多元统计理论实验与案例

王建军 宋香荣 编著

中国统计出版社
China Statistics Press

多元统计理论实验与案例

王建军 宋香荣 编著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

多元统计理论、实验与案例 / 王建军，宋香荣编著. —— 北京 ：中国统计出版社，2014.1

ISBN 978－7－5037－7025－8

Ⅰ. ①多… Ⅱ. ①王… ②宋… Ⅲ. ①多元分析－研究 Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字(2013)第 279673 号

多元统计理论、实验与案例

作　　者/王建军　宋香荣
责任编辑/陈悟朝　姜　洋
封面设计/黄　晨
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn
印　　刷/三河市双峰印刷有限公司
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/280 千字
印　　张/16.75
印　　数/1－2000 册
版　　别/2014 年 1 月第 1 版
版　　次/2014 年 1 月第 1 次印刷
定　　价/34.00 元

前　言

统计学是分析数据信息的科学。多元统计是同时分析多个变量的统计方法，是分析多维数据的重要工具。

信息时代统计学的应用背景发生了巨变，数据的存储方式由纸质到软盘、硬盘再到云盘，数据的远距离传输方式由邮寄到网络传输，数据的搜集在交易中可以瞬间完成，加之随着计算机与统计软件的升级，数据的分析工具功能正变得越来越强大。这使得统计分析方法从理论发明到实际应用的时间大幅缩短，统计的应用领域无限扩展。

多元统计教学模式正在不断改革。现实世界的复杂性和影响因素的关联性，使得多元统计必然成为描述复杂事物、多角度分析实际问题的一个非常重要的统计分析方法。多元统计的内容不断扩充和丰富，使得在统计软件已经解决了多元统计复杂计算的背景下，统计学的教学模式必然相应发生变化，教学方法要突出统计课程的应用性特点。教学模式由数学性质推导为主演变为以培养解决实际问题的能力为主，统计实验实践操作是重要的教学环节，教学中增加了实验教学比例，突出综合案例的作用，强调通过综合型、设计型实验提高学生的动手能力，锻炼学生将多元统计理论与正确应用多元统计方法研究分析实际问题结合的能力。

本书以国际著名多元统计教材为蓝本，内容有描述分析、矩阵运算、主成分分析、因子分析、多元回归模型、聚类分析、判别分析、典型相关与对应分析的理论、实验与案例。本书的特点在于：将多元统计计算和统计软件 SAS 有机结合，突出分析结果的解释。运用 SAS 软件编写多元统计实验和案例，培养学生编写程序的能力。附有 R 软件与 SPSS 软件的多元统计分析的程序。

本书设想读者为财经类统计专业学生，精选经济管理类的案例，各章选择的案例以宏观经济数据为主，具有真实背景的指标来源于统计年鉴，实验所用统计方法是综合性的，分析结论具有一定的代表性

和实际意义，能够提高学生分析和解决问题的能力。

书中内容是作者多年从事多元统计教学经验和科研的总结，由浅入深，由简到繁，由易到难，由验证实验到综合实验，注重能力培养。多年教学经验告诉我，本科学生应用多元统计的困难是如何选择研究的变量，如何提炼统计学的问题，而用多元统计解决实际问题的关键是如何选择多个相关变量。本书中强调了统计方法与变量选择的关系。

本书的完成得到了统计专业研究生刘定安、李如意、杨辉平的大力帮助，特在此表示谢意。书中观点如有不妥，恳请广大读者不吝赐教，以便及时修正，书中程序和数据可通过 xjcdtjx@sina.com 与作者联系。

本教材获得新疆 2012 年度“自治区高等教育地方特色和民文教材建设计划”专项经费资助。

王建军

2013 年 8 月

目　　录

第一章

多元变量选择与描述分析

1.1 多元变量选择与数据

1.1.1 多元统计的概念

统计学(Statistics)研究数据的搜集、整理、分析、解释和展示。统计学是分析数据信息的科学。

多元统计(multivariate statistics)是同时观测与分析两个以上变量的统计分析方法。多元统计分析是研究多个随机变量之间的相互依赖关系以及内在统计规律的一门统计学科。如今已广泛运用于工业、农业、医学、气象、环境以及经济、管理等诸多领域中。

