多元统计分析与SPSS应用

汪冬华 ◎编著

華東理工大學出版社
EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS

多元统计分析与SPSS应用

汪冬华 编著

華東理工大學出版社
EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS

图书在版编目(CIP)数据

多元统计分析与 SPSS 应用/汪冬华编著. —上海：华东理工大学出版社，2010.9
ISBN 978-7-5628-2874-7

Ⅰ.①多… Ⅱ.①汪… Ⅲ.①多元分析：统计分析-软件包，SPSS 12.0 Ⅳ.①O212.4

中国版本图书馆 CIP 数据核字(2010)第 154303 号

多元统计分析与 SPSS 应用

编　　著/汪冬华
责任编辑/马夫娇
责任校对/张　波
封面设计/陆丽君
出版发行/华东理工大学出版社
　　社　址：上海市梅陇路 130 号，200237
　　电　话：(021)64250306(营销部)　64251137(编辑部)
　　传　真：(021)64252707
　　网　址：press.ecust.edu.cn
印　　刷/上海展强印刷有限公司
开　　本/787 mm×1092 mm　1/16
印　　张/19.5
字　　数/521 千字
版　　次/2010 年 9 月第 1 版
印　　次/2010 年 9 月第 1 次
印　　数/1-3 000 册
书　　号/ISBN 978-7-5628-2874-7/F·228
定　　价/38.00 元

前　言

随着科技进步和社会发展，在工业、经济、农业、生物和医学等领域的实际问题中，需要处理多个变量的观测数据，以及研究多个随机变量之间的相互依赖关系和内在统计规律性。因此，对多个变量进行综合处理的多元统计分析(multivariate statistical analysis)方法显得尤为重要。随着电子计算机技术的普及，以及社会、经济和科学技术的发展，过去被认为具有数学难度的多元统计分析方法，已越来越广泛地成为管理学、经济学、生物学、人口学、社会学等学科分析、处理多维数据不可缺少的重要工具。

多元统计分析是从经典统计学中发展起来的一个分支，是一种综合分析方法，应用很广泛。然而，现已出版的多元统计分析的相关教材和著作，多数侧重于数理推导和证明，关于数学方法在实践中的应用介绍较少，且案例偏重于自然科学，适合经济管理类专业学生学习的教材较少。基于此，作者总结多年从事经济管理类专业的多元统计分析的教学经验，结合学生实际的学习特点和需求，编著了本书。

本书的特点有以下几点。

1. 加强基本原理和基本方法的理解。面对枯燥的数学理论，本书侧重于在实际案例解决分析过程中，加强对多元统计分析的基本原理和基本方法的理解。

2. 加强多元统计分析方法在实际经济管理问题中的应用。本书在介绍完基本方法后，通过利用多元统计分析的方法解决实际经济管理的案例，强调方法的应用和解决问题的能力。

3. 加强 SPSS 在多元统计分析中的应用。为了提高读者的多元统计分析理论方法的实践应用能力和可操作性，本书强调依据多元统计方法利用 SPSS 现代统计软件对实际案例进行数据处理和统计分析，并在每章结合实例概要介绍了 SPSS 软件的实际操作和实现过程。

全书共十三章，主要内容包括：多元描述统计分析、均值的比较检验、方差分析、正交试验设计、相关分析、回归分析、聚类分析、判别分析、主成分分析、因子分析、对应分析、典型相关分析和定性数据的统计分析等。

本书是华东理工大学校级精品课程“应用统计学”的建设成果之一，由华东理工大学商学院金融学系汪冬华组织编写，教材编写大纲和写作要求是经过本书全体作者多次讨论而定，最

后由汪冬华统一定稿。参与编写的人员主要有：汪冬华(第1、2、3、4、7、8章)，马艳梅(第5、9、10、11、12、13章)，任飞(第6章)。本书出版的动力一部分来自商学院金融学系朱邦毅老师不断的鞭策和帮助，在此表示感谢。同时需感谢吴雅婷、黄康等研究生的辛勤工作，感谢华东理工大学教务处以及华东理工大学出版社，本书受到他们的大力资助。感谢商学院金融学系刘建国教授为本书的出版所提供的帮助和支持。

