

实用统计分析方法与 SPSS 应用

(初稿)

张文璋

zhangwenzhang@sina.com.cn



二〇〇〇年十月

目录

详细目录	2
第一章 概论	5
第二章 SPSS 软件基础	12
第三章 统计数据的收集、整理与描述	34
第四章 总体与样本的描述	54
第五章 由样本推断总体	77
第六章 方差分析	100
第七章 相关分析	112
第八章 回归分析	121
第九章 含虚拟自变量的回归分析	178
第十章 Logistic 回归分析	186
第十一章 非参数检验	198
第十二章 聚类分析	221
第十三章 主成分分析	241
第十四章 因子分析	270
第十五章 事件史分析	298
附录一 Excel 在统计分析中的应用	308
附录二 常用统计表	367
参考文献	- 379 -

详细目录

详细目录	2
第一章 概论	5
第一节 市场经济呼唤统计学	5
第二节 统计学的研究对象及其学科分类	5
第三节 实用统计分析方法概述	8
第二章 SPSS 软件基础	12
第一节 统计分析软件简介	12
第二节 SPSS 简介	14
第三节 SPSS 基本操作	21
第三章 统计数据的收集、整理与描述	34
第一节 统计数据的来源	34
第二节 统计数据的收集	35
第三节 统计数据的整理	38
第四节 统计数据的描述	45
第五节 统计数据的探索性分析	51
第四章 总体与样本的描述	54

第一节	总体、样本与随机变量	54
第二节	总体与随机变量的描述	56
第三节	样本的描述	63
第四节	抽样分布——总体与样本的连接点	66
第五章	由样本推断总体	77
第一节	抽样	77
第二节	估计	81
第三节	检验	87
第六章	方差分析	100
第一节	单因素方差分析	100
第二节	多因素方差分析	107
第三节	案例:证券信息的定量分析	110
第七章	相关分析	112
第一节	简单相关分析	112
第二节	偏相关分析	115
第三节	其它相关系数分析	117
第八章	回归分析	121
第一节	一元线性回归分析	121
第二节	一元线性回归模型估计量的性质与分布	129
第三节	一元线性回归模型的检验	131
第四节	多元线性回归基本概念	135
第五节	多元线性回归模型的估计和检验	137
第六节	非线性回归与曲线回归	143
第七节	多重共线性	150
第八节	异方差	154
第九节	自相关	161
第十节	回归模型的应用	165
第十一节	案例 1:我国经济增长持续性的实证研究	167
第十二节	案例 2:中德人口老龄化水平之比较	170
第九章	含虚拟自变量的回归分析	178
第一节	虚拟变量回归模型的基本概念	178
第二节	包含一个质因素的虚拟变量模型	178
第三节	包含多个质的因素的虚拟变量模型	183
第四节	案例:虚拟变量在新股上市模型中的应用	183
第十章	Logistic 回归分析	186
第一节	Logistic 回归基本概念	186
第二节	Logistic 回归模型的估计与检验	187
第三节	案例:审计意见预测模型的构建	193
第十一章	非参数检验	198
第一节	非参数检验基本概念	198
第二节	非参数检验方法	199
第十二章	聚类分析	221
第一节	聚类分析概述	221
第二节	数据变换处理	223

第三节	聚类统计量.....	225
第四节	聚类方法.....	230
第五节	案例：汽车市场需求情况定量研究.....	236
第十三章	主成分分析.....	241
第一节	主成分分析的基本思想.....	241
第二节	总体主成分.....	243
第三节	样本主成分.....	247
第四节	案例：新兴股市的多因素模型.....	258
第十四章	因子分析.....	270
第一节	因子分析模型.....	270
第二节	因子分析模型估计方法.....	276
第三节	因子旋转.....	285
第四节	因子得分.....	288
第五节	案例：研究生院规模的因子分析.....	291
第十五章	事件史分析.....	298
第一节	事件史分析方法的源流.....	298
第二节	事件史分析方法的内容概述.....	299
第三节	事件史案例分析.....	303
附录一	Excel 在统计分析中的应用.....	308
第一节	中文 Excel 概述.....	308
第二节	Excel 基本操作.....	314
第三节	Excel 在描述统计中的应用.....	319
第四节	Excel 在推断统计中的应用.....	325
附录二	常用统计表.....	367
参考文献		- 379 -

第一章 概论

第一节 市场经济呼唤统计学

许多人简单地认为统计 (Statistics) 就是收集数字, 其实这仅仅是统计学的原始意义。现代统计学已远远超出了这个范围, 发展成为广泛应用于社会科学、自然科学等领域的科学方法。它是研究客观事物数量特征和数量关系的方法论学科, 能够告诉人们如何通过打开几扇窗口去探索一个未知的世界, 教会人们怎样用一种新的方式来思考问题, 是一门很实用的学科。

大千世界, 万事万物, 无一不具有它的质量、数量两个方面, 都是一定质量和数量的结合和表现。在对事物质的了解基础上, 从数量方面认识事物, 把握事物的数量方面, 做到胸中有数, 是对事物认识深化的具体表现。统计作为一种强有力的定量分析方法, 在社会、经济、政治、生活等领域得到了广泛的应用, 起着日益重要的作用。大至国家的宏观决策, 小至企事业单位的微观管理, 都离不开统计的应用。现代市场经济对统计信息的需求急剧增加, 对统计理论与方法提出了更高的要求。

面对二十一世纪, 我国的人文社会科学肩负着时代的重托。社会发展问题、经济可持续发展问题、国际竞争力问题、金融风险管理问题、保险精算问题、人口与社会保障问题、环境保护问题等等, 这些都迫切地等待着我们去深入地研究。要解决这些问题, 置身于古老东方文化氛围之中的中国学者需要冷静思考。时代要求我们必须抛开偏见, 正确理解与批判地吸收建立在发达商品经济基础上的外来文化, 加强数学方法、统计学方法的学习, 提高我们的定性分析与定量分析相结合的能力。这样, 中国人才会在新的世纪里大步赶上世界发达国家。

第二节 统计学的研究对象及其学科分类

一、统计学的研究对象

1992 年 11 月, 国家技术监督局正式批准统计学为一级学科, 国家标准局颁布的学科分类标准已将统计学列为一级学科, 1998 年教育部进行的专业调整也将统计学归入理学类一级学科。建设一级学科统计学的构想反映了统计学学科建设的内在要求, 符合国际统计学发展的大趋势。所谓一级学科统计学, 指的是研究搜集和分析数据、研究客观事物数量特征和数量关系的方法论科学。一级学科统计学首先是一门方法论, 它是研究客观现象 (包括自然现象和社会现象) 数量特征和数量关系、具有明确对象的方法论科学。统计方法论性质是指它作为一门认识方法论科学, 为人们提供一套从不确定的现象中探索现象规律性的理论和方法。这里作为统计学研究对象具体体现的“数据”, 是指进行各种统计 (指统计工作) 计算、科学研究或技术设计等所依据的数值。

统计数据所具有的不同特点, 使得统计学百花园色彩纷呈, 各具特色。数据中的实验数据主要来自自然技术现象, 如对产品配方检验得到的数据等等, 这类数据大多在可控条件下通过物理测量取得, 这类数据的搜集、整理工作并不复杂, 研究的重点在于数据分析。另一

类是观察数据,它主要来自社会经济现象,如国内生产总值(GDP)数据、某年度的货币购买力数据等等。由于社会经济现象的复杂性,尤其是不能通过一定条件下的物理或化学实验进行研究,致使观察数据的搜集往往十分困难,统计学不仅要研究观察数据的整理、分析技术,而且要花很大力气研究观察数据的调查搜集技术。正因为实验数据和观察数据有不同特点,所以以实验数据作为研究对象的自然技术统计学,如生物统计学、统计力学等等,和以观察数据作为研究对象的社会经济统计学,如农业统计学、工业统计学等等,就表现出很不相同的特点。社会经济统计学利用统计指标、统计分组方法,不厌其详地研究数据搜集的技术,研究资料来源、指标口径和计算方法,至于数据整理、尤其是数据分析的技术,则由于社会经济各专门统计的共同特点,出于简化篇幅的考虑,一般安排在社会经济统计学原理中作统一研究。自然技术统计学的生物统计学等等,与社会经济统计学的农、工业统计学则恰恰相反,它的研究重点往往放在对数据所作的各种分析上,至于数据搜集、整理的技术,则考虑到自然技术各专门统计所具有的共同特点,一般放到作为自然技术统计学原理的数理统计学中作简要讨论(之所以往往仅作简要讨论,是因为实验数据的搜集和整理远比观察数据的搜集整理简单)。从上面的分析中不难看出,自然技术统计学和社会经济统计学本没有不可逾越的鸿沟,两者只是由于研究对象所具有的不同特点,才产生了不同的理论体系和学科特色。建设一级学科统计学的构想,兼容自然技术统计学与社会经济统计学,反映了统计学发展的内在要求,对促进自然技术统计学和社会经济统计学各自的发展,都具有重要的意义。

二、统计学的学科分类

统计学作为一门研究客观事物数量特征和数量关系的方法论科学,其内容构成错综复杂,既有层次性,又有交叉性,所以对其学科的分类迄今未得到合理的解决。较为流行的划分是把统计学分为社会经济统计学和数理统计学,或者分为描述统计与推断统计。这些分类都无法完全包括现代意义上的统计学内容,是不妥当的。与一级统计学相对应,我们把统计学分为理论统计学、应用统计学、与其他统计学等(如图 1-1 所示)。

理论统计学包括各种统计基础理论,又可以分为描述统计学和推断统计学。描述统计学指以总体全面资料或非随机性局部资料为基础的统计理论与方法体系,包括统计总体论(有关总体、指标和分组等理论)、统计设计、统计调查、统计整理、统计指数、动态分析理论、统计平衡理论、统计数据库等等,不同于仅研究如何整理和概括大量数据的“描述统计学”。推断统计学指依据随机样本推断总体特征的理论与方法体系,也就是数理统计学,又可以分为理论数理统计学和应用数理统计学。理论数理统计学侧重于统计方法的数理基础,包括概率论、经典统计理论、贝叶斯理论、统计判决理论等。应用数理统计学(现代意义上的数理统计学)则侧重于统计方法的应用形式,包括抽样技术、试验设计、相关分析、方差分析、多重应答分析、多元统计分析、序贯分析、线性统计模型、时间序列分析、非参数统计等。这里的描述统计学与推断统计学并无“普通统计学”与“高级统计学”之分,实际上,推断统计学的某些内容是非常初等的,而描述统计学中的某些方法(如统计指数理论)却具有相当的理论深度和复杂性。

