高等院校“十三五”规划教材

SPSS统计分析

——应用案例教程

廖小官　主编

南京大学出版社

十三五 高等院校"十二五"规划教材

SPSS统计分析
——应用案例教程

主　编　廖小官
副主编　李文霞　贾晓燕

南京大学出版社

图书在版编目(CIP)数据

SPSS统计分析：应用案例教程 / 廖小官编. — 南京 ：南京大学出版社，2016.7
高等院校“十三五”规划教材
ISBN 978-7-305-16530-6

Ⅰ. ①S… Ⅱ. ①廖… Ⅲ. ①统计分析—软件包—高等学校—教材 Ⅳ. ①C819

中国版本图书馆CIP数据核字(2016)第037085号

出版发行 南京大学出版社
社　　址 南京市汉口路22号　　邮　编 210093
出 版 人 金鑫荣

丛 书 名 高等院校“十三五”规划教材
书　　名 SPSS统计分析——应用案例教程
主　　编 廖小官
责任编辑 惠　雪　吴　华　　编辑热线 025-83597087

照　　排 南京南琳图文制作有限公司
印　　刷 常州市武进第三印刷有限公司
开　　本 787×1092 1/16　印张 21.25　字数 525千
版　　次 2016年7月第1版　2016年7月第1次印刷
ISBN 978-7-305-16530-6
定　　价 42.00元

网址：http://www.njupco.com
微信服务号：njuyuexue
销售咨询热线：(025) 83594756

前　言

统计分析方法是进行问题研究必不可少的工具。一直以来，对于数学基础较薄弱，对概率论、数理统计缺乏系统学习的学习者来说，掌握和运用统计分析方法，分析问题，揭示现象的特征及其规律，变得非常困难。本书的教学安排，以案例的方式呈现统计方法运用，不管是否具有统计学基础，都能基本掌握统计方法以及软件的运用。本书适用于本科生、研究生和科研工作者统计分析学习参考。

根据分析目的，本书分成五个部分。第一部分为特征描述，主要分析方法包括单变量描述统计、复合变量描述统计(分类变量与分类变量，分类变量与尺度变量)和多重响应变量的描述统计。第二部分为特征检验，分析方法主要为参数检验(单样本、独立样本、成对双样本的均值检验，单因素方差分析)和非参数检验(卡方检验、二项检验、游程检验、单样本 K－S 检验、独立双样本、相关双样本、独立多样本和相关多样本检验)。第三部分为关系分析，主要包括相关分析、一般线性回归分析、有序因变量回归、Logit 回归分析、广义线性回归分析、神经网络分析。第四部分为特征判别，主要有聚类分析、判别分析、控制图形与 ROC。第五部分为数据问题分析，涉及可靠性分析、缺失值分析和多重插入。由于篇幅所限，一些统计方法如时间序列分析、面板数据模型、非参数与半参数回归、多方程系统模型等需要参考其他书籍。

本书由江西农业大学廖小官担任主编，武昌理工学院李文霞、郑州工程技术学院贾晓燕担任副主编。全书由廖小官进行总纂、修改和定稿。

所涉及的案例数据均来自 SPSS 软件自带的数据，操作采用 SPSS17.0 完成，软件操作说明均参照 SPSS 相关的英文版使用手册。一些引起歧义的地方请参照原版。本书的使用仅限用于教学。望读者在学习使用过程中购买正版软件，他们将提供相应的技术支持和资料下载。

编　者

2016 年 5 月

目　录

第三部分 关系分析

第四部分 特征判别

第五部分 数据问题

第1章 绪 论

统计理论与方法作为定量分析的学科，与其他学科一样，都是描述世界，揭示世界规律的一门学科，只不过，其角度和思维范式不一样而已。为此本章首先介绍统计学科的理论范式，为了区别于其他学科，概要地展示该学科的角度和思维。为了更好地运用其理论与方法，必须遵循分析的步骤，而定量分析的步骤就是统计方法所遵循的步骤，这将在本章的第二部分进行介绍。统计理论与方法作为定量分析的工具学科，首先要将对象进行量化。这是因为研究对象特征不一样，量化形成的数据类型也就不一样，所运用的统计方法也不一样。第三部分将介绍数据类型。第四部分则介绍基于数据类型和分析目的粗略地构建统计方法的框架，统计方法的框架很重要，用户可以根据分析的目的，研究对象的特征，来选择合适、恰当的方法。最后简单地介绍本书的安排。

