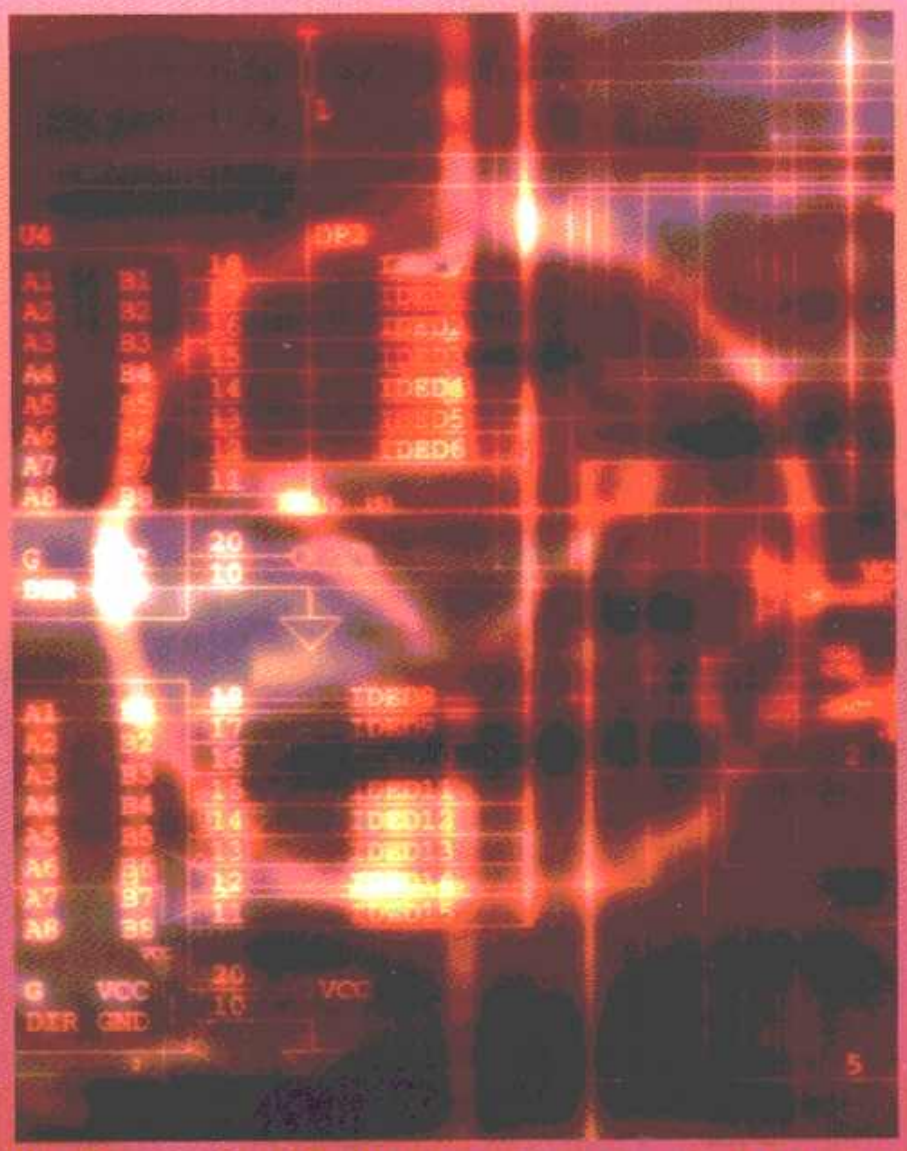
国外工商管理硕士(MBA)优秀教材译丛

Applied Multivariate Statistical Analysis

实用多元统计分析

（第四版）

[美] Richard A. Johnson
Dean W. Wichern 著 陆璇 译



清华大学出版社 · PRENTICE HALL
http://www.tup.tsinghua.edu.cn http://www.phregents.com

国外工商管理硕士(MBA)优秀教材译丛

Applied Multivariate Statistical Analysis

实用多元统计分析

（第四版）

Richard A. Johnson
Dean W. Wichern 著

陆 璇 译

清 华 大 学 出 版 社

(京)新登字 158 号

内 容 简 介

本书为多元统计分析领域内的经典著作,对社会科学和自然科学的许多学科中常要运用多元统计方法来分析数据的研究者是一本很好的专业参考书,同时也可以作为高等院校研究生学习应用统计类课程的教学参考书。本书的内容十分丰富,涵盖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,分为四大部分:

第一部分:预备知识,其中包括多元分析概述,矩阵代数与随机向量,样本几何与随机抽样,多元正态分布;

第二部分:关于多元均值与线性模型的推断,其中包括关于均值向量的推断,多个多元均值向量的比较,多元线性回归模型;

第三部分:协方差结构分析,其中包括主成分分析,因子分析与对协方差矩阵结构的推断,典型相关分析;

第四部分:分类和分组方法,其中包括判别与分类,聚类、距离方法与多维标度变换。

北京市版权局著作权合同登记号:01-98-1482

书　名:实用多元统计分析(第四版)
作　者:Richard A. Johnson, Dean W. Wichern　著　陆　璇　译
出版者:清华大学出版社(北京清华大学学研大厦,邮编 100084)
http://www.tup.tsinghua.edu.cn
印刷者:北京市清华园胶印厂
发行者:新华书店总店北京发行所
开　本:787×1092 1/16　**印张**:38.5　**字数**:889 千字
版　次:2001 年 4 月第 4 版　2001 年 4 月第 1 次印刷
书　号:ISBN 7-302-04347-7/F・313
印　数:0001～4000
定　价:69.50 元(含盘)

第一部分

实用多元统计分析

预备知识

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。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[7]和[8]，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。下面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)[19]的一段话为多元分析做一简要的综述。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[10]和[21]。)
- 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，可被看作是一个二维的海岸线的图像。(参见[22]。)
- 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[14]。)

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[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[25]。)
- 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[30]。)

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[13]。)
- 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[5]。)
- 表示 10 次奥运会上十项全能运动的结果的变量的数据,被用来确定担负十项全能项目取得成功的身体因素。(参见[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[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[11]。)
- 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[9]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[27]。)
- 利用关于繁缕苗的几个变量数据,构造一种方法用来预报新苗的种类。(参见[4]。)

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,教育的本质是否造成发觉风险的任何差异,由测验分数来度量。(参见[26]。)
- 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[16]和[24]。)

- 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[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,我们将涉及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 $\mathbf{X}$

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

于是,阵列 $\mathbf{X}$ 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量。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作是这两个变量的四组测量值。假定数据如下所示:

变量 1(销售美元):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 **X** 由 4 行 2 列组成,即

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

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。在这方面,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,一般地,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。一般地,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。

第二，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[3]）。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} 也可看作是一个样本**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。标准化值是可公度的，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，而另一个值小于它的平均值；$r>0$ 则表示这样的趋势，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_{rk} 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用于计算与分析。当数据没有显示出明显的非线性结合模式且没有出现杂乱的观测值时，它们能提供有说服力的结合的数值概括。

必须考虑可疑的观测值，以便改正明显的记录错误并根据所发现的原因而采取相应措施。s_{ik} 和 r_{ik} 的值应该同时用它们的观测值或同时不用这些观测值来引证。

来自均值的偏差的平方和与偏差的交叉乘积的和常常是有用的。这些参量是：

$$w_{kk}=\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2 \qquad k=1,2,\cdots,p \tag{1-6}$$

和

$$w_{ik}=\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k) \qquad i=1,2,\cdots,p,\quad k=1,2,\cdots,p \tag{1-7}$$

由 p 个变量的 n 组测量值计算出的描述统计量也可用阵列来构成。

基本的描述统计量阵列

$$\text{样本均值}\quad \bar{\mathbf{x}}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix}$$

$$\text{样本方差和协方差}\quad \mathbf{S}_n=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1p}\\ s_{21} & s_{22} & \cdots & s_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ s_{p1} & s_{p2} & \cdots & s_{pp}\end{bmatrix} \tag{1-8}$$

$$\text{样本相关系数}\quad \mathbf{R}=\begin{bmatrix}1 & r_{12} & \cdots & r_{1p}\\ r_{21} & 1 & \cdots & r_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ r_{p1} & r_{p2} & \cdots & 1\end{bmatrix}$$