应用统计学只涉及某一特定现象领域的统计研究,又可以分为核算统计学和实验统计学。核算统计学是通过核算手段研究社会现象及其过程的数量特征或统计规律性的理论与方法体系,包括经济统计学、社会统计学、科技统计学、环境统计学等等。而实验统计学是运用实验手段研究自然现象自身及其过程的数量特征或统计规律性的理论与方法体系,包括统计物理学、生物统计学、天文统计学、气象统计学、心理统计学、农业试验统计学、工程技术统计学等等。

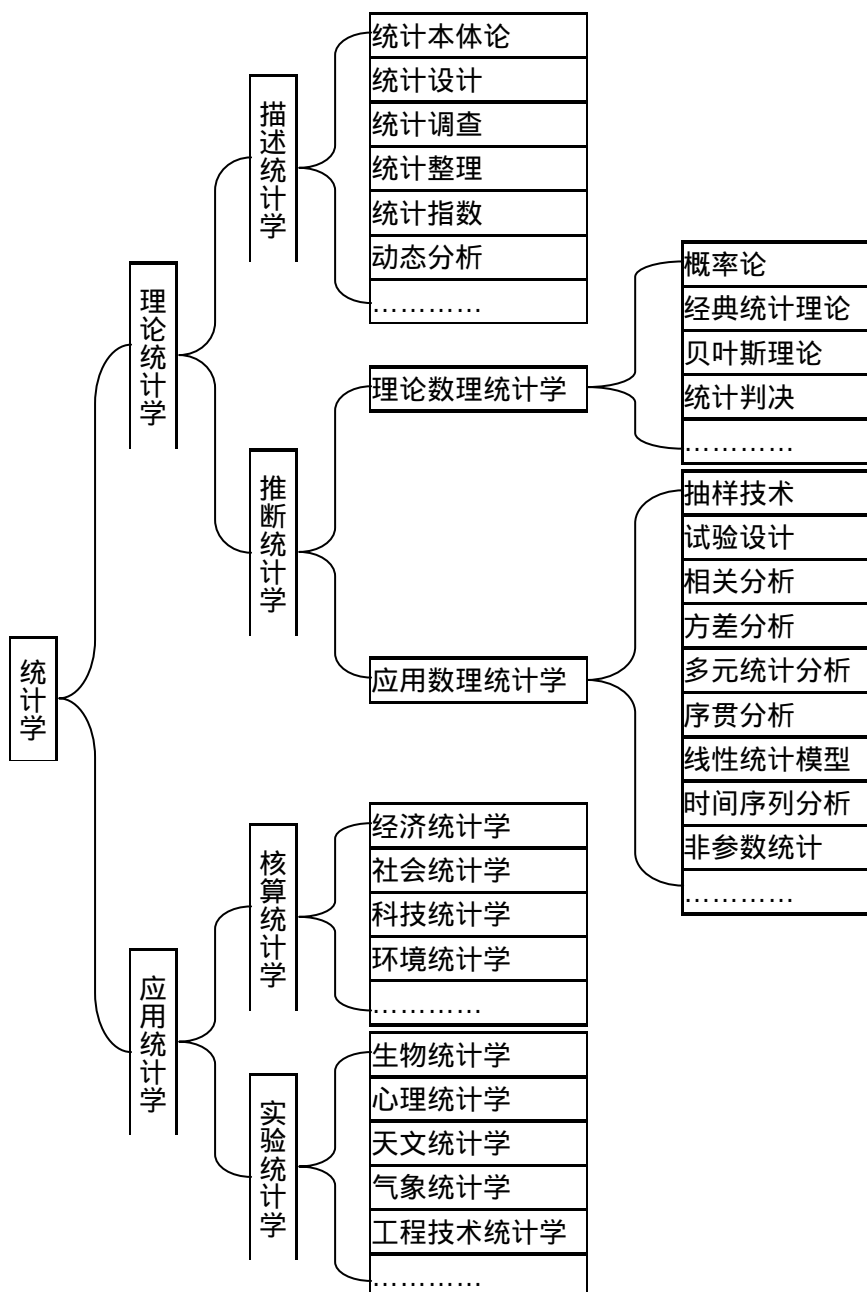
杨灿:《统计学基本问题研究》,《统计研究》,1993 年第 3 期;

黄良文、黄沂木:《大学科统计刍议》,《统计研究》,1995 年。

除了理论统计学和应用统计学外，还有统计史学、统计法学、比较统计学等其他统计学科，以及经济计量学、保险精算学、运筹学、信息论等边缘学科。

从统计学的学科分类可以看出，统计学的内容是十分丰富的，其研究和应用的领域非常广泛。本书主要是为非统计专业的学生和统计工作者提供一本关于实用统计分析方法的读物，所以，主要包括了应用数理统计的一些内容。本书强调统计分析方法的基本思想和应用条件，培养用计算机进行统计计算的能力，并希望通过案例分析提高学生的解决实际问题的能力。

图 1-1 统计学分类



第三节 实用统计分析方法概述

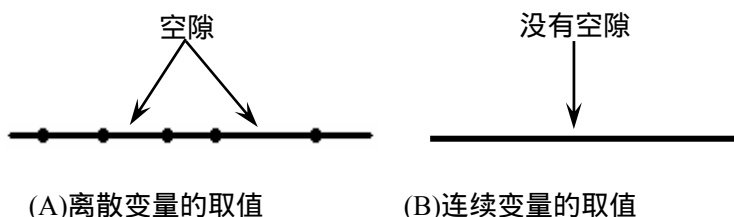
一、变量(Variable)的分类

要进行统计分析，离不开统计数据。在搜索数据之前，必须首先了解数据的种类。数据涉及到变量的取值，通常用变量的取值来描述数据。变量可按多种方法分类，这些分类有助于选择适当的统计分析方法作进一步的分析与研究。下面按三种方法对变量进行分类：按间隙分类、按作用分类和按测量尺度分类。

(一) 按间隙(gaps)划分

根据一个变量紧挨着的两个观测值之间是否有空隙（缺口），可以把变量分为两类：离散型变量(discrete variable)和连续型变量(continuous variable)。如果一个变量的观测值之间有空隙，该变量称为离散型变量，否则称为连续型变量，如图 1-2 (A)所示。更准确地说，当一个变量的任意两个可能取值之间没有其他取值时，该变量是离散的；当一个变量的任意两个可能取值之间还有其他可能取值时，该变量是连续的。例如，性别(设男性取值为 0，女性取值为 1)、企业数目、分组情况（设 A 组取值为 1，B 组取值为 2 等）等为离散型变量；身高、体重、血压、GDP 等为连续型变量。

图 1-2 离散型变量与连续型变量



需要指出的是，由于分析的需要，离散型变量经常作为连续型变量处理。而连续型变量也可以作为离散型变量处理，如可以把“血压”变量分为“低”、“中”、“高”三组变为离散型变量。

(二) 按作用划分

根据一个变量在分析时的作用，可以把变量分为因变量(dependent variable)或自变量(independent variable)。如果一个变量由其他变量来描述，该变量称为因变量或反应变量(response variable)；如果一个变量与其他变量一起用于描述因变量，该变量称为自变量或预测变量(predictor variable)。例如，在分析家庭收入、性别等因素对消费支出的影响时，收入变量和性别变量是自变量，消费支出变量是因变量。

一个变量是因变量还是自变量，与统计分析的目的有关。同一个变量在某种分析中作为因变量，而在其它分析中可能作为自变量。

(三) 根据测量尺度划分

根据变量测量精度不同，可把变量由低到高分四种尺度：定类变量、定序变量、定距变量和定比变量。

1、定类变量

定类变量又称为名义(nominal)变量。这是一种测量精确度最低、最粗略的基于“质”因素的变量，它的取值只代表观测对象的不同类别，例如“性别”变量、“职业”变量等都是定类变量。定类变量的取值称为定类数据或名义数据。定类数据的共同特点是用不多的名称来加以表达，并由被研究变量每一组出现的次数及其总计数所组成，这种数据是枚举性的，即由计数一一而得。唯一适合于定类数据的数学关系是“等价关系”。因而，在定类数据中，同一组内各单位是等价的，同时若更换各不同组的符号并不会改变数据原有的基本信息。因

此，最常用来综合定类数据的统计量是频数、比率或百分比等。

2、定序变量

定序变量又称为有序(ordinal)变量、顺序变量，它的取值的大小能够表示观测对象的某种顺序关系（等级、方位或大小等），也是基于“质”因素的变量。例如，“最高学历”变量的取值是：1—小学及以下、2—初中、3—高中、中专、技校、4—大学专科、5—大学本科、6—研究生以上。由小到大的取值能够代表学历由低到高。定序变量的取值称为定序数据或有序数据。适合于定序数据的数学关系是“大于(>)”和“小于(<)”关系。在定序数据中，同一组内各单位是等价的，相邻组之间的单位是不等价的，它们存在“大于”或“小于”的关系。而且，并进行保序变换（或称单调变换），则不改变数据原有的基本信息即等级顺序。最适合用于综合定序数据取值的集中趋势的统计量是中位数。

3、定距变量

定距变量又称为间隔(interval)变量，它的取值之间可以比较大小，可以用加减法计算出差异的大小。例如，“年龄”变量，其取值 60 与 20 相比，表示 60 岁比 20 岁大，并且可以计算出大 40 岁（60-20）。定距变量的取值称为定距数据或间隔数据。定距数据是一些真实的数值，具有公共的、不变的测定单位，可以进行加减乘除运算。定距数据的基本特点是两个相同间隔的数值的差异相等，例如，年龄的 60 岁与 50 岁之差等于 40 岁与 30 岁之差。对于定距数据，不仅可以规定“等价关系”以及“大于关系”和“小于关系”，而且也可以规定任意两个相同间隔的比值或差值。如果将每个数值分别乘以一个正的常数再加上一个常数，即进行正线性变换，并不影响定距数据原有的基本信息。因此，常用的统计量如均值、标准差、相关系数等都可直接用于定距数据。

4、定比变量

定比变量又称为比率(ratio)变量，它与定距变量意义相近，细微差别在于定距变量中的“0”值只表示某一取值，不表示“没有”。例如，人的身高就是一个定比变量，如果身高值为“0”米，则表示这个人不存在。而定比变量的“0”值表示“没有”。而在测定温度的摄氏表中，0℃并不表示没有温度，因为还有在零点以下的温度。定比变量的取值称为定比数据或比率数据。定比数据也同样可进行算术运算和线性变换等。通常对定距变量和定比变量不需再加以区别，两者统称为定距变量或间隔变量。