本书可作为经济与管理类专业本科生统计分析课程的教材，也可作为研究生和MBA的教材或参考书，同时也适合作为从事社会、经济、管理等研究和实际工作的从业人员进行数据分析的参考书。

本书参考了国内外大量的相关书籍和文献，由于篇幅有限未能一一列出，谨向这些作者表示感谢。本书也是作者长期教学和研究的经验结晶，由于作者水平有限，疏漏之处在所难免，恳请读者批评指正，以便于再版修订时，不断完善。

目　录

·第1章·
多元描述统计分析

相关实例

➢ 在管理学理论中，企业文化是企业的灵魂，是推动企业发展的重要因素，是多维的、多层次的。国内外很多学者对此进行了大量的定性、定量研究，提出了自己的观点和不同的文化测度模型。如荷兰学者霍夫斯坦特(Hofstede)从管理心理学的角度来研究企业文化，提出了权力距离、风险规避、个人主义倾向和对抗性四个维度；美国学者奎因(Quinn)和卡迈隆(Cameron)发现组织中的主导文化、领导风格、管理角色、人力资源管理、质量管理及成功的判断准则等因素共同构成了企业文化的测度；美国学者德尼森(Denison)构筑了一个四维文化测度模型，由适应性、使命、一致性和投入四个文化特质构成，其中每个文化特质都对应着三个子维度。在国内，以清华大学张德教授为首的中国学者结合东方的文化特征，提出了中国特色的文化测度模型，包括领导风格、能力绩效导向、人际和谐、科学求真、凝聚力、正直诚信、顾客导向、卓越创新、组织学习、使命与战略、团队精神、发展意识、社会责任、文化认同等十四个因素。

➢ 在经济学中，要研究一些经济问题，往往需要综合大量的经济数据，形成各种经济指数进行分析。如研究一个地区的经济发展水平，就需要分析该地区的生产总值、工农业产值、税收、居民收入、居民消费指数、商品价格指数、生活费用指数等。

上述例子中，我们都需要用多个观察变量来描述一些现实生活中的社会经济现象，而这些现象往往都具有多维性，需要用多个指标进行测量和分析。在本章中我们就要学习多元数据的描述方法，以便更直观地观察数据之间的关系。

1.1 多元描述统计量

1.1.1 数据的组织

在研究各种生产生活或者经济管理现象时，我们会搜集多个变量的测量值，形成多元数

据，然后从这几类数据中获取信息。这些测量值以不同的方式排列和显示，可以比较清晰地描绘数据的某些特征。

我们选择 p 个变量来记录事物的特征，对于每个个体或单位，记录下这些变量的测量值。我们用记号 x_{ij} 表示第 i 个样本上第 j 个变量的测量值，即

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

因此，p 个变量的 n 个测量值就可以表示如下。

表 1.1 数 据 表

	变量 1	变量 2	…	变量 j	…	变量 p
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
⋮	⋮	⋮	…	⋮	…	⋮
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
⋮	⋮	⋮	…	⋮	…	⋮
n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

表 1.1 中，第 i 行表示第 i 个样本 p 个变量的测量值，第 j 列表示第 j 个变量在各个样本中的测量值。

我们也可以用一个 n 行 p 列的矩阵列来表示这些数据，记为 $\boldsymbol{X}$。

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

矩阵 $\boldsymbol{X}$ 包含了全部变量的所有测量值。

例 1.1 消费者物价指数(CPI)是反映与居民生活有关的产品及劳务价格统计出来的物价变动指标，通常作为观察通货膨胀水平的重要指标。商品零售价格指数是反映一定时期内商品零售价格变动趋势和程度的相对数。两者都能为研究市场流通、进行国民经济核算提供依据。表 1.2 为某地区四个主要城市的消费者物价指数和商品零售价格指数。

表 1.2 某地区主要城市的消费者物价指数和商品零售价格指数(上年＝100)