样本均值阵列由 $\bar{\mathbf{x}}$ 表示，样本方差和协方差阵列用大写字母 $\mathbf{S}_n$ 表示，而样本相关系数阵列用 **R** 表示。在阵列 $\mathbf{S}_n$ 中下标 n 作为助记符用来提醒你 n 被当作元素 s_{ik} 的除数。所有阵列的大小取决于变量的个数 p。

阵列 $\mathbf{S}_n$ 和 **R** 由 p 行 p 列组成。$\bar{\mathbf{x}}$ 是个只有 1 列 p 行的阵列。阵列 $\mathbf{S}_n$ 和 **R** 中每个元素的第 1 个下标表示行，第 2 个下标表示列。由于 $s_{ik}=s_{ki}$ 和 $r_{ik}=r_{ki}$ 对所有的 i,k 成立，阵列 $\mathbf{S}_n$ 和 **R** 中与主对角线对称位置上的元素是相同的，并且这样的阵列称为是**对称**的。

例 1.2 （关于二元数据的阵列 $\bar{\mathbf{x}}$，$\mathbf{S}_n$ 和 **R**）

考虑例 1.1 中所用数据。每张收据产生一对测量值，总销售金额和售书总数。求阵列 $\bar{\mathbf{x}}$，$\mathbf{S}_n$ 和 **R**。

由于有 4 张收据，共得到每个变量的 4 个测量值。

样本均值是

$$\bar{x}_1=\frac{1}{4}\sum_{j=1}^{4}x_{j1}=\frac{1}{4}\times(42+52+48+58)=50$$

$$\bar{x}_2=\frac{1}{4}\sum_{j=1}^{4}x_{j2}=\frac{1}{4}\times(4+5+4+3)=4$$

$$\bar{\mathbf{x}}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\end{bmatrix}=\begin{bmatrix}50\\ 4\end{bmatrix}$$

样本方差和协方差是

$$s_{11}=\frac{1}{4}\sum_{j=1}^{4}(x_{j1}-\bar{x}_1)^2$$

$$=\frac{1}{4}\times[(42-50)^2+(52-50)^2+(48-50)^2+(58-50)^2]=34$$

$$s_{22}=\frac{1}{4}\sum_{j=1}^{4}(x_{j2.}-\bar{x}_2)^2$$

$$=\frac{1}{4}\times[(4-4)^2+(5-4)^2+(4-4)^2+(3-4)^2]=0.5$$

$$s_{12}=\frac{1}{4}\sum_{j=1}^{4}(x_{j1}-\bar{x}_1)(x_{j2}-\bar{x}_2)$$

$$=\frac{1}{4}\times[(42-50)(4-4)+(52-50)(5-4)$$

$$+(48-50)(4-4)+(58-50)(3-4)]=-1.5$$

$$s_{21}=s_{12}$$

和

$$\mathbf{S}_n=\begin{bmatrix}34 & -1.5\\ -1.5 & 0.5\end{bmatrix}$$

样本相关系数是

$$r_{12}=\frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}}=\frac{-1.5}{\sqrt{34}\times\sqrt{0.5}}=-0.36$$

$$r_{21}=r_{12}$$

故

$$\mathbf{R}=\begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

图解法

作图是进行数据分析的重要的辅助手段，但常常被人们忽略。尽管不可能同时对几个变量的所有测量值作图并研究它们的形状，但单个变量的图形和一对变量的图形仍可提供很多信息。高级计算机程序和显示设备使我们可以相对容易地从一维、二维或三维空间轻松而直观地检查数据。另一方面，利用纸和笔绘图，也可从数据中领悟许多有价值的东西。表示数据的简单、优美且有效的方法可以从[28]得到。为一对对的变量作图并直观地考察连结的模式，这是一种很好的统计学做法。考虑以下两个变量的 7 对观测值：

变量 1(x_1)：3　4　2　6　8　2　5

变量 2(x_2)：5　5.5　4　7　10　5　7.5

这些数据在二维平面上用 7 个点表示(每个坐标轴代表一个变量)，如图 1.1 所示。每个点的坐标由一对测量值确定：(3,5)，(4,5.5)，…，(5,7.5)，所得到的二维图形称为**散布图表**或**散布图**。