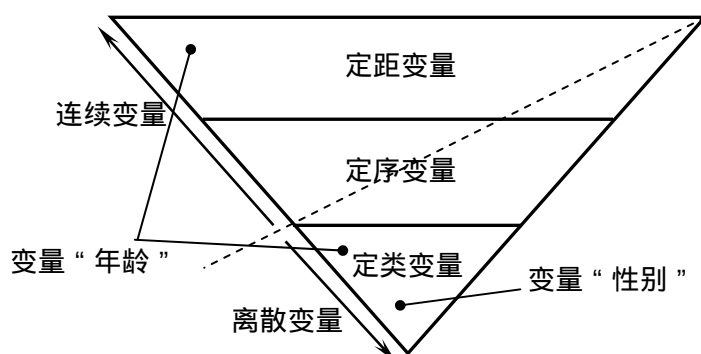
一般地，定类变量和定序变量用于描述定性数据，属于定性变量；而定距变量和定比变量用于描述定量数据，属于定量变量。

同其他分类标准一样，一个变量在不同分析中可当作不同尺度的变量。例如，“年龄”在某些分析中（如回归分析）当作定距变量，而在另外一些分析中（如方差分析）可通过分组作为定类变量处理。

另外，较高尺度的变量包含了较低尺度变量的性质。定序变量包含了定类变量的所有特征，定距变量同时包含了定序变量和定类变量的特征。这种性质允许在分析数据时把一些较高尺度变量作为较低尺度变量处理。例如，定距变量可当作定类变量或定序变量看待，而定序变量可作为定序变量分析。

以上通过三种不同方法对变量进行分类。这些分类是可以重叠的。一个变量可能是离散型变量、自变量、定类变量（如“最高学历”），也可能是连续型变量、因变量、定距变量（如“血压”）。按间隔分类和按测量尺度分类的重叠如图 1-3 所示。

图 1-3 变量分类的重叠



因为自变量与因变量是根据分析目的而不是按变量本身性质来划分的，所以图 1-3 中没有包括这种分类。从图 1-3 中可以看出，定类变量必须是离散变量，而定距变量和定序变量可以是离散变量或连续变量；连续变量必须是定序变量或定距变量。例如，变量“性别”是离散变量又是定类变量；变量“年龄”可当作定距变量、连续变量，也可以作为定类变量、离散变量。

二、统计分析方法的分类与选择

对数据进行统计分析时，选择正确的分析方法是非常重要的。选择统计分析方法时，必须考虑许多因素，主要有：(1) 统计分析的目的，(2) 所用变量的特征，(3) 对变量所作的假定，(4) 数据的收集方法（即抽样过程）。选择统计分析方法时一般考虑前两个因素就足够了。

（一）根据统计分析目的不同进行分类

统计分析方法根据统计分析目的的不同，可以分成四大类：相关分析方法、结构简化方法、分类分析方法、预测决策方法。

（二）根据变量特征的不同进行分类

根据变量的分类不同分类方法，把变量分为因变量、自变量以及定量变量、定性变量，可把统计分析方法一一进行归类（如表 1-1 所示），这是正确选择统计分析方法的一种有效方法。

表 1-1 统计分析方法分类表

变量类型		统计分析方法	统计分析目的
因变量	自变量		
定量	定量	回归分析(或线性模型)、相关分析	描述一个或多个自变量与一个因变量之间的因果依存关系,或变量之间的相关关系。
定量	定性	<i>T</i> 检验、方差分析	描述一个连续型因变量与一个或多个定类自变量之间的关系。
定量	定性、定量	协方差分析(或线性模型)	描述在控制了一个或多个连续型自变量的影响下一个连续因变量与一个或多个定类自变量之间的关系。
定性	定性	列联分析,Logit 模型	描述定性变量之间的相互影响关系。
定性	定量	Logistic 回归分析、判别分析、聚类分析	描述多个定量变量与定性变量之间的依赖关系。
定性	定性、定量	对数线性模型	描述定性或定量变量与分类变量之间的关系。
定性、定量	定性、定量	/	/
相依模型		主成分分析、因子分析、对应分析等。	描述变量、样品或类型之间的结构关系。

第二章 SPSS 软件基础

第一节 统计分析软件简介

进行统计分析时,涉及到的变量和样本数据很多,计算量很大。靠手工方法进行统计计算是不现实的,不借助于计算机往往难以实现,只有计算机才能快速得到精确的结果。在微机使用的统计软件有许多种,在实际工作中应用比较普遍的主要有 SPSS、SAS、TSP、EViews、BMDP、TPL、CENTS、DET、SP、SARP、Excel、Lotus 1-2-3、Matlab、S-Plus、Minitab 等,为帮助读者了解选择和使用这些软件,我们在此作一简单介绍,具体的应用技术和操作方法请参阅相应的软件说明书及有关书籍。

(一) SAS 统计分析系统

SAS(Statistical Analysis System)软件是为处理数据而研制的大型统计分析系统,是融数据管理和统计分析于一体,由多个子软件构成的一个大型软件。该软件 1972 年由美国 SAS 软件研究所推入市场以来,经过不断完善,已成为当今世界上最有影响的统计分析系统之一,它具有完备的数据访问、数据管理、数据分析以及数据呈现能力。其中,强大的数据分析能力是使 SAS 成为业界著名应用软件的重要因素。SAS 支持多种软硬件平台,广泛地运行在各种型号的大、中、小型机和微型计算机上。SAS 系统中提供的主要分析功能包括统计分析、经济计量分析、时间序列分析、决策分析、财务分析和全面质量管理工具等等。

SAS 系统是一个组合软件系统,它由多个功能模块组合而成,其基本部分是 BASE SAS 模块。BASE SAS 模块是 SAS 系统的核心,承担着主要的数据管理任务,并管理用户使用环境,进行用户语言的处理,调用其他 SAS 模块和产品。也就是说,SAS 系统的运行,首先必须启动 BASE SAS 模块,它除了本身所具有数据管理、程序设计及描述统计计算功能以外,还是 SAS 系统的中央调度室。它除可单独存在外,也可与其他产品或模块共同构成一个完整的系统。各模块的安装及更新都可通过其安装程序非常方便地进行。SAS 系统具有灵活的功能扩展接口和强大的功能模块,在 BASE SAS 的基础上,还可以增加如下不同的模块而增加不同的功能:SAS/STAT(统计分析模块)、SAS/GRAPH(绘图模块)、SAS/QC(质量控制模块)、SAS/ETS(经济计量学和时间序列分析模块)、SAS/OR(运筹学模块)、SAS/IML(交互式矩阵程序设计语言模块)、SAS/FSP(快速数据处理的交互式菜单系统模块)、SAS/AF(交互式全屏幕软件应用系统模块)等等。

在统计功能方面(SAS/STAT),SAS 可以完成以下任务:

- (1) 方差分析:单因素、多因素方差分析和单变量、多变量方差分析。
- (2) 离散型数据的分析:二维列表分析、分层分析、对数线性模型、Logistic 模型。
- (3) 回归分析:多元线性回归、多项式回归、逐步回归、非线性回归、正交回归等。
- (4) 生成分析:生命表及 Cox 回归模型。
- (5) 时间序列分析。
- (6) 多元统计分析:相关分析、样品聚类、变量聚类、判别分析、因子分析、对应分析。
- (7) 一般线性模型。

SAS 有一个智能型绘图系统,不仅能绘各种统计图,还能绘出地图。

SAS 提供多个统计过程,每个过程均含有极丰富的任选项。用户还可以通过对数据集的一连串加工,实现更为复杂的统计分析。此外,SAS 还提供了各类概率分析函数、分位

数函数、样本统计函数和随机数生成函数，使用户能方便地实现特殊统计要求。

SAS 提供两种非交互式运行方式（批处理方式、程序方式）和两种交互式（命令行方式、菜单方式），以适应不同的应用场合和不同层次的使用者。非交互式适用于大批量、经济性统计分析和用户应用系统。交互方式则适用于临时性统计分析和程序调试。其中菜单方式只需用户在屏幕上显示的程序框架中填入合适的参数，尤其适于不熟悉 SAS 的使用者。SAS 多窗口技术提供多种系统定义窗口，使运行情况一目了然。此外，用户还可自己定义各种窗口，使用户研制的系统更为方便、“友善”。

SAS 的通讯功能允许用户与主机进行数据及程序交换，可实现 SAS 数据文件与 SQL Server、Access、Excel 等互相交换数据。

SAS 系统简单易学、使用方便、即使是没有编程经验甚至不太熟悉计算机的用户，也可以在很短的时间内学会使用 SAS 系统作基本的数据分析和统计工作。对统计人员来说，SAS 系统是一个得心应手的工具，所有的工作都可以在本系统内完成，而不象有的统计软件那样，需要先在一个系统内作数据管理工作，再在另一个系统内作数据分析和统计工作，从而简化了处理过程。

最近 SAS 软件研究所又发布了 SAS 系统 8.2 新版本。与以往的版本比较，8.2 版的 SAS 系统除在功能和性能方面得到增加和提高外，GUI 界面也进一步加强。SAS 通过对 ODBC、COLE 和 MailAPIs 等业界标准的支持，大大加强了 SAS 系统和其它软件厂商的应用系统之间相互操作的能力，为各应用系统之间的信息共享和交流奠定了坚实的基础。有关 SAS 系统的最新动态新参见 SAS 主页 <http://www.sas.com>。

(二) Micro TSP 时间序列软件包

Micro TSP(Time Series Processor)是原国家教委所推荐的功能强大的经济计量分析软件，主要用于时间序列分析，当然也包括了基本的统计运算。其主要功能为：(1)基本统计运算，如平均数、方差、标准差等。(2)相关分析。(3)回归分析，包括简单回归分析和多元回归分析。(4)统计预测，即根据回归模型进行历史外推。(5)季节数据整理。(6)ARMA 模型的建立。(7)文件及数据管理。(8)统计图形与图形打印。(9)联立方程模型估计求解功能。该软件可采用对话式操作，也可用命令编程运行。Micro TSP for DOS 的最高版本为 V6.53，其 Windows 版改名为 EViews。

(三) EViews 软件

EViews 是 Econometric Views(经济计量视图)的缩写，为 Micro TSP 的 Windows 版本。EViews 充分利用 Windows 操作系统的强大功能，引入了全新的面向对象概念，通过操作对象实现各种分析功能。EViews 提供了在运行 Windows 的微机上进行复杂的数据分析、回归和预测的强大工具。用 EViews 可以快速地建立起数据间的统计模型，并用此模型进行预测。EViews 的版本有 V1.0、V2.0、V3.0、V3.1 和 V4.0 等。有关 EViews 软件的最新动态见 <http://www.eviews.com>。

(四) MiniTab for Windows

MiniTab for Windows 统计软件比 SAS、SPSS 等小得多，但其功能并不弱，特别是它的试验设计及质量控制等功能。MiniTab 提供了对存储在二维工作表中的数据进行分析的多种功能，包括：基本统计分析、回归分析、方差分析、多元分析、非参数分析、时间序列分析、试验设计、质量控制、模拟、绘制高质量三维图形等。可在其主页(<http://www.minitab.com>)上查询最新动态或下载 30 天全功能试用版。

