导却比在一元情形下需要更多的数学知识. 我们选择的做法是只提供基于代数概念的解释,避开需要用到多元微积分学的统计结果的推导. 我们的目标是以一种清晰的方式,利用大量说明性的例子和最低限度的数学,向读者介绍几种有用的多元方法. 不过某些数学上的复杂知识仍是需要的,也要求读者具有进行定量思考的愿望.

我们的主要侧重点在于对那些不受控制或操纵的变量所提供的测量值进行分析,只是在第6和第7两章中,我们才处理少数几个实验设计方案,以产生人们主动操纵重要变量时才会出现的数据. 尽管实验设计通常是一项科学研究中最重要的部分,但要在某学科中控制适当数据的生成通常是不可能的. (情况的确是这样,例如在商业、经济学、生态学、地质学及社会学中就是如此.)实验设计原理的详情可参考文献[6]和[7],幸运的是,这些文献的内容也适用于多元情形.

许多多元方法的基本依据是一种被称为多元正态分布的基本概率模型,这点以后将看得越来越清楚. 另一些方法就性质而言属于特殊方法,其正确性要由逻辑或常识方面的论据来证明. 无论多元方法的来源如何,都必须在计算机上实现. 计算机技术的最新进展已产生出一些相当复杂的统计软件包,从而使实现步骤变得比较容易.

多元分析是一个"混合包". 很难为多元方法建立一个分类体系,使之既能被广泛接受,又能指明这种方法的合适性. 有一种分类法可将旨在研究相互依赖关系的方法同旨在研究从属关系的方法区分开来. 另一种方法则根据所研究总体的个数及变量集的数目对多元方法进行分类. 本书根据各种处理均值的推断、协方差结构的推断以及分类或分组技术来对各章进行分节. 但决不应将这种做法看成是试图将每一种方法定位. 相反,方法的选择和分析类型的采

用很大程度上取决于研究目标. 在 1.2 节中我们列出少量实际问题,用以说明统计方法的选择与研究目标之间的联系. 这些问题加上书中的例子将对多元方法在不同领域的适用性提供正确评价.

多元方法能最自然地发挥作用之处,便是下列科学研究目标:

1. **数据简化或结构简化** 在不损失有价值信息的情况下尽可能简单地将被研究的现象描述出来. 希望这样能使解释变得容易些.

2. **分类与分组** 根据所测量的特征将一些"类似的"对象或变量分组. 另外,或许需要一些分类规则,以便将对象归入明确定义的各组.

3. **变量间依赖性的研究** 人们对变量间关系的本质感兴趣. 是否所有变量都相互独立? 还是有一个或多个变量依赖于其他变量? 如果是后者,那又是怎样依赖的?

4. **预测** 为了根据某些变量的观测值预测另一个或另一些变量的值,必须确定诸变量之间的关系.

5. **假设的构造与检验** 对以多元总体参数形式陈述的多种特殊统计假设进行检验. 这样做可以验证某些假设或增强事先建立的信念.

我们引用 F. H. C. 马里奥特(Marriott)的一段话为多元分析做一简要的综述(参见文献[19]). 这段话是在讨论聚类分析时说的,不过我们觉得它适用于更多的方法. 无论何时,在你尝试或阅读某种数据分析方法时,应该记住这一点. 它可以使你保持正确的眼光,不致被理论中某些漂亮词句所左右.

> 要是所得结果与你所获悉的看法不一致,不要接受一个简单的逻辑解释,也不要在某种图形表示中清楚显露出来,因为这些结果可能是错的. 数值方法并非魔术,招致失败的原因有很多. 它们对数据解释来说不过是有用的帮手,不是能将大量数字自动转变成一组科学事实的碎肉灌肠机.

1.2 多元方法的应用

有关多元方法应用的出版物近年来有了惊人的增加. 现在已很难像本书先前那些版本那样用简短的讨论来概括这些方法各式各样的实际应用了. 然而,为了说明多元方法的实用价值,我们对来自若干学科的研究成果作一个简要介绍. 所介绍的内容是依据前一节中给出的目标类别组织的. 当然,我们的许多例子具有多面性,可适用于一个以上的类别.