	消费者物价指数(CPI)	商品零售价格指数
A 市	106.3	104.8
B 市	102.5	101.4
C 市	103.2	102.5
D 市	105.8	105.3

引入上述定义，就有

$$\boldsymbol{X}=\begin{bmatrix}106.3 & 104.8\\ 102.5 & 101.4\\ 103.2 & 102.5\\ 105.8 & 105.3\end{bmatrix}$$

用矩阵的形式来表示多元数据，是一种有序且有效的方法，简化了对问题的说明，有利于数据的变换和处理。

1.1.2 描述统计量

在现实生活中，诸多社会、经济等实际问题往往都是很复杂的，我们通过抽样调查等方式获得大量庞杂的数据，而这些数据中包含了许多信息，不能直观地表现出来。为了从这些数据中提取有效的信息，可以通过计算一些通称为描述统计量的概括性数字来对样本数据进行分析，进而推断总体特征。

常用的描述统计量有样本均值、样本协方差、样本相关系数等。

1. 样本均值

样本均值是反映样本数据集中趋势的统计量，是对单个变量样本数据取值一般水平的描述。

设 x_{11}，x_{21}，…，x_{n1} 是变量 1 的 n 个测量值，则这些测量值的算术平均值为

$$\bar{x}_1=\frac{1}{n}\sum_{i=1}^{n}x_{i1}$$

$\bar{x}_1$ 就称为变量 1 的样本均值。

在多元统计中，一般存在多个变量，因此可计算出 p 个变量的样本均值

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}\quad j=1,2,\cdots,p$$

样本均值可用矩阵的形式表示为

$$\bar{\boldsymbol{x}}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix}$$

2. 样本协方差

样本协方差是反映数据离散趋势的统计量，协方差分析是利用线性回归的方法消除混杂因素的影响后进行的方差分析，其功能就是消除方差分析中不可控因素的影响。样本数据的分布程度即可由样本协方差来描述。

样本方差

变量 1 的样本方差可表示为

$$s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2$$

式中，$\bar{x}_1$ 为变量 1 的样本均值。对于 p 个变量，其样本方差为

$$s_j^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\quad j=1,2,\cdots,p$$

由于在样本协方差矩阵中，各变量的样本方差位于矩阵的主对角线上，为了方便表达，我

们使用双下标来标记样本方差，因此引入记号 s_{kk} 来表示 s_j^2，即

$$s_{kk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2 \quad k = 1, 2, \cdots, p$$

样本协方差

p 个变量中，任意两个变量：变量 j 和变量 k 之间的协方差为

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j), \quad j = 1, 2, \cdots, p, k = 1, 2, \cdots, p$$

我们可以发现，当 $i = k$ 时，样本协方差就等于样本方差。此外，对于所有的 j 和 k，都有 $s_{jk} = s_{kj}$。

用矩阵形式来表示样本协方差

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

3. 样本相关系数

样本相关系数，又称皮尔逊(Pearson)积距相关系数，是样本协方差的标准化形式，反映两个现象之间相关关系密切程度。

样本相关系数一般用 r 表示。定义变量 j 和变量 k 的样本相关系数为

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

式中，$i = 1, 2, \cdots, p$；$k = 1, 2, \cdots, p$。此外，对于所有的 j 和 k，都有 $r_{jk} = r_{kj}$。

用矩阵形式来表示样本相关系数

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

关于样本相关系数，有以下几点性质。

① r 的值必在 -1 与 $+1$ 之间；

② r 表示两个变量之间的相关程度，r 的绝对值越大，相关程度越高：$r = 1$，完全正相关；$r = -1$，完全负相关；$r = 0$，不相关；$0 < r < 1$，正相关；$-1 < r < 0$，负相关。

1.2 多元数据的图形表示

利用图形的方法来表现多元数据是进行数据分析的重要辅助手段。现在，高级计算机软件的

发展取代了用纸和笔作图的传统方法，可以方便快捷地绘制出各种统计图表，清晰直观地展现数据的特征和关系，帮助我们从数据中提取信息进行处理和分析。正如俗话说的那样，一图胜千言。