(五) NCSS 2000 for Windows

NCSS for Windows 是一个十分优秀的统计软件，其界面友好，功能齐全。其主要功能有：描述性统计、相关及回归分析、试验设计、质量控制、生存及可靠性分析、多元分析、

时间序列分析及预测、统计图表绘制等。其主页 (<http://www.ncss.com>) 上有全功能 30 天试用版可下载。

(六) DPS For Windows

这是一款国产的数据处理软件,除了输出结果较为简单外,其功能十分齐全,是一个“通用多功能数理统计和数学模型处理软件”。它是用 Delphi 开发的,采用 TideStone 公司的 FormulaOne 控件作为其电子表格。与国外同类专业统计分析软件(如 SAS、SPSS、STAT、STATISTICA 等)相比,DPS 系统是独特的,它在使用时不必拘泥于一般电子表格的行列规定,行和列由系统辨认。DPS 在统计分析及模型模拟方面功能齐全,易于掌握,尤其是对广大中国用户。其配套书实际上是一本难得的统计分析资料,因为书中对各种统计过程的原理都作了较深入的介绍。

(七)其它统计分析软件其网址。

Statistica, <http://www.statsoft.com>

BMDP, <http://www.spss.com>

SYSTAT, <http://www.spss.com>

StatMost, <http://www.datamost.com>

Stata, <http://www.stata.com>

S-Plus, <http://www.mathsoft.com/splus>

SimStat, <http://www.simstat.com>

SHAZAM, <http://shazam.econ.ubc.ca>

DataDesk, <http://www.datadesk.com/datadesk>

Matlab, <http://www.matlab.com>

第二节 SPSS 简介

一、SPSS 概述

SPSS 是英文 Statistical Package for the Social Science (社会科学统计软件包)的缩写。20 世纪 60 年代末,美国斯坦福大学的三位研究生研制开发了最早的统计分析软件 SPSS,同时成立了 SPSS 公司,并于 1975 年在芝加哥组建了 SPSS 总部。20 世纪 80 年代以前,SPSS 统计软件主要应用于企事业单位。1984 年 SPSS 总部首先推出了世界第一个统计分析软件微机版本 SPSS/PC+,开创了 SPSS 微机系列产品的开发方向,极大地扩充了它的应用范围,并使其能很快地应用于自然科学、技术科学、社会科学的各个领域,世界上许多有影响的报刊杂志纷纷就 SPSS 的自动统计绘图、数据的深入分析、使用方便、功能齐全等方面给予了高度的评价与称赞。SPSS 名为社会科学统计软件包,这是为了强调其在社会科学应用的一面(因为社会科学研究中的许多现象都是随机的,要使用统计学来进行研究),而实际上广泛应用于经济学、社会学、生物学、教育学、心理学、医学以及体育、工业、农业、林业、商业和金融等各个领域。

SPSS 现已推广到多种各种操作系统的计算机上,它和 SAS、BMDP 并称为国际上最有影响的三大统计软件。和国际上几种统计分析软件比较,它的优越性更加突出。在众多用户

对国际常用统计软件 SAS、BMDP、GLIM、GENSTAT、EPILOG、MiniTab 的总体印象分的统计中，其诸项功能均获得最高分。在国际学术界有条不紊的规定，即在国际学术交流中，凡是用 SPSS 软件完成的计算和统计分析，可以不必说明算法，由此可见其影响之大和信誉之高。

SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等等。SPSS 统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等几大类，每类中又分好几个统计过程，比如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、两阶段最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 也有专门的绘图系统，可以根据数据绘制各种图形。

二、SPSS for Windows 的不同版本

到目前为止，SPSS 已具有适合于 DOS、Windows、Unix、Macintosh、OS/2 等多种操作系统使用的产品，国内常用的是其 DOS 和 Windows 版本。SPSS for DOS 通常称为 SPSS/PC+，现已较少使用。由于 SPSS for Windows 界面友好，功能强大，使用者越来越多。SPSS for Windows 的主要版本有 SPSS V7.0、SPSS V7.5、SPSS V8.0、SPSS V9.0、SPSS V10.0、SPSS V11.0 等，SPSS V10.0 以上有服务器(Server)与本地(Local)/客户版本之分。SPSS 各个版本的主要新增功能如表 1-2 所示。表 1-2 SPSS for Windows 不同版本新增特性

汤旦林、王松柏：《几种国际通用统计软件的比较》，《数理统计与管理》，1996.1。

参见 SPSS 帮助文件和 SPSS 公司的网址（<http://www.spss.com/>）。

SPSS 版本	主要新增特性
SPSS V7.0	1、充分利用了 Windows 95 的强大功能,提供了新的输出界面和灵活的帮助; 2、添加了 Summarize 和 GLM(一般线性模型)等统计分析过程。
SPSS V7.5	1、首次加入了脚本引擎,可以采用与 Visual Basic 完全兼容的 Sax Basic 语言编写脚本程序定制输出或自动运行某些任务; 2、可以把输出结果以 HTML 文件格式导出; 3、新增了 Statistics Coach(统计教练)帮助新用户选择合适的统计分析过程以及 Variance Components Analysis 等。
SPSS V8.0	1、新增了动态的交互式图表; 2、增强了方差分析、探索性分析、均值分析、可靠性分析、生存分析、回归分析等过程的功能; 3、增强了输出结果管理、数据管理和帮助系统的功能。
SPSS V9.0	1、对界面作了一些改动,如用[Analyze(分析)]菜单项代替以前各版本的[Statistics(统计)]等; 2、添加了多种交互式图表类型; 3、增强了可靠性分析、交叉表分析、回归分析和 ROC 曲线过程等; 4、提供了新的文件管理。
SPSS V10.0	1、新的数据管理功能允许对大数据文件进行分析,减少了分析时间和所需的临空间; 2、新的数据编辑器使得数据的录入、检查、组织更为方便; 3、简化了与 SQL 数据库、Excel 等的交互,可以直接分析 Excel 文件中的数据,并支持最新的 XML 文件格式; 4、增强了图表和输出结果管理等功能; 5、新增了一些统计过程,如非线性主成分分析、PLUM 等; 6、改进了 Logistic 回归分析和 Cox 回归分析的输出; 7、增加了分布式分析,即把数据提交给 SPSS 10.0 的服务器版本进行分析,大大提高了效率。

本书以运行于 Windows 9X/NT/2000 上的 SPSS 10.0 for Windows 本地版本为例,并简称 SPSS。

三、SPSS 的运行环境

(一) SPSS 的硬件环境

能运行 Windows 95/NT/2000 或以上版本的微机。

(二) SPSS 的软件环境

目前 SPSS 还没有简体中文版,SPSS 能在中英文 Windows 9X、Windows NT 4.0、Windows 2000 及更高版本的 Windows 操作系统上运行。

四、SPSS 的安装

如果您的计算机中没有安装 SPSS,则按下列步骤进行 SPSS 的安装:

1、启动 Windows 后,把 SPSS 系统安装软盘(或光盘)插入软驱(或光驱),并找到 SPSS 的安装程序的可执行文件 Setup.exe。

- 2、双击 Setup.exe 文件,安装程序向导将给出每一步操作的提示。在出现[Welcome(欢迎)]窗口后,选择[Next]进入下一步。
- 3、安装程序显示[Software License Agreement]对话框时,选择[Yes]接受显示的协议条款。
- 4、选择把 SPSS 安装到哪个文件夹(目录),默认文件夹为程序文件目录下的 SPSS 目录(如“ C:\Program Files\SPSS ”)。如果要改变安装目录,按[Browse]选择新的目录。然后单击[Next]按钮。
- 5、在[User Information]窗口中输入[Name(姓名)]、 [Organization(组织单位)]、 [Serial Number (产品序列号)] (如图 2-1 所示), 然后单击[Next]按钮。

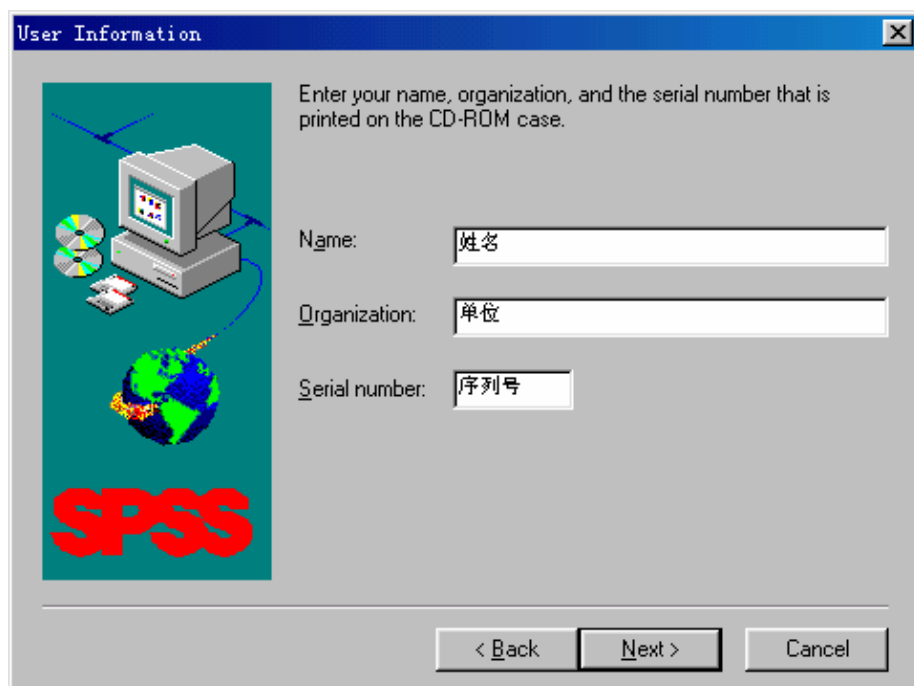


图 2-1

- 6、根据需要选择安装类型: [Typical(典型安装)]、 [Compact(最小安装)]、 [Custom(定制安装)] (如图 2-2 所示)。这里假设要进行定制安装,所以选择[Custom]。按[Next]进入下一步。