表 1—1 多元统计与一元统计的比较

	一元统计	多元统计
变量	研究一个随机变量,一维	同时研究多个随机变量,p 维随机向量
样本	一维数据,n 个数组成的向量	多维数据,$n \times p$ 的二维矩阵
均值(期望)	均值(期望)是一个数	均值(期望)是一个向量
方差	方差是一个数	协方差矩阵 Σ,$p \times p$
变量间关系	假设变量间独立	假设变量间有相关关系
分布的关系	一元正态	多元正态
	卡方分布	Wishart 分布
	T 分布	Hotlling T^2 分布
	F 分布	Wilks 分布
统计方法	参数估计,假设检验,方差分析	多元回归,聚类\判别,因子\主成分,对应分析,典型相关 ,多维标度
密度函数	一个变量,密度曲线	联合密度函数曲面
研究的关系	研究样本点间关系	研究样本点间关系与变量间关系

多元统计分析的一般步骤：

1. 选择研究与解释的社会经济现象或自然现象；

2. 选择现象可量化的变量：先选因变量，再选影响因变量的自变量；

3. 搜集数据：选择样本点，同时观测样本点的多个变量的值。社会经济主要是不可控制的观测数据，而非实验数据。变量反映的是同一样本的不同方面，所以变量间一般存在相关性；

4. 选择多元统计方法分析数据：多个变量同时参与分析，根据研究目的选择多元分析方法。可运用的统计分析软件有 R、SAS、SPSS、MINITAB、Matlab 等；

5. 解释结果，获取信息。

多元统计方法的选择如图 1－1 所示。

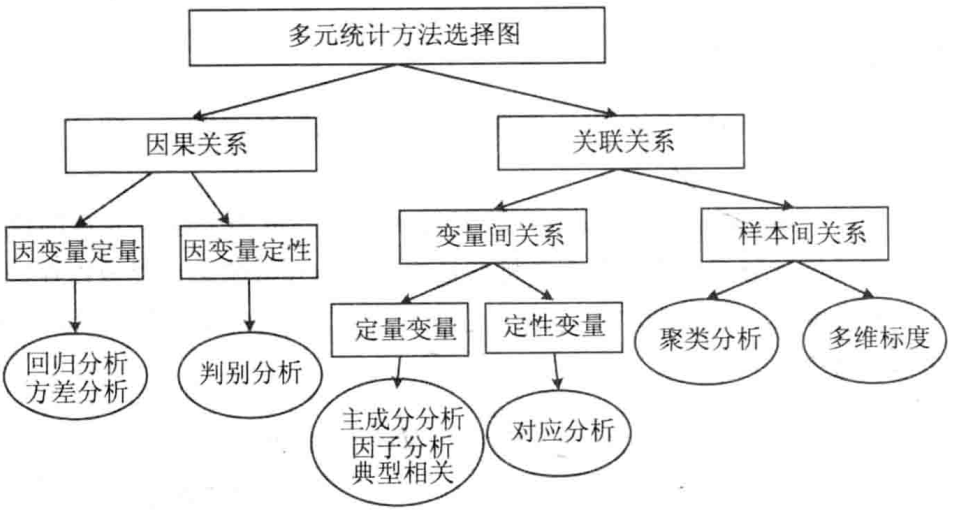

图 1－1　多元统计分析方法选择图

1.1.2　数据的分类与选取

统计数据(data)是对研究对象的某些现象进行测量的结果，也就是对统计变量进行测量的结果。比如：对居民收入情况测量可以得到居民收入数据，对农业总产值进行测量可以得到农业总产值数据。统计学的数据主要用二维表格形式来显示，如表 1－2 所示。