在多元统计中有很多种不同的图形分析方法。根据图的维数不同，可以分为一维图、二维图、三维图等；根据图的形状不同，有直方图、饼图、散点图、箱线图、茎叶图、雷达图、脸谱图等。现在，常见的统计分析软件也有很多，常用的有 Excel，SPSS，SAS，Matlab，Eviews，Stata 等，这些纷繁多样的软件也给我们提供了更多不同的方法来进行统计研究，在本书中我们主要介绍如何用 SPSS 来进行统计分析。

1.2.1 散点图

散点图，又称为散布图或相关图，是直观反映变量间相关关系的一种统计图形。与其他统计图相比，散点图更能表现数据的原始分布情况。从散点图中，可以根据点的位置来判断测量值的大小、变动趋势和变动范围，从而深入了解变量间的关系。

我们使用得比较多的是二维的简单散点图，它是将二维平面上的数据用点在坐标中表示绘制而得的。其中每个坐标轴代表一个变量，每个测量值的坐标确定一个点。这样得到的散点图可以直观地表示出两个变量之间的相关关系，便于我们观察数据间的相关性，剔除异常数据，提高准确性。

更复杂一点的散点图是在简单散点图的基础上的扩展，用同样的方法，我们可以将二维散点图扩展到三维。而对于多个变量的问题，我们则可以用矩阵散点图来解决。

1. 简单散点图

例 1.2　我们以 SPSS 中自带的数据文件 employee data. sav 作为例子。该文件是某商业银行员工有关基本情况的数据。文件包含了 474 名员工的员工编号(id)、性别(gender)、出生日期(bdate)、受教育年限(educ)、工作类别(jobcat)、工资水平(salary)、初始工资水平(salbegin)、工作时间(jobtime)、来银行以前时间(prevexp)、是否少数民族(minority)等信息。