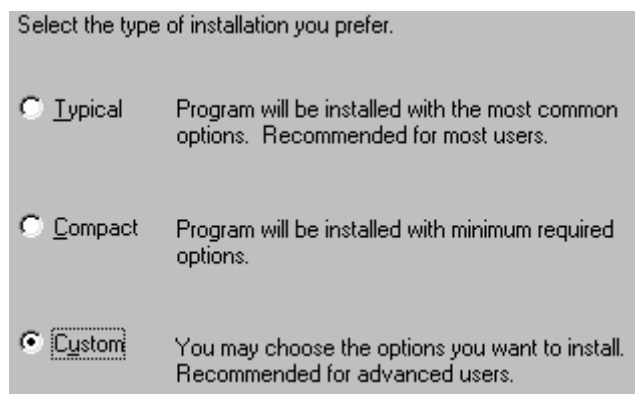


图 2-2

- 7、选择要安装的部件 (如图 2-3 所示)。

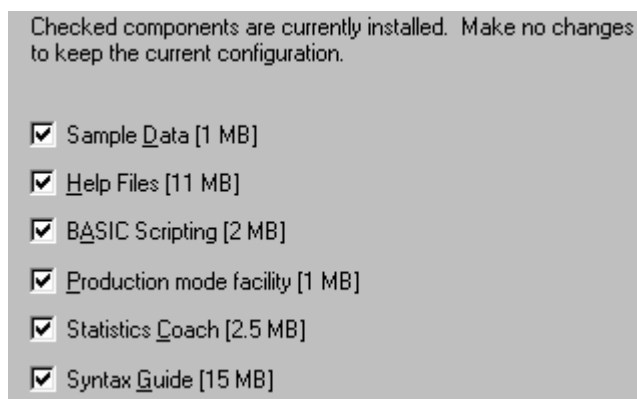


图 2-3

8、根据授权情况选择个人安装或共享安装（如图 2-4 所示）。

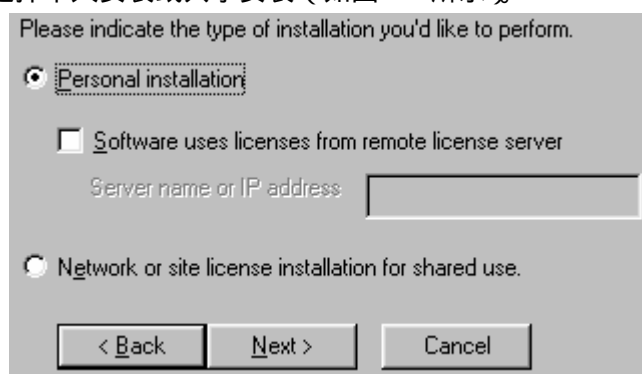


图 2-4

9、输入许可证号（如图 2-5 所示），单击[Next]。

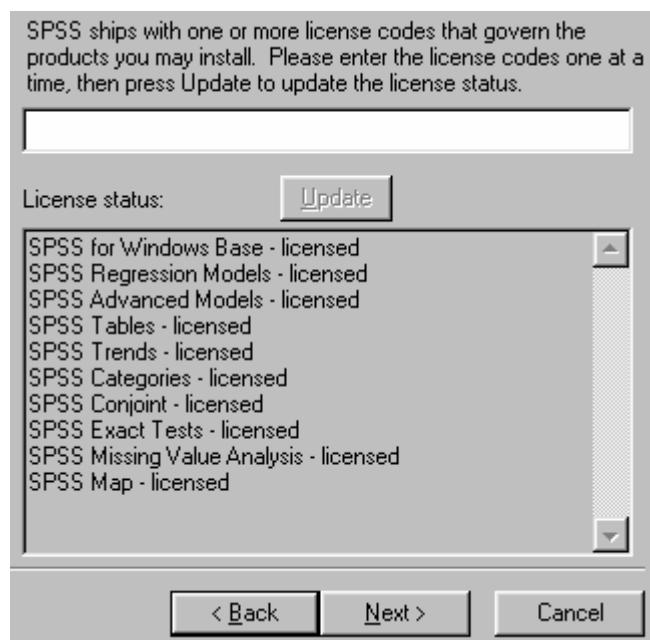


图 2-5

10、选择要安装的模块（如图 2-6 所示），然后单击[Next]。系统将根据用户的选择安装有关文件。

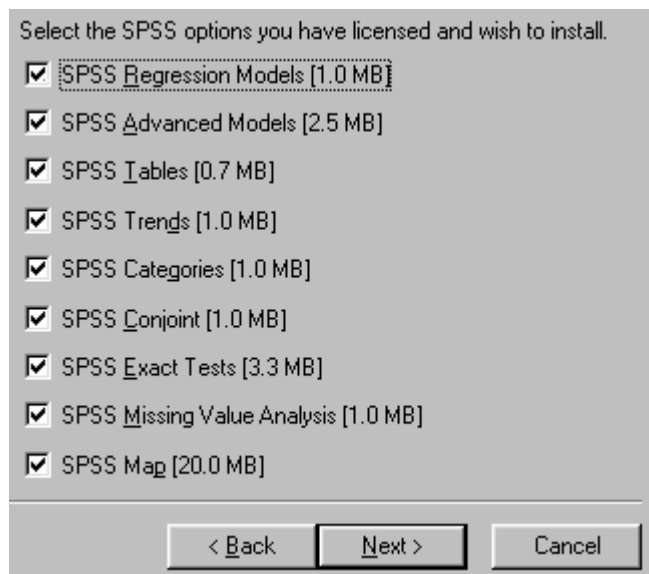


图 2-6

11、安装完文件后，SPSS 显示如图 2-7 所示的对话框。SPSS 提供了 SPSS 命令语法说明文件，这些文件是以 PDF 格式保存的，如果要阅读这些文件，就必须安装 Adobe 公司的 Acrobat Reader 软件。这里，选择[Do not reinstall Adobe Acrobat Reader]，单击[Next]。

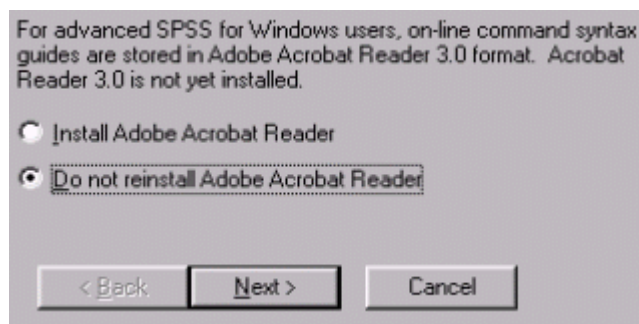


图 2-7

12、SPSS 安装程序显示如图 2-8 的对话框，如果不重装 ODBC 驱动程序，选择[Do not reinstall ODBC]，然后单击[Next]。

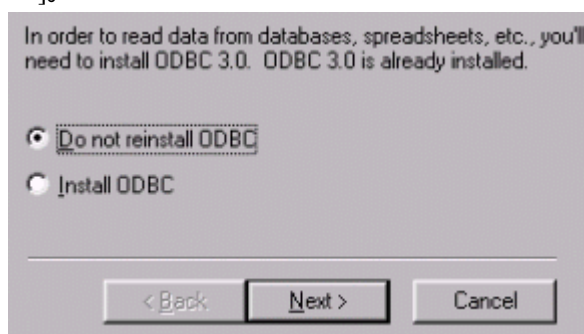


图 2-8

13、安装程序显示如图 2-9 所示的对话框，表明 SPSS 安装成功。图 2-9 的对话框中有两个选项[Launch tutorial now? (单击[Finish]后马上启动 SPSS 教程)]和[Display the ReadMe file now? (马上显示 SPSS 自述文件吗?)]。单击[Finish]按钮结束 SPSS 安装过程。

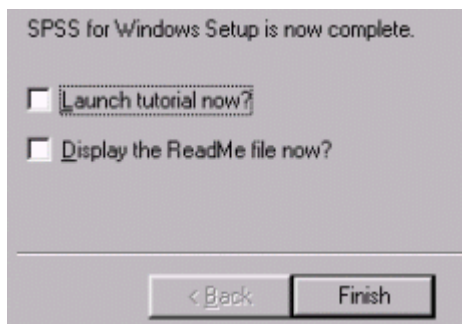


图 2-9

五、SPSS 的运行方式

SPSS 运行方式灵活，主要有四种方式：

（一）批处理方式

把已编写好的程序（语句程序）存为一个文件，提交给[开始]菜单上[Spss for Windows]=>[Production Facility]程序运行。

（二）完全窗口菜单运行方式

这种方式通过选择窗口菜单和对话框完成各种操作。用户无须学会编程，简单易用。

（三）程序运行方式

这种方式是在语句(Syntax)窗口中直接运行编写好的程序或者在脚本(Script)窗口中运行脚本程序的一种运行方式。这种方式要求掌握 SPSS 的语句或脚本语言。

（四）混合运行方式

混合运行方式指以上各种方法的结合方式。

本书采用“完全窗口菜单运行方式”。

六、SPSS 的启动、主界面和退出

（一）启动 SPSS

单击 Windows 的[开始]按钮（如图 2-10所示），在[程序]菜单项[SPSS for Windows]中找到[SPSS 10.0 for Windows]并单击。



图 2-10 启动 SPSS

（二）SPSS 的主界面

启动 SPSS 后，出现 SPSS 主界面（数据编辑器）。同大多数 Windows 程序一样，SPSS 是以菜单驱动的。多数功能通过从菜单中选择完成。主菜单包括十个菜单项（所图 2-11所示）：

File：“文件”菜单用于新建 SPSS 各种类型文件，打开一个已存在的文件，从文本文件

或其它数据源读入数据。

Edit：“编辑”菜单用于撤消操作、剪切、复制、粘贴、查找、改变 SPSS 默认设置等。

View：运用“视图”菜单显示或隐藏状态行、工具栏、网络线、值标签和改变字体。

Data：运用“数据”菜单对 SPSS 数据文件进行全局变化，例如定义变量，合并文件，转置变量和记录，或产生分析的观测值子集等。

Transform：“转换”菜单在数据文件中所选择的变量进行变换，并在已有变量值的基础上计算新的变量。

Analyze：“分析”菜单在以前版本中为“统计 (Statistics)”，可进行各种统计分析，包括各种统计过程(Procedure)，如回归分析、相关分析、因子分析等等。