1.1 统计学科的范式

人类通过智慧，不断传承、探索和创新，形成了认识世界的成果——林林总总的各种学科。不同的学科以其独特的研究对象、视角和范式来描述世界。比如微观经济学看待物品和现象，看到的是市场和交易、供给和需求、成本和收益、价格和满足。甚至“爱”，经济学分析的则是爱的效用、爱的付出，以及爱的交易——婚姻或者出轨。而数学的语言，则是界定爱的定义，衍生出爱的性质，在假设的基础上推导出爱的定理，诸如什么因素影响爱——爱的函数。而文学的语言脱离数理的严谨逻辑，以歌唱之，以诗咏之，以舞跳之。统计学科则是众多学科的一个门类，以其独特的范式去描述世界。统计学科的视角主要表现在以下方面。

1. 量的方式

以量的方式去描述世界，是统计方法分析现象的基础。统计学看世界，首先看到的是量。看到研究对象，看到其不同的维度和特征，统计学首先要在定性分析基础上把它们量化。一个人身体是否健康，统计学要做的是在构造健康维度及其指标的基础上收集这些相关数据；身高高不高，用的是数据说话；爱得够不够深，需要构造爱的指数；幸福不幸福，构造的是幸福指数；宏观经济形势好不好，根据宏观经济理论，用的是GDP增速、CPI、PPI、消费者信心指数、投资者信心指数、经济景气指数等。相应的数值多少，反映了相应研究对象的状况。量化是统计学科描述世界的独特视角。当然这种量化必须建立在定性分析的基础上。不同类别的特征量化方式也不同。单维的特征，如性别用0、1或1、2(名义变量，nomial)表示，喜欢程度用如1、2、3、4(有序变量，ordinal)表示，身高、体重用的是具体数值(尺度变量，scale)；而综合多维的特征则用的是指数方式，或者数据降维，如因子分析。

2. 随机的方式

统计学看世界，除了一些如人数用精确的量的方式表示以外，更多的是用随机的视角。看到的爱，不再是“要么爱，要么不爱”，而是一个不确定性的区间分布。看到居民可支配收

入，不仅仅是平均收入，而会看其服从何种分布，也会看其中等收入、分位数收入、标准差等。这种随机的视角是建立在概率论的基础上。概率论为其提供了可供量化的理论基础，而统计学则是其理论的实现。这种随机的范式不断革新，推动统计方法不断更新。作为随机现象的描述，随机的范式表现在以下方面：

（1）分布函数和密度函数。看到随机现象，其数据的出现，会试图用某种具体的分布去描述。如公交站等候的人数服从泊松分布，收入或成绩服从正态分布，股票的价格用对数正态分布。多维用的是联合分布函数。统计方法为其估计和检验提供了支持。

（2）矩和分位数的方式。一阶矩如期望，二阶矩如方差、协方差，三阶矩如偏态，四阶矩如峰态。分位数是指数据分布的区间点，如中位数收入表明的是排名50%的收入水平。

（3）随机过程。对于时间序列数据，为描述不同时间数据的产生，试图构造一个随机过程去描述，或者判断它是否平稳。如股票价格的波动过程，认为它是一个布朗运动的随机过程。

围绕以上随机描述的视角，统计学科构建参数或非参数方法，采用收集的数据，构造一系列统计量去估计或检验或推断，试图真实反映世界的存在。但或许我们只能接近真相，永远也达不到真相。比如食用的谷物中含有一种有毒物质——黄曲霉毒素。有多种谷物，需要甄别何种谷物符合标准。黄曲霉毒素含量是随机的，不能根据某一含量去臆断哪种谷物符合标准，哪种谷物不符合。统计方法根据随机范式，为我们提供解决的方案和步骤。这种随机现象，先用矩的方式如样本均值、标准差、峰态、偏态描述其数据分布特征，用参数或非参数的方法，构造相应的统计量，如卡方统计量，T统计量等，检验其是否服从正态分布，它的含量是否符合标准（低于20PPB）等。