数据缩减或简化

- 用一些进行放射治疗的癌症患者的变量数据构造一个进行放射治疗的患者的简单测量法(参见练习 1.15).
- 用许多国家和地区的径赛运动记录为男女运动员建立一个成绩标准(参见文献[8]和[22]).
- 由高精度扫描仪获得的多光谱图像数据被简化成一种形式,可被看做是一个二维的海岸线的图像(参见文献[23]).
- 利用与产量及蛋白质含量有关的几个变量的数据,建立一个选择母体的标准以改善下几代豆类植物(参见文献[13]).
- 由职业仲裁人获得的数据所导出的策略相似性矩阵. 通过这个矩阵,可帮助职业仲裁人确定评价解决争端所用策略的维数(参见文献[21]).

分类与分组

- 使用与计算机用途有关的几个变量的数值，按计算机工作的类型分成几组，以便更好地决定如何利用现存的(或计划的)计算机资源(参见文献[2]).
- 一些生理学的变量的测量值可用来产生一种筛选法将酗酒者与不酗酒者区分开(参见文献[26]).
- 有关对视觉刺激的反应的数据可用来形成一条规则区分由多发性硬化症引起的视觉疾病患者与未曾患病者(参见习题1.14).
- 美国国内税务署根据从所得税申报单上收集的数据，将纳税人分为两类：需要进行审计的和不需要审计的(参见文献[31]).

变量间依赖性的研究

- 几个变量的数据被用来识别令委托人成功地雇用外来顾问的因素(参见文献[12]).
- 对于创新以及商业环境和商业组织的有关变量的测量，使我们可以发现为什么一些公司实现了产品创新，而另一些公司却没有(参见文献[3]).
- 对于纸浆纤维和用其所生产的纸的特征的测量，可被用来研究纸浆纤维和所生产出的纸张的性质之间的关系.目的是确定生产高质量纸张所需的纤维(参见文献[17]).
- 对高水平企业经理的冒险倾向的测量与社会经济学特性的测量结合起来，以评估冒险行为与个人业绩间的联系(参见文献[18]).

预测

- 利用考试得分以及几个高中成绩变量与几个大学成绩变量之间的联系，构造用来预测在大学里会成功与否的指标(参见文献[10]).
- 关于沉积物分布尺寸的几个变量数据可用来建立预报不同的沉积物的周围环境的规则(参见文献[7]和[20]).
- 利用会计财务方面的变量的测量值，可构造识别潜在的无偿还能力的财产保险公司的方法(参见文献[28]).
- cDNA微序列实验(基因表示数据)越来越广泛地用于研究癌肿瘤中的分子变化情况.肿瘤的可靠分类方法对成功的诊断和治疗癌症起着本质的作用(参见文献[9]).

假设检验

- 测量一些与污染有关的变量，以确定一个大城市地区的污染程度是在一周中大致保持不变，还是在工作日与周末之间会有明显的不同(参见练习1.6).
- 用几个变量的实验性数据可看出，教育的性质是否造成发觉风险的任何差异，由测验分数来度量(参见文献[27]).
- 利用众多变量的数据来研究美国职业结构的差别，以决定支持两个对立的社会学理论中的哪一个(参见文献[16]和[25]).
- 利用几个变量的数据来确定在新兴工业化国家中不同类型的公司是否呈现出不同的改革模式(参见文献[15]).

前面的描述使我们看到了多元方法在各种领域中的广泛应用.

1.3 数据的组织

在本书中，我们将分析由几个变量或特征构成的测量值. 这些测量值(通常称为**数据**)常常必须以不同方式排列和显示. 例如，曲线图和列表方式是数据分析的重要手段. 概括数字可定量地描绘数据的某些特征，对于任何描述都是必要的.

现在我们介绍一些初步的概念，作为数据组织的第一步.

阵列

每当一个研究者试图了解一个社会现象或自然现象时，他会选择 $p(p\geqslant 1)$ 个**变量**或**事物的特征**来进行记录，从而出现多元数据. 对每个个别的**项目**、**个体**或**实验单元**，这些变量的值都被记录下来.