图 1.1 还分别显示了由变量 1 和变量 2 的观测值单独作的散点图。这种图称为(**边缘**)**点图**。它们可以从原始观测值直接得到，也可将点投影到散布图的每个坐标轴上得到。

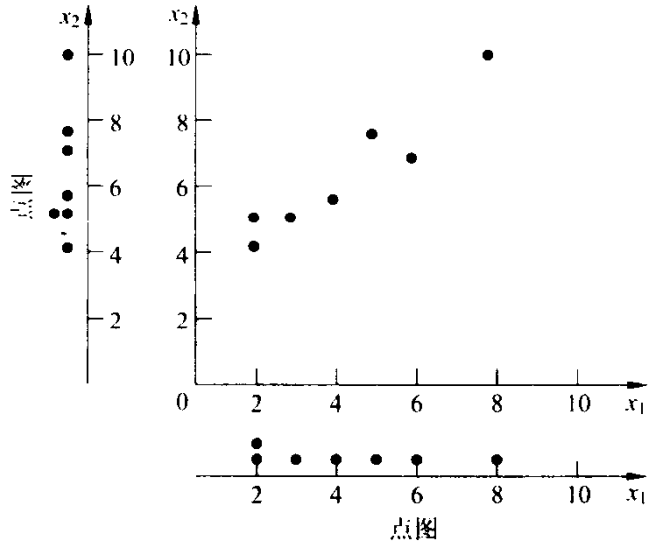


图 1.1　散布图和边缘点图

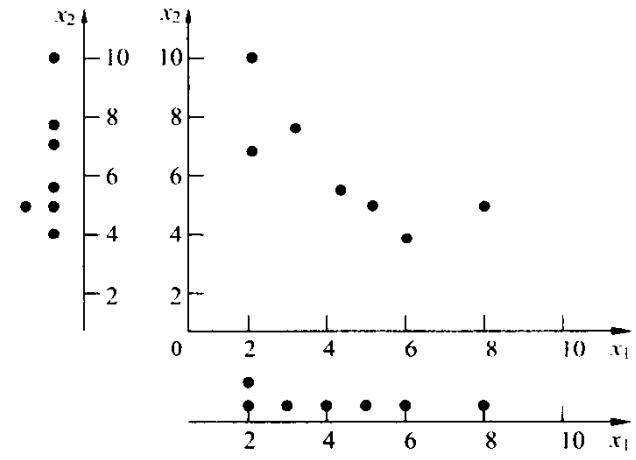


图 1.2　数据重新排列后的散布图和点图

单变量点图中所包含的信息可被用来计算样本均值 $\bar{x}_1$ 和 $\bar{x}_2$ 及样本方差 s_{11}和 s_{22}(参考练习 1.1。)散布图表明了点的方位，这些点的坐标可用来计算样本协方差 s_{12}。在图 1.1 的散布图中，x_1 的大值与 x_2 的大值一起出现，而 x_1 的小值又与 x_2 的小值一起出现。因此，s_{12}将是正的。

点图与散布图包含了不同类型的信息。边缘点图中所含信息不足以构造出散布图。作为一个说明，假设对前面图 1.1 中的数据重新配对之后，使得变量 x_1 和 x_2 的测量值如下：

变量 1 (x_1)：　5　4　6　2　2　8　3

变量 2 (x_2)：　5　5.5　4　7　10　5　7.5

(我们已简单地重新排列了变量 1 的数值。)“新”数据的散布图和点图在图 1.2 中。比较图 1.1 和图 1.2，我们发现边缘点图是相同的，而散布图却明显不同了。在图 1.2 中，x_1 的大值与 x_2 的小值一起出现，而 x_1 的小值又与 x_2 的大值同时出现。因此，描述统计量中单独的变量 $\bar{x}_1$，$\bar{x}_2$，s_{11}和 s_{22}保持不变，而度量一对变量间结合的样本协方差 s_{12}现在却变为负数。

图 1.1 和图 1.2 中数据的不同取向无法通过单独的边缘点图辨别。然而，在两种情况下边缘点图是相同的，这个事实也无法立即从散布图中显示出来。这两种图解法不是对立的，而是相互补充的。

接下来两个例子可进一步说明通过图形表示可以传递的信息。

例 1.3　(异常的观测值对样本相关系数的影响)