3. 总体的方式或大多数法则

统计方法所揭示的特征及其推断的特征代表的是总体的特征和规律，而非个案。尽管收集的数据都来自个案，不管是小样本精确性质，还是大样本近似性质都需要足够数量的样本，才能保证统计方法的有效性。“一花独放不是春，万紫千红春满园”。统计方法不是仅根据某个案的特征，而是根据大多数法则，在相应置信水平下推断总体上变量的特征和统计依赖关系。即使把统计规律用于推断个体，这种推断也是在一定的置信水平下的置信区间。

1.2 定量分析的步骤

运用理论分析问题有其独特的范式，从而衍生出其基本的分析步骤。定量分析的步骤是运用定量分析方法分析问题遵循的一般步骤，是运用分析范式实现分析目的的步骤。

1. 定性分析基础上确定变量

这是统计分析首先要面临的问题。定性分析是定量分析的基础，没有定性认识，量的分析就失去依靠，统计分析无从谈起。定性认识怎么样，直接决定分析能否触及事物的本质，选择的变量是否重要，所揭示的特征和关系是否是基本且重要的，这直接决定着构建统计模型的成效，决定着统计分析的成效和意义。比如，评价一个人身体健康状况，没有健康的定性认识，就不能很好地构造出反映健康状况的指标，就不能获得相关的数据信息，也就不能准确地评价身体正常与否。因此，在做定量分析之前，必须先做定性分析，以反映事物的本质特征和重要关系，从而确定重要变量及它们之间的关系，为有效的统计分析打下基础。变量及其关系的选择，是以是否基本全面反映研究对象为标准，在全面性、模型的易处理性和分析目的三者之间进行权衡。分析精确度要求不高的，选择的重要变量少一些，关系简单一些，也容易处理；而精

确度要求高的，考察的重要变量就多一些，关系就复杂一些，虽能很好地反映研究对象，但也具有处理的复杂性。因此，统计分析做得好不好，先要看定性分析做得怎么样，是否包含重要变量和重要关系，是否能较好地反映实际事物。

2. 收集数据

变量确定后，接着就是收集数据。数据的质量怎么样，样本容量是否足够大，数据是否全面客观、具有代表性，是否基本反映实际情况，这些都关系到统计分析的有效性。数据的获得有两种方式：一是直接获得。通过设计调查方案，设计好调查问卷，选择合适的调查组织形式，组织人员进行调查获得数据。二是间接获得，收集官方公布或已调查好的数据。这些内容相关的统计类图书都有详细介绍。这里针对数据的收集需要补充几点：① 收集数据之前要先做好定性分析、调查报告或者研究论文的理论分析。这样可以避免当有理论创新时收集的数据遗漏了重要变量，进而失去统计分析的意义。② 在调查问卷或设计正式定稿前，要做预调查。调查问卷设计得好不好，直接关系到数据的质量。通过预调查，可以发现设计中存在的一系列问题。③ 问卷要附着问卷说明。问卷设计者与调查人员或是填报人员往往不相同，问卷中的问题往往会产生歧义，而问卷说明则能很好地解决这一问题。④ 间接数据的获得，要注意指标的内涵与外延(口径)，以及时效性、客观性和可比性。

3. 问卷审核，录入数据

为保证数据的质量，需要对问卷的质量进行审核，避免录入虚假数据。问卷审核一般有以下常用方法：

(1) 问卷设计中的疑问框法。在问卷的不同部分设计一些相同的但表述方式不一样的客观的问题，如果答案不一致，对该问卷的数据需谨慎。

(2) 逻辑推理法。这是常用的方法。例如，一个25岁的人绝不可能有20年的工龄。

(3) 经验法。这种方法来自问卷调查的经验和常识。主观问题的答案，比如政策满意度，若具有高度的一致性，则应怀疑该调查人员调查的问卷质量。例如，一户农村家庭年收入20万，又没有其他特别的支出项目，且居住的房子破破烂烂，则可以怀疑该问卷的质量；再比如一户农村家庭仅有3亩地，用来种粮，年收入10万元。这些问卷都不合常理。

(4) 信度分析，又叫可靠性分析。这是SPSS软件自带的方法，认为相同特征的样本有相似的数据。如果Cronbach's alpha值显著且较高，则认为问卷的可信度较高。