Graphs：“图表”菜单产生条形图、饼图、直方图、散点图和其它全颜色、高分辨率的图形，以及动态的交互式图形。有些统计过程也产生图形，所有的图形都可以编辑。

Utilities：“工具”菜单可以显示数据文件和变量的信息，定义子集，运行脚本程序，自定义 SPSS 菜单等。

Window：“窗口”菜单用于选择不同窗口和最小化所有窗口。

Help：“帮助”菜单包含 SPSS 帮助主题、SPSS 教程、SPSS 公司主页、统计教练等菜单项。

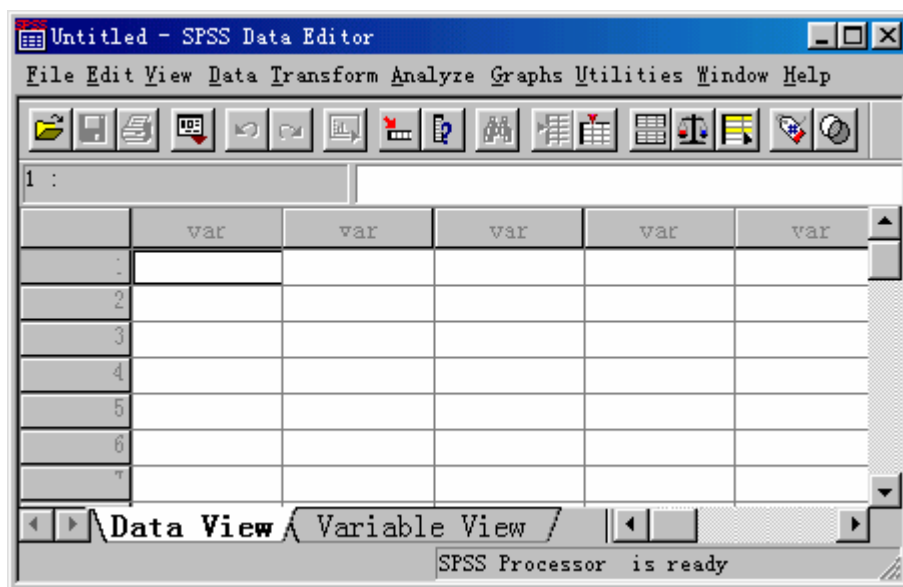


图 2-11 SPSS 主界面（数据编辑器）

（三）退出 SPSS

选择数据编辑器的[File]菜单中的[Exit]或单击标题栏上的关闭按钮退出 SPSS。

第三节 SPSS 基本操作

使用 SPSS 进行统计分析时，首先要录入数据或者打开一个已经存在的数据文件，根据需要进行数据转换；然后选择合适的统计分析过程，选择统计分析所采用的方法和参数；最后分析 SPSS 输出的结果，并保存结果。

一、数据管理(Data Management)

启动 SPSS 后，出现的界面是数据编辑器窗口（如

图 2-11所示），它的底部有两个标签：[Data View（数据视图）]和[Variable View（变量视图）]，它们提供了一种类似于电子表格的方法，用以产生和编辑 SPSS 数据文件。[Data View]

用于查看、录入和修改数据，[Variable View]定义和修改变量的定义。如果使用过电子表格如 Microsoft Excel 等，那么数据编辑窗口的许多功能应该已经熟悉。但是，还有一些明显区别：(1) 列是变量，即每一列代表一个变量(Variable)或一个被观测量的特征。例如问卷上的每一项就是一个变量。(2) 行是观测，即每一行代表一个个体、一个观测、一个样品，在 SPSS 中称为事件(Case)。例如，问卷上的每一个人就是一个观测。(3) 单元包含值，即每个单元包括一个观测中的单个变量值。单元(Cell)是观测和变量的交叉。与电子表格不同，单元只包括数据值而不能含公式。(4) 数据文件是一张长方形的二维表。数据文件的范围是由观测和变量的数目决定的。可以在任一单元中输入数据。如果在定义好的数据文件边界以外键入数据，SPSS 将数据长方形延长到包括那个单元和文件边界之间的任何行和列。

如果要分析的数据还没有录入，可用数据编辑器来键入数据并保存为一个 SPSS 数据文件（其默认扩展名为.sav）。

（一）定义变量。

输入数据前首先要定义变量。定义变量即要定义变量名、变量类型、变量长度（小数位数）、变量标签（或值标签）和变量的格式，步骤如下：单击数据编辑窗口中的[Variable View]标签或双击列的题头(Var)，显示如所示的变量定义视图，在出现的变量视图中定义变量。每一行存放一个变量的定义信息，包括[Name]、[Type]、[Width]、[Decimal]、[Label]、[Value]、[Missing]、[Columns]、[Align]、[Measure]等。

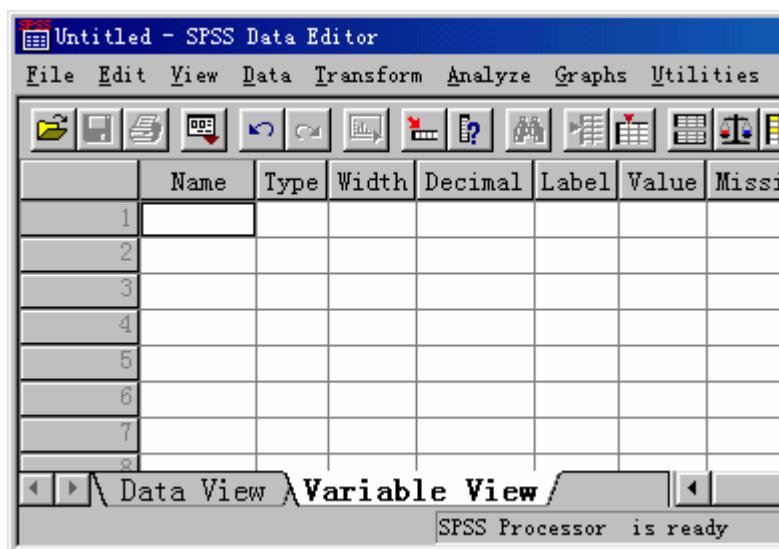


图 2-12 定义变量

1、[Name]：定义变量名

变量名必须以字母或字符@开头，其他字符可以是任何字母、数字或_、@、#、\$等符号。变量名总长度不能超过 8 个字符（即 4 个汉字）。

2、[Type]：定义变量类型

SPSS 的主要变量类型有：Numeric（标准数值型）、Comma（带逗点的数值型）、Dot（逗点作小数点的数值型）、Scientific Notation（科学记数法）、Date（日期型）、Dollar（带美元

SPSS10.0 的变量定义方式与 SPSS 9.0 以下版有所不同。

为方便起见，在本书中用方括号“[]”表示菜单项名称或者对话框中的标签等，并把菜单选择简记为 []=>[]，如[File]=>[Exit]表示：先单击主菜单的[File]项，然后在其下拉菜单中单击[Exit]菜单项。

如果在中文版 Windows 下运行英文版 SPSS，那么对话框中的部分文字可能无法完整显示出来。可在文字标签或按钮上单击右键，通过其帮助信息了解其含义。下同。

符号的数值型)、Custom Currency(自定义型)、String(字符型)。单击[Type]相应单元中的按钮,显示如图 2-13 所示的对话框,选择合适的变量类型并单击[OK]。

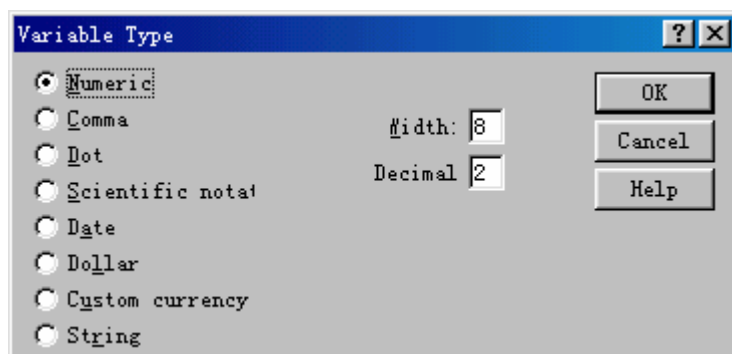


图 2-13 定义变量类型对话框

3、[Width]: 变量长度

设置数值型变量的长度,当变量为日期型时无效。

4、[Decimal]: 变量小数点位数

设置数值型变量的小数点位数,当变量为日期型时无效。

5、[Label]: 变量标签

变量标签是对变量名的进一步描述,变量只能由不超过 8 个字符组成,8 个字符经常不足以表示变量的含义。而变量标签可长达 120 个字符,变量标签对大小写敏感,显示时与输入值完全一样,需要时可用变量标签对变量名的含义加以解释。

6、[Value]: 变量值标签

值标签是对变量的每一个可能取值的进一步描述,当变量是定类或定序变量时,这是非常有用的。单击[Value]相应单元,在如图 2-14 所示的对话框中进行设置。

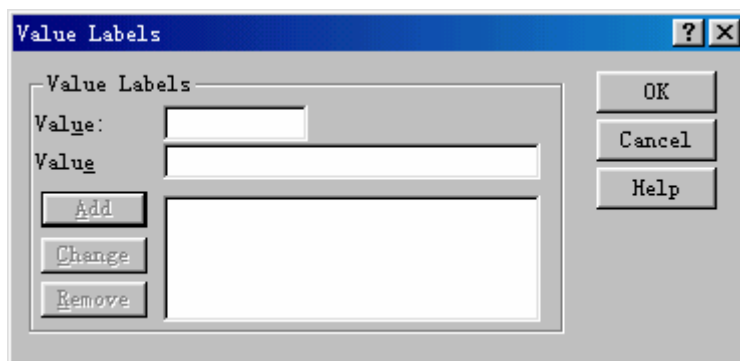


图 2-14 修改变量标签和值标签

7、[Missing]: 缺失值的定义方式

SPSS 有两类缺失值:系统缺失值和用户缺失值。在数据长方形中任何空的数字单元都被认为系统缺失值,用点号(.)表示。SPSS 可以指定那些由于特殊原因造成的信息缺失值,然后将它们标为用户缺失值,统计过程识别这种标识,带有缺失值的观测被特别处理。默认值为[None]。单击[Value]相应单元中的按钮,可改变缺失值定义方式,如图 2-15 所示。

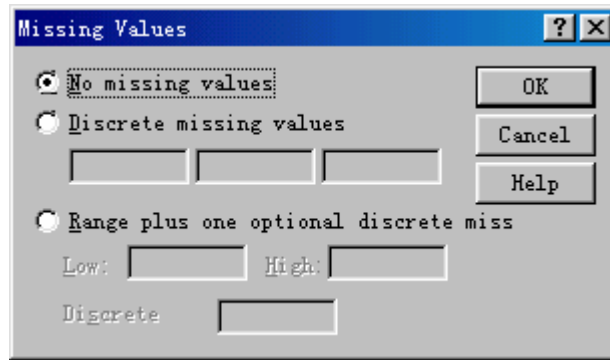


图 2-15 改变缺失值的定义方式

8、[Column]：变量的显示宽度

输入变量的显示宽度，默认为 8。

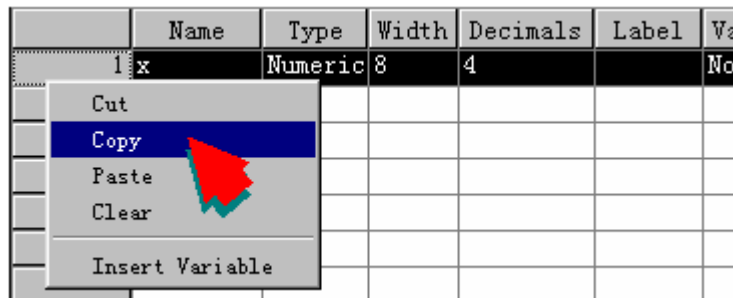
9、[Align]：变量显示的对齐方式

选择变量值显示时的对齐方式：[Left（左对齐）]、[Right（右对齐）]、[Center（居中对齐）]。

10、[Scale]：变量的测量尺度

正如前面所说的，变量按测量精度可以分为定类变量、定序变量、定距变量和定比变量，定距变量和定比变量经常不加以区别。如果变量为定距变量或定比变量，则在[Scale]相应单元的下拉列表中选择[Scale]；如果变量为定序变量，则选择[Ordinal]；如果变量为定类变量，则选择[Nominal]。