表 1—2 数据形式表

国家	Country	国内生产总值（亿美元）	资本形成率（%）	人口（万人）	万美元国内生产总值能耗（吨标准油/万美元）
印 度	India	13806	36.5	115534.8	1.95
日 本	Japan	50330	20.5	12755.8	1.26
哈萨克斯坦	Kazakhstan	1153	30.5	1592.5	3.96

注：数据来源于《国际统计年鉴 2012》

(1)分类数据、顺序数据、数值型数据

按照所采用的计量尺度，我们常把统计数据分为字符型数据和数值型数据。

字符型数据(character data)也称为**定性数据**(qualitative data)，是指只能分为某些类别的数据。字符型数据又分为分类数据和顺序数据。**分类数据**(categorical data，nominal data)指只能归于某一些类别的非数字数据，比如说：按性别分为男、女两类；企业按行业属性分为金融业、制造业、建筑业等；按经济的三次产业划分为一、二、三产业；按经济成分分类有公有经济与非公有经济。**顺序数据**(ordinal data)是指只能归于某一有序类别的非数字型数据。例如说学生成绩分为优、良、一般和差；一个人对自己的收入满意度为：非常满意、比较满意、一般、比较不满意和相当不满意；学历，职务等。字符型数据可以通过人为指定数据代码成为虚拟数值数据，如：用 1 代表男性，0 代表女性，但不计算中位数、标准差、平均值等统计量。字符型数据是经济和社会研究中，特别是调查问卷研究中最常用的数据。字符型数据的统计分析方法主要有频数分布表、比例结构分析，比较分析，列联表关联分析，对应分析，饼图和条形图展示等。

数值型数据(metric data)是指按数字尺度测量的观测值，其结果表现为具体的数值。比如一个人的身高、一个学生某学科的考试分数、一个国家某年的国内生产总值。数值型数据主要统计分析方法有散点图、箱线图和直方图展示、相关分析、回归分析、因子分析、主成分分析、典型相关、聚类分析和判别分析等。

(2)截面数据和时间序列数据

按照被分析的现象空间与时间的关系，可以把统计数据分为截面数据和时间序列数据。

截面数据(cross-sectional data)是指在相同或近似相同的时间点上收集的数据，这些数据通常是在不同的空间地点获得的，用于描述现象在某一时刻的不同情况。比如 2011 年全国各省市的 GDP 值是截面数据。

时间序列数据(time-series data)是在相同空间不同的时间上收集的数据,这些数据是按时间顺序排列的,用于描述现象随时间变化的情况。比如 1995～2010 年中国的棉花总产量是时间序列数据;某种股票十年来的股价也是时间序列数据。还有一种数据被称为**面板数据**(panel data),是截面数据与时间序列数据综合起来的一种数据类型。

多元统计分析选取的数据主要是截面数据,这是因为时间序列数据不平稳的较多,可能存在伪相关、自相关等现象,样本间的相关性会对多元统计分析造成许多麻烦,甚至可能导致一些错误的结论。

1.1.3 变量的分类与选取

统计学的变量(variable)是研究中样本的特征,这些特征是可测量的。与数学中的变量概念有差异,数学中的变量是指未知数或可变量的数,统计的变量是数据表中的项目,说明样本某些特征的概念,其特点是从一次观测到下一次观测可能会呈现出差别和变化。比如"商品的销售额"、"受教育程度"、"性别"等都是变量。

统计数据是统计变量的样本的取值。变量分为字符型变量和数值型变量。**字符型变量**(qualitative variable)也称定性变量,可分为**分类变量**(categorical variable)和**有序变量**(rank variable),常见的字符型变量比如有"性别","成绩档次","学历水平"。**数值型变量**(metric variable)可以分为连续变量,如"国民生产总值",离散变量,如"人口数量"等。在统计学中连续型变量的值是测量或计算得到的,一般允许有测量误差,如身高与体重。

个体也称为观测(case, observation)或记录(record),个体是搜集样本数据的实体。样本是个体的集合,用于代表总体(所有个体),对每个个体测量多个变量就是多元数据。