我们将用记号 x_{jk} 表示第 k 个变量在第 j 项上或第 j 次试验中的观测值，即

$$x_{jk} = \text{第 } k \text{ 个变量的第 } j \text{ 项测量值}$$

因此，p 个变量的 n 个测量值可以表示如下：

	变量 1	变量 2	…	变量 k	…	变量 p
项目 1：	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
项目 2：	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
⋮	⋮	⋮		⋮		⋮
项目 j：	x_{j1}	x_{j2}	…	x_{jk}	…	x_{jp}
⋮	⋮	⋮		⋮		
项目 n：	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

或者我们可用一个有 n 行 p 列的矩形阵列来表示这些数据，称为 $\boldsymbol{X}$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

于是，阵列 $\boldsymbol{X}$ 包含了全部变量的所有观测值.

例 1.1 (一个数据阵列)

从一所大学的书店选出四张收据来了解书籍的销售情况. 每张收据提供了售书数量及每笔买卖的总金额. 用第一个变量表示总销售金额，第二个变量表示售出书的数量. 然后我们可将收据上相关数据看做是这两个变量的四组测量值. 假定数据如下所示：

变量 1(销售金额)：42　52　48　58

变量 2(书的总数)：4　5　4　3

用我们刚才引入的记号，就有

$$x_{11} = 42 \quad x_{21} = 52 \quad x_{31} = 48 \quad x_{41} = 58$$

$$x_{12} = 4 \quad x_{22} = 5 \quad x_{32} = 4 \quad x_{42} = 3$$

而数据阵列 $\boldsymbol{X}$ 由 4 行 2 列组成，即

$$\boldsymbol{X}=\begin{bmatrix}42 & 4\\ 52 & 5\\ 48 & 4\\ 58 & 3\end{bmatrix}$$ ■

以阵列格式考虑数据，简化了对问题的说明，并使数字计算以一种有序且有效的方式进行. 从以下两点得到的是双重益处：(1)以阵列运算**描述**数字计算；(2)以计算机为**实现**计算的工具，现在在计算机上用多种语言及统计程序包来进行阵列运算. 在第2章我们考虑数字阵列的变换. 目前，我们只考虑它们作为展示数据方法的价值.

描述统计量

一个大的数据集是很庞大的，而正是它的庞大严重地干扰了任何从中提取适当信息的企图. 包含在数据中的许多信息，可以通过计算某些通称为**描述**统计量的概括数字进行估计. 例如，算术平均值，或样本均值，是一种描述统计量，它提供了一种定位测量——即一个数集的“中心值”. 而从均值到所有数的距离的平方的平均值提供出用这些数的分布程度或变差.

我们将尽量依靠描述统计量来测量定位、变差以及线性结合. 这些量的正式定义如下：

设 $x_{11},x_{21},\cdots,x_{n1}$ 是第一个变量的 n 个测量值，则这些测量值的算术平均值是

$$\bar{x}_1=\frac{1}{n}\sum_{j=1}^{n}x_{j1}$$

如果这 n 个测量值代表被观测的全部测量值集合的一个子集，则 $\bar{x}_1$ 也称为第一个变量的**样本均值**. 我们采用这个术语是由于本书的大部分内容是用于研究为了分析来自较大聚集的测量值样本而设计的计算方法.

样本均值可用 p 个变量中的每一个变量的 n 个测量值计算出来，因此，一般的，将有 p 个样本均值：

$$\bar{x}_k=\frac{1}{n}\sum_{j=1}^{n}x_{jk}\qquad k=1,2,\cdots,p \tag{1-1}$$

分布的程度由**样本方差**给出，对第一个变量的 n 个观测值定义为

$$s_1^2=\frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2$$

其中 $\bar{x}_1$ 是 x_{j1} 的样本均值. 一般的，对于 p 个变量，我们有

$$s_k^2=\frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2\qquad k=1,2,\cdots,p \tag{1-2}$$

有两点说明. 首先许多作者定义样本方差时，用 $n-1$ 而不是 n 作为除数. 后面我们将看到这样做是有理论上的理由的，并且当测量值的数目 n 较小时，这样做尤其适当. 样本方差的两种形式总能用适当的式子区分开来.