如果有许多个变量的类型相同，可以先定义一个变量，然后把该变量的定义信息复制给新变量。具体操作为：先定义一个变量，在该变量的行号上单击右钮，弹出如图 2-16（A）所示的快捷菜单，选择[Copy]；然后用鼠标右钮选择多行，弹出如图 2-16（B）所示的快捷菜单，选择[Paste]；再把自动产生的新变量名称（如 Var0001、Var0002、Var0003、……）改为所要的变量名称。



(A) 复制

	Name	Type	Width	Decimals	Label	Value Labels
1	x	Numeric	8	4		No.
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

(B) 粘贴

图 2-16 复制变量定义信息

定义了所有变量后，单击[Data View]即可在数据视图输入数据。

(二) 数据的输入与编辑

定义了变量后就可以输入数据了，数据窗口如图 2-17所示。

当前观测号 变量名

单元编辑器

变量

3 : bdate

07/26/29

	id	gender	bdate	educ	jobcat
1	1	m	02/03/52	15	3
2	2	m	05/23/58	16	1
3	3	f	07/26/29	12	1
4	4	f	04/15/47	8	1
5	5	m	02/09/55	15	1
6	6	m	08/22/58	15	1
7	7	m	04/26/56	15	1

观测

当前单元格

图 2-17 数据文件格式

由于各种原因，已经输入的数据往往会有错误，这就需要进行编辑。用 Windows 的基本操作方式可实现对数据的编辑，例如，可用方向键或鼠标移动到要修改的单元，键入新值。如果数据文件较大且知道要修改的数据单元的行号，可通过选择[Data]⇒[Go to Case]打开如图 2-18 所示的对话框，在对话框中[Case Number]的右框输入行号来查找特定观测（行）。如果要查找某变量中的特定值或值标签，选择该变量，再选择[Edit]⇒[Find]或者按 Ctrl+F 打开如图 2-19 所示的对话框，在[Search for]右框中输入要查找的数值或标签。

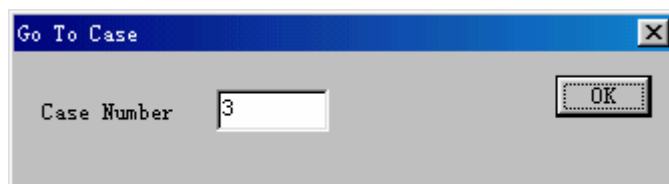


图 2-18 指向观测对话框

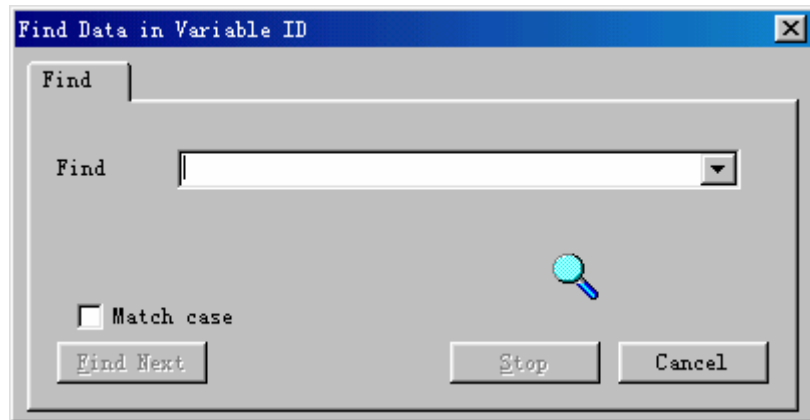


图 2-19 查找数据对话框

（三）数据转换

在理想情况下，输入的原始数据完全适合要执行的统计分析类型，遗憾的是，这种情况很罕见，经常需要通过数据转换来提示变量之间的真实关系。利用 SPSS 可进行从简单到复杂的数据转换。例如：

1、根据已存在的变量建立新变量

选择[Transform]>[Compute]，打开如图 2-20 所示的[Compute Variable (计算变量)]对话框。在对话框中的[Target Variable (目标变量)]下框中输入符合变量命名规则的变量名，目标变量可以是现存变量或新变量。对话框中[Numeric Expression (数值表达式)]下的文本框用于输入计算目标变量值的表达式。表达式能够使用左下框列出的现存变量名、计算器板列出的算术运算符和常数和[Functions (函数)]列表框显示的各种函数等。可以在文本框中直接输入和编辑表达式，也可以使用变量列表、计算器板和函数列表将元素粘贴到文本框中。

计算器板包括数字、算术运算符、关系运算符和逻辑运算符，可以象使用计算器一样使用它们。计算器板上的算术运算符有+（加）、-（减）、*（乘）、/（除）、**（指数）、（）（运算符顺序）；关系运算符有<（小于）、>（大于）、<=（小于等于）、>=（大于等于）、=（等于）、<>（不等于）等；逻辑运算符有&（and，与运算，A、B 两关系均为真时 A&B 才为真）、|（or，或运算，A、B 任一关系为真时 A|B 即为真）、~（not，非与算，颠倒表达式的真假结果，A 为真则~A 为假，A 为假则~A 为真）。

函数表 70 多个函数，包括算术函数、统计函数、分布函数、逻辑函数、日期和时间汇总与提取函数、缺失值函数、字符串函数、随机变量函数等等，例如自然对数 LN()、绝对值对数 ABS()、求和函数 SUM()等。

计算器板下面有一个[IF]按钮，单击该按钮打开条件表达式对话框。在条件表达式对话框中指定一个逻辑表达式，一个逻辑表达式对每一个观测(case)返回真、假或缺失值。如果一个逻辑表达式的结果是真，就把转换应用于那个观测；如果结果是假或缺失值，就不对那个观测应用转换。

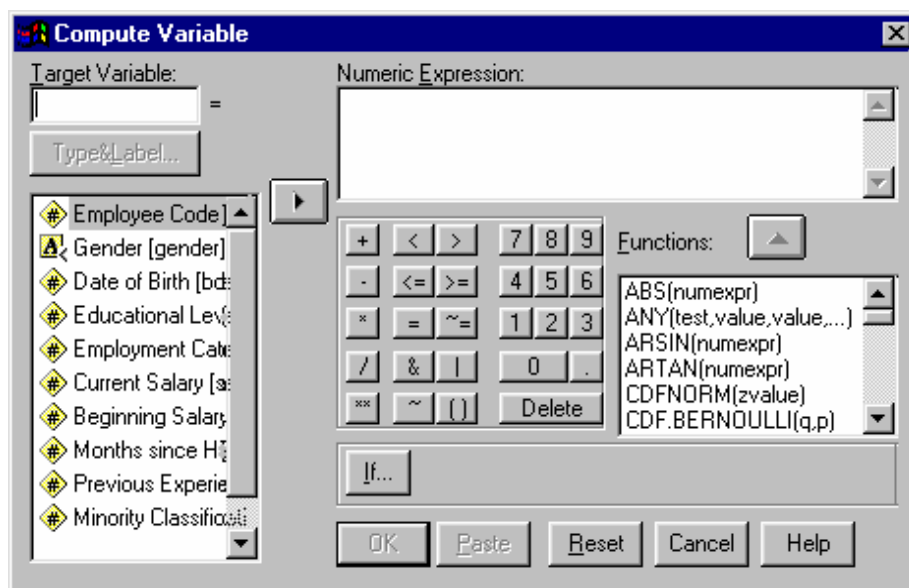


图 2-20 计算变量对话框

2、对观测 (case) 记录进行排序

在数据文件中, 可根据一个或多个排序变量的值重排观测的顺序。选择[Data]=[Sort Cases], 打开[Sort Cases]对话框, 如图 2-21所示。

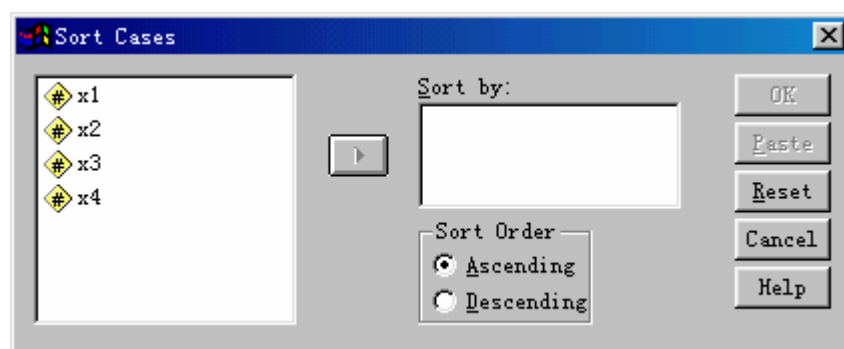


图 2-21 观测排序对话框

3、观测或变量转置

SPSS 中将行作为观测, 列作为变量。对那些观测和变量的行列关系与此相反的数据文件, 可以选择[Data]=[Transpose]将行列互换, 对话框如图 2-22所示。

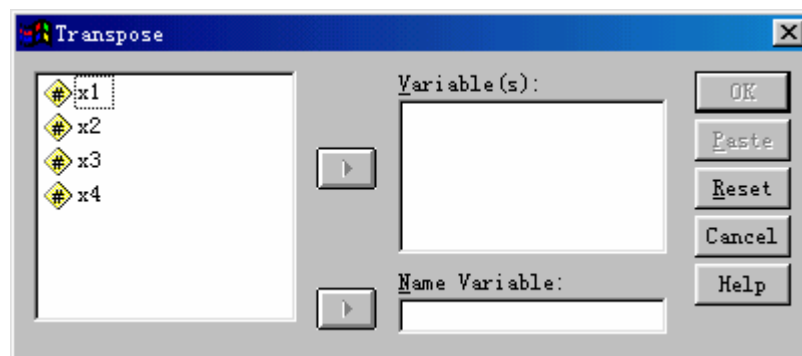


图 2-22 转置对话框

4、文件合并

可以将两个或更多个数据文件合并在一起, 即可将具有相同变量但观测不同的文件合

并,也可将观测相同变量不同的文件相合并。选择[Data]=>[Merge Files]=>[Add cases]从第二个文件即外部 SPSS 数据文件向当前工作数据文件追加观测。选择[Data]=>[Merge Files]=>[Add Variables]合并含有相同观测但不同变量的两个 SPSS 外部文件。