多数统计软件中是以表格形式存放数据,一行(row)是一个个体,一列(column)是一个变量,变量代表一列数据集,变量也是列的名称。

需要注意的是,在进行多元统计分析时,当研究的变量确定以后,需要收集样本数据,多元统计中样本量大于变量个数时,可以保证协方差矩阵非负定。**样本量大于变量个数是多元统计中的最低要求**!样本的个数理论上最好是变量个数的 5～10 倍以上。选择变量数 5～10 倍的样本数一是因为选择样本数据一定要对总体有代表性,代表性越好,结论的意义就越大;二是因为样本数越多,偶然因素的影响被抵消的可能就越大,研究现象的统计规律就更加容易显示出来,统计模型的检验结果才有意义。有人认为样本量大于 30 就是大样本,其实这是一种误解,在计

算工具落后的时代，当样本量达到 30 时的 T 分布可用正态分布近似，这是数学上的近似，并不适用于统计学研究的情况，因此不能理解为样本量达到 30 就是大样本，就足够了。样本量需要的多少一般与统计方法有关，如建立统计回归模型需要的样本量一般要多于聚类分析需要的样本量；样本量也与变量的多少和估计精度有关。变量越多，需要的样本量就越多，统计估计精度要求越高，样本量需要越多。

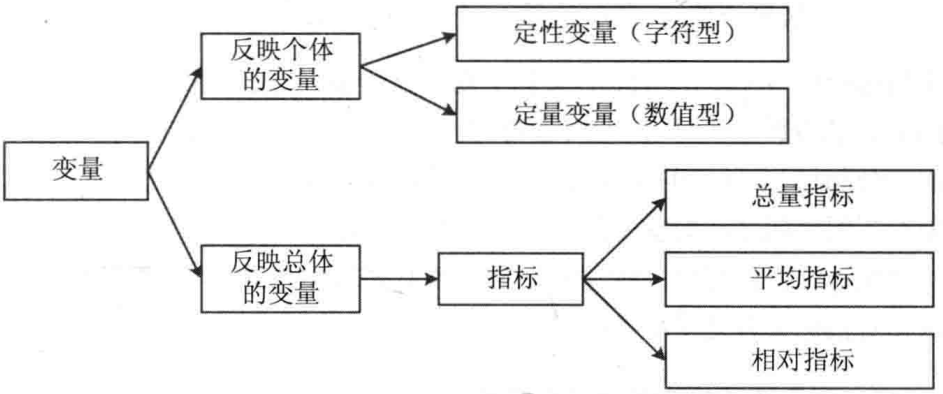

图 1—2　变量分类选择图

1.1.4　统计指标的分类与选取

统计指标(statistical indicator)简称指标，是变量的一种重要形式，是研究社会与经济宏观现象的数据，是经过汇总的数据，是反映社会经济现象总体(population)或子总体(subpopulation)的变量和具体数值。统计指标按照其反映的内容或其数值表现形式，一般分为总量指标、相对指标和平均指标三类。

总量指标(aggregate indicators)是反映社会经济现象总规模、总水平或工作总量的统计指标。它的数值随着统计范围的大小而变化。总量指标用绝对数形式表现，一般带有量纲计量单位。例如货币发行量、土地面积、国内生产总值、财政收入和企事业单位个数等。总量指标是计算平均指标和相对指标的基础。同时总量指标也可以用来进行总量的统计趋势预测分析，总量与总量的相关关系分析。但是各样本的总量指标之间不适合做比较。比如甲班有 60 个人，乙班有 40 个人，拿两个班的数学总成绩做比较是没多少意义。同样，把两个国家的 GDP 相加后再除 2，这样求平均也没意义，总量指标慎做平均。

相对指标(relative indicator)是两个总量指标之比。如第三产业占比、人均收入、人均国内生产总值等。相对指标通过两总量指标的对比，可以显示事物的相对水平，能更深刻地说明事物之间的联系，提供事物之间共同的比较基础。比如说男