(5) 复调查。随机选择一些初调查的样本进行再次调查，根据其误差来调整最终的结果。

只有录入数据满足软件的要求，才能进行统计整理和分析。但在录入数据过程中缺乏经验的人往往会犯一些错误。① 样本忘记输入编号。给每份问卷进行编号，在输入数据时也要输入相应的编号，便于在数据输入错误时，进行核对。② 数据带单位输入。数据带有单位，软件则认为数据是字符串，不能进行相应的统计分析。SPSS软件自带数据的格式，只要输入数据，就可显示，如货币符号等。③ 选项输入字母。一般选项输入数字，数字表示的含义可以在软件中界定变量标签。如果是字母，则认为是字符串，大多数情况下如变量为满意程度，就不能进行相应的分析。字母大小写不一致则认为不是同一答案。④ 同一称谓文字不相同。如果必须输入文字的，如地址，相同的地址必须是相同的文字。SPSS软件中，认为“北京”与“北京市”不是同一地址。⑤ 缺失数据，空白表示。当然这种空白，SPSS软件也作为缺失数据处理。但有时候不能区别是遗漏的还是缺失的。一般缺失数据会用特别的数值表示，并在软件中界定。⑥ 多项选择的输入在一个单元格。如选项1、3，在一个单元格输入。这样输入软件

只能将其视为一个字符串，只能统计1、3的频数，而不能统计选1的频数，可统计的项数会非常多，因此没有多大意义。输入方法有两种：一种是标题栏为各选项，每个样本的数据输入，在相应的选项输入0或1，其中0表示未选该项，1表示选择该项；另一种是限制项数的排序选择，标题栏输入第一选项、第二选项等，每个样本的数据输入，在下面单元格中输入相应选择的项。关于如何界定多重响应集，详见3.4节多重响应分析。

4. 数据分析

在数据录入完成后，需要进行数据分析。如何用SPSS软件来实现数据分析？这是本书所要讲述的主要内容。数据分析一般包括以下方面：

(1) 整理数据。用图表、指标的方式揭示数据分布特征。这是统计分析首先要做的。软件提供了不同程序显示适合不同数据类型的图表，如分类数据可以用饼图描述，而连续变量则不可以；分类变量之间的复合分组，可以用交叉表显示，而分类变量与尺度变量之间的则不行。

(2) 特征检验与推断。特征主要是指均值、方差、分位数、分布形态等。有单样本、双样本和多样本的，有参数检验和非参数检验。比如打包机工作是否正常，标准是100 kg。抽取一批数据，如果数据服从正态分布，就可以用SPSS软件中均值程序进行检验。

(3) 关系检验与推断。对变量之间是否存在显著的关系，关系多大，或者是某变量的数据是如何被其他变量决定所产生(数据的产生机制)，所做的统计分析。常见的有相关分析和回归分析。不同的数据类型，所满足的假设条件不一样，采用的分析方法和程序也不一样。

(4) 特征判别。根据多维变量数据，采用一定的距离标准和选择规则，对样本或变量进行分类。比如银行根据客户的多维特征判别出不同风险类别的客户。为更好地描述样本的综合特征，可以采用因子分析或主成份分析，把多个变量的数据特征用较少的因子或主成份表示。

1.3 数据类型

对研究对象的特征量化，得到变量数据，这是统计分析的前提。不同的数据类型适用的分析方法也不一样，为此在做统计分析时，需要明确数据类型。数据类型有不同分类方法，常见的有以下两种。

(1) 根据对象特征，也就是单位标志不同，所量化形成的不同数据类型。对象特征一般划分以下几类：一是如性别、企业性质的属性特征；二是如满意程度、喜好程度的状态程度特征；三是如身高、体重可以具体测量的数量标志。因此，相对应形成的数据类型一般称之为：名义(nomial)数据、有序(ordinal)数据和尺度(scale)数据。

(2) 根据单位(或是样本)特征发生的时间和空间划分为横截面数据、时间序列数据和面板数据。

第一种是根据单位标志衡量尺度(measurement)分成名义(nomial)数据、有序(ordinal)数据和尺度(scale)数据，对应的变量则分为名义变量、有序变量和尺度变量。名义变量和有序变量又称为分类变量。SPSS软件的数据处理方法大多以此分类。① 名义数据。名义数据指的是表示样本性质类别特征的数据，此数据只能计数，不能排序、加减和乘除。如表示性别，企业类型，选择与否等特征的数据。这些特征量化后的数据，两类别的一般用虚拟变量(取值0或1变量)，多类别的一般用1,2,3,…表示。如性别，男=0，女=1，也可以男=1，女=2；选择与否，是=1，否=0；事件发生，发生=1，不发生=0。它们作为因变量，此时取值虚拟变量更合