这里我们选取受教育年限和工资水平两个变量，用 SPSS 绘制成简单散点图，如图 1.1 所示。

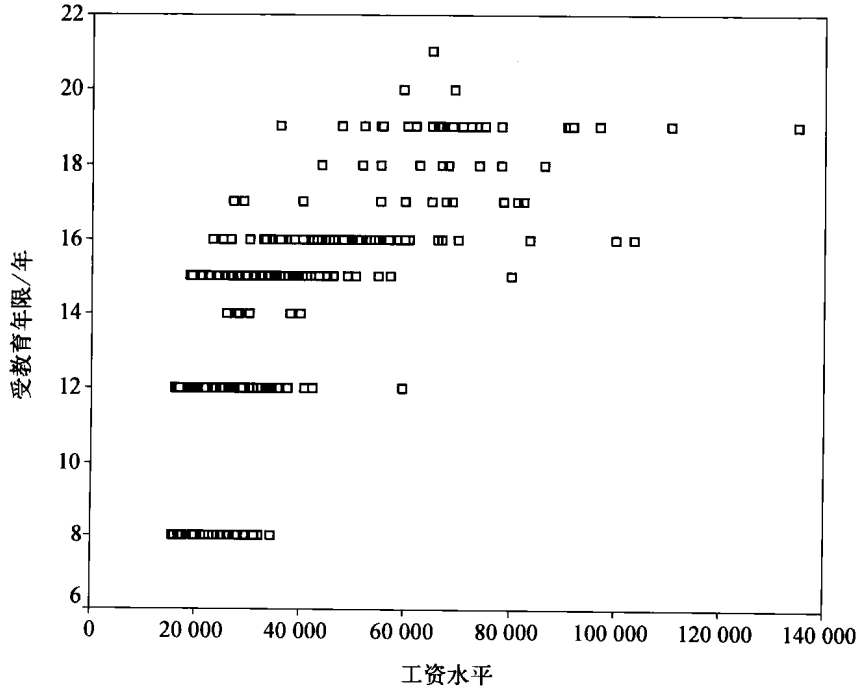

图 1.1　受教育年限和工资水平的简单散点图

从图 1.1 中可以看出，这些点不是均匀地分布在坐标轴中，但可以直观地观察出，这些点的分布存在一定规律，工资水平和受教育年限呈现正相关关系。

2. 三维散点图

选取例 1.2 中的工资水平、受教育年限和工作时间三个变量，分别作为 x 轴、y 轴和 z 轴，用 SPSS 绘制成三维散点图，如图 1.2 所示。

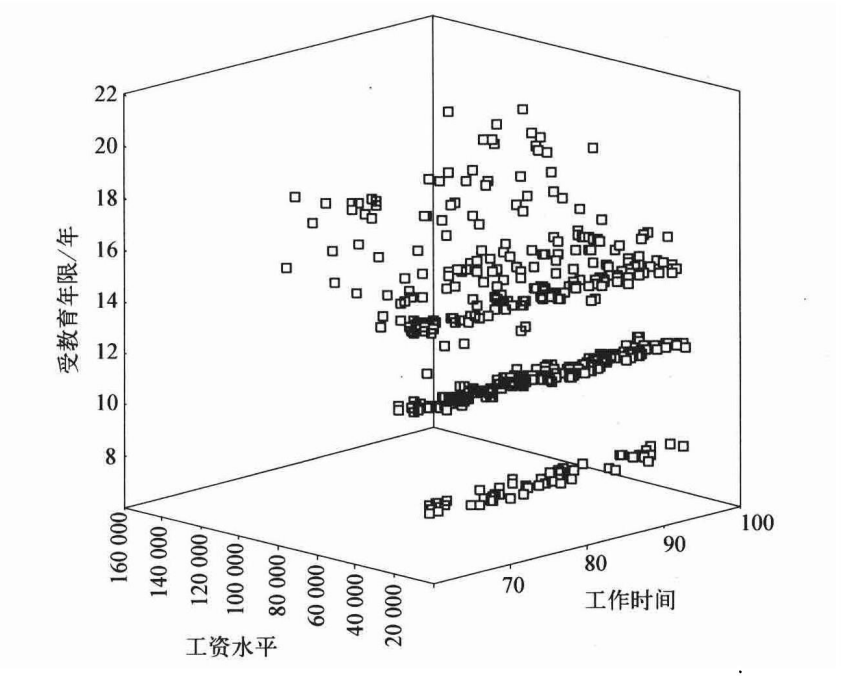

图 1.2　工资水平、受教育年限和工作时间的三维散点图

在三维散点图中，各个点在坐标轴上对应的值很难直观地看出来。由于人们的空间想象能力有限，三维散点图中数据的分布特征和变化趋势并不容易被观察到。

3. 矩阵散点图

矩阵散点图在一定程度上改进了三维散点图的不足，并能处理三个以上变量的问题，表示出多个变量间两两之间的关系。

选取例 1.2 中的受教育年限、工资水平、初始工资水平和工作时间四个变量，用 SPSS 绘制成矩阵散点图，如图 1.3 所示。

根据变量的个数，有 n 个变量就可以绘制成 n 行 n 列的矩阵散点图，从左上角到右下角分别是变量的名称，每个单元格就是一个简单散点图。从图 1.3 中可以看出，工资水平和初始工资水平存在比较明显的正相关关系，其他变量间的关系相对比较模糊。

1.2.2　箱线图

箱线图，又称箱须图、方盒图、盒须图，是处理连续多元数据的一种常用图形。它能同时显示每一个变量的中位数、四分位差(第 3 个四分位数与第 1 个四分位数之差)、异常值以及最大值和最小值，因此能直观地表现出未分组或已分组的变量值的分布，可以粗略地看出数据的对称性、分散性以及异常情况等。

从外观上看，箱线图中变量的每个分组都由一个箱子形状的封闭矩形框和上下两段线段组成。“箱子”为箱线图的主体，箱子的下边缘线表示变量的第 25 个百分点，上边缘线表示第