女比率可以认识总体的性别结构，经济增长率可以很清楚地反映现实经济的发展趋势，合格率可以评估产品的生产是否科学合理。注意在运用相对指标时，两个总量指标必须具有可比性，且对比有一定的意义。比如说用汽车的总数除以水库总数，这样的相对指标就没多少意义。相对指标的计算举例如下：

居民消费占 GDP 比重指的是居民消费支出与 GDP 的比值，计算公式为：

$$\text{居民消费占 GDP 比重}=\frac{\text{居民最终消费支出}}{\text{GDP}}\times 100\%$$

平均指标(average indicator)又称平均数或均值，它反映的是现象在某一空间或时间上的平均数量状况，如人均工资等。平均指标经常用来进行同类现象在不同空间、不同时间条件下的对比分析，从而反映现象在不同地区之间的差异，或揭示现象在不同时间之间的发展趋势。平均指标主要用于研究变量间的关系。

多元统计分析的聚类分析与判别分析不宜选取总量指标，一般选取可以进行比较的相对指标或者是平均指标。

1.1.5 多元统计的变量选择与指标体系

多元统计的多个变量不是随意选取的，一组变量要有实际的意义，要有相关性、关联性或因果关系。多元统计分析变量的选择影响分析结果，选取变量要先进行理论分析，如果分析结果与经验和实事太不相符时，多数是与选择变量时缺少重要的关键的变量有关。变量选择可以借鉴社会统计与经济统计经常用到的指标体系。若干个相互联系的统计指标组成一个整体就称为统计指标体系。用多元统计分析做经济研究时，常用到的宏观和微观指标如表 1—3，表 1—4 所示。

表 1—3 宏观社会经济指标体系

指标体系	指标名称
经济增长	人均 GDP，GDP 指数
结构优化	服务业增加值占 GDP 比重，居民消费占 GDP 比重，高技术产品产值占工业总产值比重，城镇化率
发展质量	财政收入占 GDP 比重，全社会劳动生产率
收入分配	城乡居民收入占 GDP 比重，基尼系数，城乡居民收入比
生活质量	城镇居民人均可支配收入，农村居民人均纯收入，人均住房使用面积，互联网普及率，每万人拥有公共汽(电)车辆，人均预期寿命
公共服务支出	人均基本公共服务支出，基本公共服务支出占财政总支出比例

续表

指标体系	指标名称
文化教育	文化产业增加值占 GDP 比重，平均受教育年限
卫生健康	5 岁以下儿童死亡率
社会保障	基本社会保险覆盖率，农村、城镇居民享受最低生活保障人口比例
资源消耗	单位 GDP 能耗，单位 GDP 水耗，单位 GDP 建设用地占用
CO_2 排放	人均二氧化碳排放量，单位 GDP 二氧化碳排放量
环境治理	环境污染治理投资占 GDP 比重，工业“三废”处理达标率，城市生活垃圾无害化处理率，城镇生活污水处理率
科技投入	万人 R&D 人员全时当量，R&D 经费支出占 GDP 比重
科技产出	高技术产品出口占总出口比例，万人专利授权数

表 1—4　微观社会经济指标体系

指标体系	指标名称
规模经济实力	市场占有率，利税占有率
投入产出能力	全员劳动生产率，成本费用利润率
营运能力	流动资产周转率，产品销售率
盈利能力	总资产报酬率，净资产收益率，人均利润，成本费用利用率
偿还能力	资产负债率，营运资金比率，应收账款周转率
发展能力	利润增长率，资本保值增值率，资本增加值率，销售增长率，技术开发经费支出比率

1.2　多元数据的描述统计

1.2.1　数据中心位置的描述

数据中心位置(location)是指一组数据分布的中心值或代表值，反映了一组数据的中心点位置所在。常见的反映**中心位置**的统计量有平均数(mean)、中位数(median)、众数(mode)和分位数(quantile)等。