适一些。至于多类别的变量，如职业，不同类别分别界定为1,2,3表示。但一些分析程序，如线性回归，当自变量是分类变量时要求虚拟化，但一般作为原始数据的录入而言，名义变量还是用离散的数值表示。② 有序数据。有序数据是用来表示样本状态程度特征的数据，如满意程度、喜好、文化程度等，一般用数值1,2,3,…表示。数据可以计数、排序，但不能加减和乘除。对应有序变量一般作为分类变量处理，但有时变量间则出现线性趋势，也可作为尺度变量处理。③ 尺度数据。尺度数据用来衡量样本数量标志的数据。当单位特征可以具体测度时，用离散或连续的数值表示，如身高、体重、温度等，此时的数据称为尺度数据，其变量称为尺度变量。尺度数据包含定距数据(能进行加减，但不能乘除，如温度)和定比数据(能进行加减乘除，0没有意义，如身高)。在分析中一般不作区分，都界定为尺度数据类型。

第二种是根据主体单位特征发生的时间和空间划分，把数据分成横截面数据(cross sectional data)、时间序列数据(time series data)和面板数据(panel data, pooled data)。横截面数据是指同一时间不同主体单位同一特征的数值排列，如2015年我国各个省份的GDP数据；时间序列数据是指同一主体在不同时间某特征的数值排列，如我国历年GDP数据；而面板数据是指多个主体、多个时间某特征数值的排列，如我国多个省份历年GDP数据。在分析上把主体数量多的不同时间段的数据，称为面板数据(panel data)；把较少主体的不同时间的数据，称为混合数据(pooled data)。国内对其一般不加以区别，都称之为面板数据。这些不同数据类型在特征描述上、关系分析上(如回归分析)所采用的统计分析方法有显著的差别。比如时间序列数据在分布特征描述和检验上，看作的是一种随机过程，检验其是否平稳。而横截面数据则不存在这样的方法，更多是从集中趋势、离散趋势指标和分布形态上去描述和检验。而面板数据既要检验数据的平稳性，也要考察分析中的时间效应和主体效应。本书由于篇幅所限，关于以这种数据分类为基础的数据处理方法不能详细阐述，则以第一种分类方法为主。

1.4 统计分析方法框架

不同数据类型、不同分析目的，以及不同类别的数据问题所采用的分析方法也不一样。另外同一问题，相同的数据，也可以采用不同的方法实现同一目的。本书把统计分析方法简单地归纳如表1-1，仅供大家参考。一些分析内容如估计方法，时间序列数据、面板数据处理没有包括进去，需要大家在学习过程中加以补充和丰富。

表1-1 统计分析方法框架

分析目的	分析内容	适用情形	分析方法	分析程序	本书章节
特征描述	单变量描述统计	分类变量描述统计	描述统计：频数分布表，图(直方图、饼图、区域图)	频率(frequencies)	第2章
		尺度变量描述统计	描述统计：集中趋势、离散趋势、分布形态等指标，图(条形图、箱图、茎叶图)	频率、描述(descriptives)、P-P图和Q-Q图(分布形态检验图)	第2章
		多项选择的描述统计	描述统计：频数分布表	多重响应分析	第3章