在 1990 年 4 月 30 日的《福布斯》杂志上的一篇文章中，列出了 16 家最大的出版公司的某些财务数据，以体现其工作情况与生产率。关于 x_1＝雇员(工作)和 x_2＝每个雇员创的利润(生产率)这一对变量的数据如图 1.3 所示。我们已标明两个独特的观测值。邓·布拉德街公司从雇员人数讲是最大的公司，但从每个雇员所创利润看是很“典型的”。时代华纳公司在雇员人数方面是很“典型的”，而在人均创利润上则相当的少(负值)。

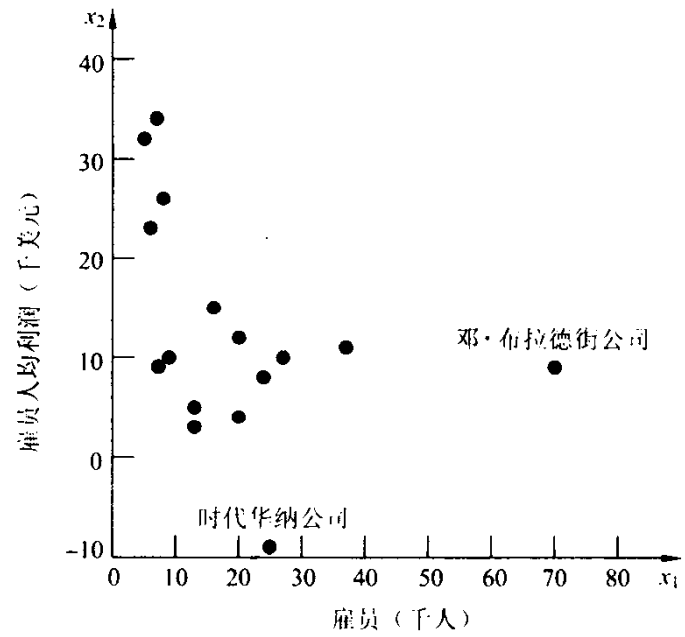


图 1.3　16 家出版公司的雇员人均利润和雇员人数

由 x_1 和 x_2 的值算出的样本相关系数是

$$r_{12}=\begin{cases}-0.39 & \text{对所有 16 个公司}\\ -0.56 & \text{对除邓·布拉德街公司之外的所有公司}\\ -0.39 & \text{对除时代华纳公司之外的所有公司}\\ -0.50 & \text{对除邓·布拉德街公司和时代华纳公司之外的所有公司}\end{cases}$$

显然异常的观测值对于样本相关系数有着相当大的影响。

例 1.4　(棒球数据的散布图)

1978 年 7 月 17 日，Sports Illustrated 杂志上一篇关于体育界工资的文章，为 x_1＝国家东部联盟棒球队为球员发放的工资总额提供了一些数据。

我们增加了 x_2＝1977 年的胜负比的数据。数据由表 1.1 给出。

表 1.1　关于国家东部联盟棒球队 1977 年工资和最终记录

队	x_1＝运动员工资	x_2＝胜负比
Philadelphia Phillies	3 497 900	0.623
Pittsburgh Pirates	2 485 475	0.593
St. Louis Cardinals	1 782 875	0.512
Chicago Cubs	1 725 450	0.500
Montreal Expos	1 645 575	0.463
New York Mets	1 469 800	0.395

图 1.4 的散布图支持了一种看法：冠军是可以买来的。当然，这种因果关系是无法证实的，因为实验不包括随机的工资总额的分配。因此，统计量无法回答这个问题：如果 Mets 队有 400 万美元来支付运动员的薪水，是否就会赢？