平均数也称均值，统计学的平均主要是指加权平均，设一组样本数据为 x_1，x_2，…，x_n，各组组值的频数分别为 f_1，f_2，…，f_n；样本量个数为 $\sum_{i=1}^{n} f_i$，则样本的加权平均数 $\bar{x}$ 的计算公式为：

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}x_if_i}{\sum_{j=1}^{n}f_j}$$

算术平均(average)是平均(mean)的特例,权重为 1,它是一组统计数据相加以后除以数据的个数得到的结果。

中位数是指将统计数据排序后处于中间位置上的变量值。用 M_e 表示。设将一组数据为:$x_1,x_2,\cdots,x_n$,按从小到大的顺序排序后为 $x_{(1)},x_{(2)},\cdots,x_{(n)}$,则中位数为:

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

中位数是从中间点将全部数据分为两部分。与中位数类似的还有四分位数、十分位数和百分位数等。

选用平均数还是选用中位数来描述数据的中心位置的原则是:如果数据分布的偏态比较大,一般使用中位数,以免受个别偏大或偏小的异常值的影响;如果数据分布的偏态比较小,基本上是对称时,一般使用平均数。

1.2.2 离散程度的描述

数据的**离散程度**(variability measure)是指数据远离其中心值的程度,反映数据的变化。数据的离散程度越大,中心值对该组数据的代表性就越差;反之,离散程度越小,中心值的代表性越强。

统计学非常注重研究数据的离散程度,统计数据最重要的特征之一是变异性,是指数据的离散程度,尺度(scale)是指度量数据离散的标准。

常见的反映离散程度的统计量有四分位距(inter-quartile range,IQR)、中位数绝对偏差(median absolute deviation,MAD)、标准差(standard deviation)、离散系数(coefficient of variation,CV)等。在度量离散程度时,我们主要用标准差,其次是四分位距和中位数绝对偏差。

四分位距表示上四分位数与下四分位数的差距,也称为内距或四分间距。用 Q_d 表示。其计算公式为:

$$Q_d=Q_U-Q_L$$

四分位距反映了中间 50%的数据的离散程度,其数值越小,说明中间数据越集中;其数值越大,说明数据越分散。

中位数绝对偏差是一个非常稳健的估计量。其计算公式为:

$$\mathrm{MAD}=med_i(|x_i-med_j(x_j)|)$$

MAD[①]一般和中位数一起使用，中位数加减一个 MAD 的距离约等于四分位距。所以它也是反映中间 50%的集散程度的统计量。一般说来中位数加减一个 MAD 的距离要小于四分位距。

方差(variance)是各变量与其平均数的离差平方的平均数。方差的算术平方根称为**标准差**。方差(或标准差)能较好地反映出数据的离散程度，是实际中应用最广的离散程度的度量值。方差用 s^2 表示，标准差用 s 表示。方差和标准差的公式为：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

方差常用于理论分析，不常用于统计研究的度量，主要是因为它与平均数等的计量单位不一样，也就是量纲不一样，它的计量单位是平均数的计量单位的平方。而标准差的计量单位是与平均数等相同的，与平均数同量纲，在统计分析里面的度量经常用到标准差。通常 $s\approx 1.48\mathrm{MAD}$。

离散系数又称变异系数，它是一组数据的标准差与其对应的平均数之比。其计算公式为：

$$v_s=\frac{s}{n}$$

离散系数是测度数据离散程度的相对统计量，主要用于比较不同样本数据的离散程度。离散系数大，说明一组数据的离散程度也大；离散系数小，说明一组数据的离散程度也小。

1.3　多元描述分析的 SAS 程序及实验指导

1.3.1　建立 SAS 数据集

用 SAS 软件分析数据，首先要将数据导入 SAS，建立一个 SAS 数据集，常用

① 统计学中还有一常见的统计量也简称为 MAD(Mean Absolute Difference)，称为平均绝对误差，主要用于评估统计预测中的预测效果。