（续表）

分析目的	分析内容	适用情形	分析方法	分析程序	本书章节
特征描述	多变量描述统计	分类变量与分类变量	描述统计：频数分布表	交叉表（crosstables）	第2章
		分类变量与尺度变量	描述统计：分组下统计指标计算，图形表示（箱图、高低图）	报告（reports）、探索性分析（explore）、均值（means）	第3章
		两个尺度变量的比率	描述统计	比率统计量（ratio statistics）	第3章
		不同类别下时间依赖事件的描述统计	生存分析	生命表（life tables）、Kaplan-Meier survival analysis	
	数据降维	多维特征用较少典型的互不相关的维度表示	因子分析、最优尺度（optimal scaling（分类变量））、对应分析（correspondence analysis（分类变量））	因子分析、optimal scaling（分类变量）、correspondence analysis（分类变量）	第4章
特征推断与检验	均值检验	正态分布假设下单个或多个样本均值检验	参数检验：均值检验	均值比较（compare means）：单样本T检验、独立样本T检验、成对样本T检验、单因素方差分析	第5章
	方差检验		参数检验：Levene statistic（方差相等检验）	单因素方差分析（方差相等检验）	第5章
	分布形态检验	单样本下服从何种分布形态的检验	非参数检验（不依赖任何分布假设）	单样本K-S检验（正态、均匀、泊松、指数分布）、P-P图和Q-Q图、卡方检验	第6章
	分布函数与密度函数估计		非参数方法		
	分布列的检验	不同类别发生的概率检验	非参数检验（不依赖任何分布假设）	卡方检验（多类别的概率检验）、二项检验（两个类别下比例检验）	第6章
	某种特征（均值、中位数、众数）下数据随机产生的检验	均值、中位数、众数的检验	非参数检验（不依赖任何分布假设）	游程检验	第6章

（续表）

分析目的	分析内容	适用情形	分析方法	分析程序	本书章节
特征推断与检验	多样本数据随机产生的一致性检验	多样本特征值是否一致的检验	非参数检验（不依赖任何分布假设）	独立双样本、独立多样本、相关双样本、相关多样本的非参数检验	第6章
关系检验/数据的产生机制：有没有关系，关系多大，因变量数据的产生如何被决定	相关分析	分类变量之间的相关分析	相关系数计算及显著性检验	交叉表（crosstables），nonlinear canonical correlation analysis（optimal scaling）	第2章
		尺度变量之间相关分析	相关系数计算及显著性检验	相关分析（简单相关、偏相关）	第7章
		分类变量与尺度变量	因素方差分析	单因素方差分析、GLM（一般线性模型）	第3、8章
	关系分析：有没有关系，关系多大，因变量预测——模型构建	自变量因变量都是尺度变量（满足经典假设）	线性回归	线性回归	第9章
		因变量是尺度变量、自变量包含分类变量	线性回归（含随机效应）	GLM（一般线性模型）	第8章
		因变量是有序变量	回归，conjiont analysis（效用函数模拟，因变量为商品特征组合偏好）	有序因变量回归、广义线性模型、conjiont analysis	第10章
		因变量是二分变量，多分变量	回归、神经网络、判别分析、树分析	二元LOGIT、probit、多元LOGIT、广义线性模型、神经网络、判别分析	第10、11、12章
		因变量与自变量之间关系非线性	回归	曲线估计（curve estimation）	第9章
		因变量非正态分布：如泊松分布、伽玛分布等	回归	广义线性模型	第11章
		因变量为时间依赖事件	回归	COX回归、含时间依赖协变量的COX回归（cox w/ time-dep cov）	
		随机项违背经典假设	回归	加权最小二乘法、两阶段最小二乘法、方差分解（variance component）、线性混合模型（linear mixed models）	
		违背多重共性假设	回归	偏最小二乘法（partial least squares）	
		违背参数线性假定：非线性模型	回归	非线性回归	
		多因变量、重复测度因变量、多方程模型	回归，神经网络	GLM multivarite、repeated measures、GLM、AMOS、神经网络	

（续表）

分析目的	分析内容	适用情形	分析方法	分析程序	本书章节
		分类变量之间的关系分析	回归、scale、贝叶斯	对数线性（Loglinear），optimal scaling（分类回归等），scale（multidimensional unfolding，multidimensional scaling），naive bayes，	
		多个输入变量与一个或多个输出变量关系分析：不受变量类型限制（分类变量的类别最好不要太多）	神经网络	神经网络	第 12 章
		多个输入变量与一个因变量预测	树分析	树	
特征判别	样本分类、变量分类	基于样本多维特征根据相近性对样本进行分类；对变量的相近似进行分类	分类、判别分析	聚类分析（两步聚类、分层聚类、最近邻元素、K-均值聚类）、判别分析、神经网络、树、二元多元 LOGIT	第 13 章
	样本特征判别	对特征是否符合标准进行判别	质量控制（quality control）	图形控制（control charts）、帕累托图形（pareto charts）、ROC	第 14 章
	分类判别	对分类的有效性进行判别	ROC	ROC	第 14 章
数据问题	数据存在质量问题的分析	数据可靠性			第 15 章
		缺失数据	插入、后验估计	AMOS、missing values analysis、multiple imputation	第 16 章
		抽样数据			
		截取数据和截断数据			