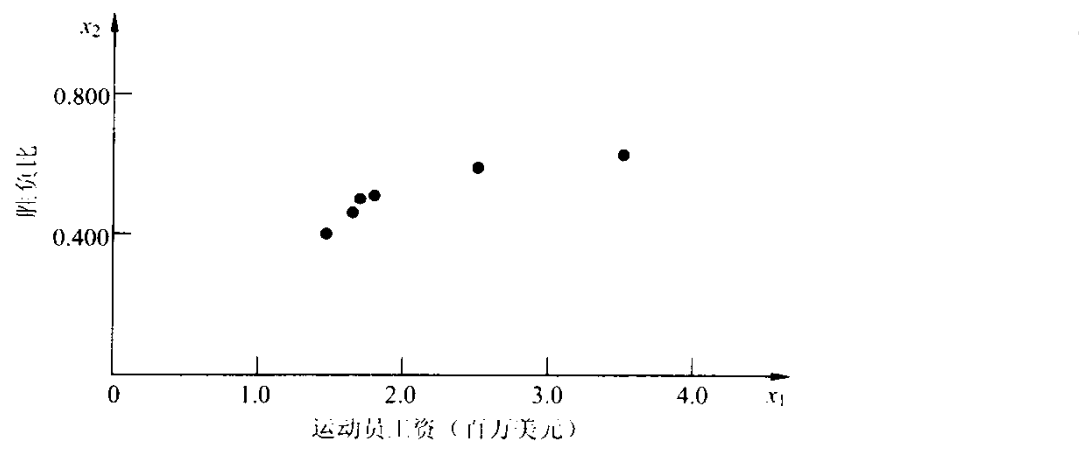


图 1.4　从表 1.1 得到的工资和胜负比

为了构造散布图，例如图 1.4，我们把表 1.1 中的 6 对观测值看作是二维空间中 6 个点的坐标。此图让我们直观地检查相对于工资总额和胜负比的分组方法。

例 1.5　（关于纸张强度观测结果的多重散布图）

纸是由几英尺宽的连续薄片制成的。由于纸中纤维的方向性，沿着机器造纸的方向测量的纸的强度，不同于沿着与机器造纸方向交叉或成直角的方向测量的结果。表 1.2 列出了下列变量的测量值。

x_1 ＝ 密度（克 / 立方厘米）

x_2 ＝ 强度（磅）：沿机器方向

x_3 ＝ 强度（磅）：沿交叉方向

一种新颖的图形用来表示这些数据，见图 1.5。散布图作为协方差阵列的非对角线元素，而盒形图作为对角线元素。用这一方法后者有不同的标度。故我们只用总的形状对称性给出信息，并为每个单独的特性提供可能的离群值。可就图形和异常的观测值来观察散布图。在图 1.5 中，有一个异常的观测值：样本 25 的密度。某些散布图表明，其观测值是相互分离的两簇。阵列的这些散布图将在下一节中新的图解方法的讨论中还会涉及。

在一般的多种响应的情况下，p 个变量同时进行 n 项记录。应该为几对重要的变量作散布图，并且如果工作量没有大到得不偿失，应该对每一对变量作图。

表 1.2　纸张性能测量结果

标本	密度	强度	
		机器方向	交叉方向
1	0.801	121.41	70.42
2	0.824	127.70	72.47
3	0.841	129.20	78.20
4	0.816	131.80	74.89
5	0.840	135.10	71.21
6	0.842	131.50	78.39
7	0.820	126.70	69.02
8	0.802	115.10	73.10
9	0.828	130.80	79.28
10	0.819	124.60	76.48
11	0.826	118.31	70.25
12	0.802	114.20	72.88
13	0.810	120.30	68.23
14	0.802	115.70	68.12
15	0.832	117.51	71.62
16	0.796	109.81	53.10
17	0.759	109.10	50.85
18	0.770	115.10	51.68
19	0.759	118.31	50.60
20	0.772	112.60	53.51
21	0.806	116.20	56.53
22	0.803	118.00	70.70
23	0.845	131.00	74.35
24	0.822	125.70	68.29
25	**0.971**	**126.10**	**72.10**
26	0.816	125.80	70.64
27	0.836	125.50	76.33
28	0.815	127.80	76.75
29	0.822	130.50	80.33
30	0.822	127.90	75.68
31	0.843	123.90	78.54
32	0.824	124.10	71.91
33	0.788	120.80	68.22
34	0.782	107.40	54.42
35	0.795	120.70	70.41
36	0.805	121.91	73.68
37	0.836	122.31	74.93
38	0.788	110.60	53.52
39	0.772	103.51	48.93
40	0.776	110.71	53.67
41	0.758	113.80	52.42

资料来源:数据经 SONOCO 制造公司许可使用。