其次，尽管记号 s^2 是样本方差传统的表示方法，我们仍将考虑一个样本方差位于主对角线的数量阵列. 在这种情况下，为了表明方差在阵列中的位置，使用双下标是比较方便的. 因此，引入记号 s_{ii} 来表示由第 i 个变量的测量值计算出的方差，并有式子

$$s_k^2=s_{kk}=\frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2\qquad k=1,2,\cdots,p \tag{1-3}$$

样本方差的平方根 $\sqrt{s_{ii}}$，称为**样本标准差**. 这个变差的度量单位和观测值的单位相同.

考虑变量1和2的 n 对测量值：

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \cdots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$$

其中 x_{j1} 和 x_{j2} 是由第 j 次试验项观测的($j=1,2,\cdots,n$). 变量 1 和 2 的测量值的线性结合由**样本协方差**给出

$$s_{12} = \frac{1}{n}\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)(x_{j2}-\bar{x}_2)$$

或是它们各自的平均值的偏差之积的平均值. 若对一个变量的大观测值与对另一变量的大观测值一起出现,而小值也一起出现,s_{12} 将为正数. 若一个变量的大值与另一变量的小值一起出现,s_{12} 将为负数. 若两个变量的值间没有什么特别的联系,s_{12} 将近似等于零.

样本协方差

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k) \qquad i=1,2,\cdots,p,k=1,2,\cdots,p \tag{1-4}$$

度量第 i 个和第 k 个变量间的结合. 我们注意到,当 $i=k$ 时,协方差就简化为样本方差. 此外,对所有的 i 和 k 都有 $s_{ik}=s_{ki}$.

在此,描述统计量最后一个要考虑的是**样本相关系数**(或**皮尔逊积矩相关系数**;参见文献[14]). 两个变量间线性结合的这个度量不依赖于测量单位,第 i 个和第 k 个变量的样本相关系数定义为

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2}} \tag{1-5}$$

其中 $i=1,2,\cdots,p$ 和 $k=1,2,\cdots,p$. 注意对所有 i,k 都有 $r_{ik}=r_{ki}$.

样本相关系数是样本协方差的一个标准化形式,其中样本方差的平方根的乘积提供了标准化. 请注意,对于 s_{ii},s_{kk} 和 s_{ik} 不论是用 n 或 $n-1$ 作为除数,r_{ik} 的值都是相同的.

样本相关系数 r_{ik} 也可看做是一个样本**协方差**. 假设初始值 x_{ji},x_{jk} 由**标准化**值 $(x_{ji}-\bar{x}_i)/\sqrt{s_{ii}}$ 和 $(x_{jk}-\bar{x}_k)/\sqrt{s_{kk}}$ 取代. 标准化值是可通约的,因为两个集合都是中心在零,并用单位标准差表达. 样本相关系数正是标准观测值的样本协方差.

尽管样本相关系数与样本协方差的记号相同,通常相关较容易解释,因为它的大小是有界的. 概括地讲,样本相关系数 r 有以下性质:

1. r 的值必定在 -1 与 $+1$ 之间.

2. 这里 r 度量的是线性结合的强度. 如果 $r=0$,这就意味着分量之间无线性结合. 另外,r 的正负号指出了结合的方向;$r<0$ 意味着一个趋势,即当一对数中一个值大于它的平均值,另一个值则小于它的平均值;$r>0$ 表示这样的趋势,即当一对数中一个值大时另一个值也大,或者两值一起小.

3. 若第 i 个变量的测量值变为 $y_{ji}=ax_{ji}+b, j=1,2,\cdots,n$,且第 k 个变量的测量值变为 $y_{jk}=cx_{jk}+d, j=1,2,\cdots,n$,假定常数 a 和 c 的正负号相同,则 r_{ik} 的值保持不变.

参量 s_{ik} 和 r_{ik} 一般不能传送有关两个变量间结合的全部信息. 可能存在不被描述统计量揭示的**非线性**结合. 协方差和相关系数提供了线性结合或沿直线结合的度量. 它们的值对其他类型的结合不能给出太多信息. 另一方面,这些参量对于"杂乱的"观测值("离群值")非常敏感,可能会表示出事实上几乎不存在的结合. 尽管有这些缺点,协方差和相关系数仍常常被