1.5 本书安排

本书根据分析目的分成 5 大部分。第一部分为特征描述（第 2～4 章），涉及分析程序主要有描述统计程序、报告程序、均值程序和数据降维中的因子分析；第二部分为特征检验（第 5、6 章），涉及分析程序为比较均值程序和非参数检验；第三部分为关系分析（第 7～12 章），涉及程序有相关分析、回归分析、一般线性模型、广义模型和广义方程估计、神经网络等；第四部分为特征判别（第 13、14 章），涉及聚类分析、判别分析、控制图形与 ROC 等；第五部分为数据问题（第 15、16 章），涉及可靠性分析、缺失值分析和多重插入。

本书中所给出的案例数据均来自 SPSS 软件自带的数据，采用的是 SPSS 17.0 版本，软件操作说明均参照 SPSS 相关的英文版使用手册。若有一些引起歧义的地方请参照英文原版。本书的使用仅限用于教学。望读者在学习使用过程中购买正版软件，相关的正版软件公司会提供相应的技术支持和资料下载。

第一部分　特征描述

特征描述是指对研究对象单维特征、复合特征和综合特征描述。单维特征和复合特征的描述主要是采用图表的方式，显示频数分布，以及相关统计指标的计算。其涉及主要描述统计分析(含多重响应分析：多选项的描述统计)。综合特征的描述是综合研究对象多维特征，用一个指标或较少的几个隐含因子表示。前者是指数分析，后者是因子分析。因子分析是试图用几个潜在因子(一般归纳为典型特征指标)来表示研究对象所有的特征空间，其涉及分析主要是数据降维(dimension reduction)。

描述统计是根据样本数据对研究对象量的特征进行描述，包括单变量和复合变量的描述统计，并采用图表的形式显示统计结果。这是统计分析的首先工作。变量类型不一样，是分类变量还是尺度变量，相应的分析程序和指标也会不一样。复合变量的描述统计主要是分类变量之间的复合分组，分类变量与尺度变量之间的复合分组。这些统计涉及的 SPSS 程序主要有报告(reports)程序、描述统计(descriptive statistics)程序、均值(means)程序和多重响应(multiple response)程序。

描述统计程序的分析功能主要如下：(1) 频率(frequencies)程序，对名义变量、有序变量和尺度变量进行描述统计。分类变量的频率分析主要计算频数、频率和显示条形图及饼图，尺度变量的频率分析主要涉及分位数、集中趋势指标、离散趋势指标、分布形态指标和显示直方图。(2) 描述(descriptives)程序，只适用于尺度变量的描述统计，可以计算均值、离散趋势、分布形态指标，并能对数据进行标准化另存为变量。(3) 探索分析(explore)，主要适用于分类变量与尺度变量之间的描述统计，可以计算不同类别下尺度变量的描述性指标，通过箱图、茎叶图、直方图、P－P 图，Q－Q 图显示数据分布结构，并对是否服从正态分布进行检验。(4) 交叉表(又称列联表，crosstabs)程序，可以通过交叉表的方式对多个分类变量之间交叉频率或频数进行统计，对它们之间的关系进行显著性检验，计算二分变量之间的相对风险指标，分析相关有序变量之间数据的一致性。

报告程序中的 OLAP，涉及多个分类变量和尺度变量之间的描述统计，可以对多层分类下的尺度变量进行描述性统计。个案汇总(case summaries)程序与 OLAP 功能基本相同，区别在于它能显示个案。均值程序也是涉及一个或多个分类变量(作为自变量)与尺度变量(作为因变量)之间的描述统计，并对它们之间线性关系和非线性关系进行显著性检验，或进行单因素方差分析。

多重响应分析，首先通过定义多重响应集的方式，对多重响应变量进行描述统计。比率统计量，对两个尺度变量之间比率进行描述统计。

数据降维，是指研究对象是多维特征空间，用几个互不相关的典型特征维度表示。数据降维程序包括因子分析(factor analysis)、对应分析(correspondence analysis)和分类变量的主成份分析——最优尺度(optimal scaling)。对应分析和最优尺度涉及的是分类变量的数据降维。因篇幅所限，本书只介绍因子分析。