5、选取观测子集

可以选择[Data]=>[Select Cases]根据包含变量和复杂的表达式的准则把统计分析限于某一特定观测子集,也可选取一个随机观测样本。这样就可以同时对不同的观测子集作不同的统计分析。

6、其它转换

数据汇总,[Data]=>[Aggregate];

数据加权,[Data]=>[Weight Cases];

数值编码,[Transform]=>[Recode];

数据求秩,[Transform]=>[Rank Cases];

产生时间序列,[Transform]=>[Create Time Series];等等。

(四) 保存数据文件

在数据文件中所做的任何变化都仅在这个 SPSS 过程期间保留,除非明确地保存它们。要保存对前面建立的数据文件进行的任何改变,选择[File]=>[Save]或按 Ctrl+S 快捷键即可。如果要把数据文件保存为一个新文件或将数据以不同格式保存,可选择[File]=>[Save As],打开如图 2-23所示的对话框。主要的保存类型有:

SPSS(*.sav), SPSS 10.0 默认格式;

SPSS7.0(*.sav), SPSS 7.0 格式;

SPSS/PC+(*.sys), SPSS/PC+格式;

Excel(*.xls), Microsoft Excel 格式;

1-2-3 Rel 3.0(*.wk3), Lotus 1-2-3 V3.0 电子表格文件;等等。

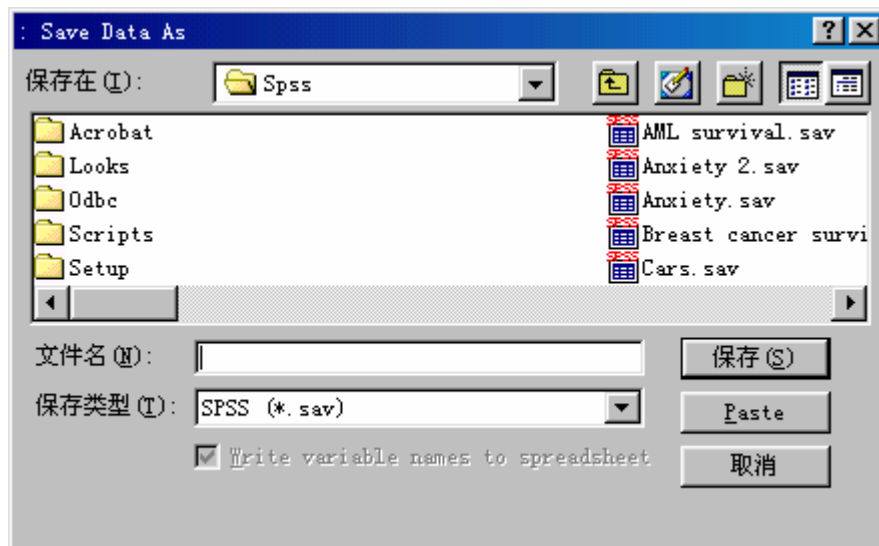


图 2-23 “另存为”对话框

(五) 打开已经存在的数据文件

选择[File]=>[Open]或按快捷键 Ctrl+O,显示[Open file(打开文件)]对话框。选择要打开的文件的文件类型和文件名,单击[打开]。

二、统计分析(Statistical Analysis)

在 SPSS 中建立了数据文件或打开一个数据文件之后,选择正确的统计分析方法,是得

到正确分析结果的关键步骤。统计分析过程在主菜单[Analyze(分析)] 中的下拉菜单中，如表 2-3所示。本书介绍的统计分析方法的 SPSS 使用参见相关章节。

表 2-3 统计分析过程

菜单项	菜单项包含的统计分析功能（子菜单项）
Reports（报告）	OLAP Cubes（OnLine Analytical Processing Cubes） Case Summaries（观测概要） Report Summaries in Rows（行形式输出报告） Report Summaries in Columns（列形式输出报告）
Descriptive Statistics（描述统计）	Frequencies（一维频数分布表） Descriptive（描述统计量计算） Explore（数据探索） Crosstabs（多维频数列表，列联分析）
Compare Means(均值比较)	Means（分组求均值） One-Sample T Test（单样本 T 检验） Independent-Samples T Test（独立样本 T 检验） Paired-Samples T Test（配对/相关样本 T 检验） One-way ANOVA（一维方差分析）
General Linear Model（GLM，一般线性模型）	Univariate（单变量 GLM） Multivariate（多变量 GLM） Repeated Measures（重复测量设计的 GLM） Variance Components（方差成分）
Correlate（相关分析）	Bivariate（两个变量的相关分析） Partial（偏相关分析） Distances（距离分析）
Regression(回归分析)	Linear（线性回归分析） Curve Estimation（曲线估计） Binary Logistic（二值 Logistic 回归分析） Multinomial Logistic（多项 Logistic 回归分析） Ordinal（有序回归） Probit（Probit 回归分析） Nonlinear（非线性回归分析） Weight Estimation（加权估计） 2-Stage Least Squares（两阶段最小二乘法回归分析）
Loglinear（对数线性模型）	General（一般对数线性模型分析） Logit（Logit 分析） Model Selection（模型选择对数线性分析）
Classify（分类）	K-Means Cluster（K 均值大样本聚类分析） Hierarchical Cluster（系统/层次聚类分析） Discriminant（判别分析）
Data Reduction（数据降维）	Factor（因子分析，主成分分析） Correspondence（对应分析）
Scale（等级分析）	Reliability Analysis（可靠性分析）

在 SPSS V8.0 或更低版本中，该菜单项的名称为[Statistics(统计)]。

菜单项	菜单项包含的统计分析功能(子菜单项)
	Multidimensional Scaling (多维等级分析)
Nonparametric Tests(非参数检验)	Chi-Square (卡方检验) Binomial (二项检验) Runs (游程检验) 1-Sample K-S (单样本 K-S 检验) 2 Independent Samples (两个独立样本非参数检验) K Independent Samples (多个独立样本非参数检验) 2 Related Samples (两个相关样本非参数检验) K Related Samples (多个相关样本非参数检验)
Time Series (时间序列)	Exponential Smoothing (指数平滑) Autoregression (自回归) ARIMA (差分自回归滑动平均模型) X11 ARIMA (X11 ARIMA) Seasonal Decomposition (季节分解)
Survival (生存分析)	Life Tables (生命表分析) Kaplan-Meier (卡普兰-梅尔分析) Cox Regression (Cox 回归分析)
Multiple Response (多重应答)	Define Sets (定义多重应答数据集合) Frequencies (多重应答频数) Crosstabs (多重应答交叉列表)
Missing Value Analysis	Missing Value Analysis (缺失值分析)

三、图形分析(Graphical Analysis)

统计图是用点的位置、线段的升降、直条的长短或面积的大小等方法来表达统计数据的一种形式,它可以把资料所反映的变化趋势、数量多少、分布状态和相互关系等形象直观地表现出来,以便于读者的阅读、比较和分析。统计图具有简明生动、形象具体和通俗易懂的特点。SPSS 的图形分析功能很强,许多高精度的统计图形可从[Analyze]菜单的各种统计分析过程产生,也可以直接从[Graphs]菜单中所包含的各个选项完成。图形分析的一般过程为:建立或打开数据文件,若数据文件结构不符合分析需要,则必须转换数据文件结构;生成图形;修饰生成的图形,保存结果。常用的统计图形有条形图、线图、面积图、圆饼图、散点图、直方图、箱线图等等,见表 2-4。其中统计图形有两种形式,一种为一般图形,另一种为交互式图形,交互式图形提供了更多的选项,可绘制出更强大的图形。

表 2-4 一般统计图形

图形名称	示意图	菜单项选择
条形图(Bar)		[Graphs]=>[Bar]
线图(Line)		[Graphs]=>[Line]
面积图(Area)		[Graphs]=>[Area]
饼图(Pie)		[Graphs]=>[Pie]

高低图(High-Low)		[Graphs]>[High-Low]
帕累托图(Pareto)		[Graphs]>[Pareto]
工序控制图(Control)		[Graphs]>[Control]
箱线图(Boxplot)		[Graphs]>[Boxplot]
误差条图(Error Bar)		[Graphs]>[Error Bar]
散点图(Scatter)		[Graphs]>[Scatter]
直方图(Histogram)		[Graphs]>[Histogram]
P-P 正态概率图 (Normal P-P)		[Graphs]>[P-P]
Q-Q 正态概率图 (Normal Q-Q)		[Graphs]>[Q-Q]
时序图(Sequence)		[Graphs]>[Sequence]
自相关图 (Autocorrelations)		[Graphs]>[Time Series] =>[Autocorrelations]
互相关图 (Cross-Correlations)		[Graphs]>[Time Series] =>[Cross-Correlations]

表 2-5 交互式统计图形

图形名称	菜单项选择
条形图(Bar)	[Graphs]>[Interactive]>[Bar]
点图(Dot)	[Graphs]>[Interactive]>[Dot]
线图(Line)	[Graphs]>[Interactive]>[Line]
带状图(Ribbon)	[Graphs]>[Interactive]>[Ribbon]
点线图(Drop-Line)	[Graphs]>[Interactive]>[Drop-Line]
面积图(Area)	[Graphs]>[Interactive]>[Area]
饼图(Pie)	[Graphs]>[Interactive]>[Pie...]
箱线图(Boxplot)	[Graphs]>[Interactive]>[Boxplot]
误差条图(Error Bar)	[Graphs]>[Interactive]>[Error Bar]
直方图(Histogram)	[Graphs]>[Interactive]>[Histogram]
散点图(Scatterplot)	[Graphs]>[Interactive]>[Scatterplot]

四、输出管理(Output Management)

不管是统计分析还是图形分析，其结果都输出到新的窗口——Viewer 窗口或 Draft Viewer 窗口，SPSS 默认输出窗口为 Viewer 窗口（如图 2-24 所示）。Viewer 窗口的左边是输出大纲视图（如图 2-25 所示），可以单击统计过程名称左边的“+”和“-”展开或收缩输出

大纲，也可以拖动输出内容项目改变项目的位置。Viewer 窗口的右边显示具体的输出内容（如图 2-26 所示），一般通过文字、表格、图形显示统计计算结果。许多输出结果以数据透视表(Pivot Table)的表格形式显示，数据透视表功能强大，便于用户自行定义所需格式。如果要查看数据透视表中某个统计术语的含义，双击该数据透视表，右击术语，在弹出的快捷菜单中选择[What's This]，就可获得该术语的简单定义。用户可通过与操作 Windows 应用程序一致的方法使用 Viewer 窗口，这里不详细介绍。

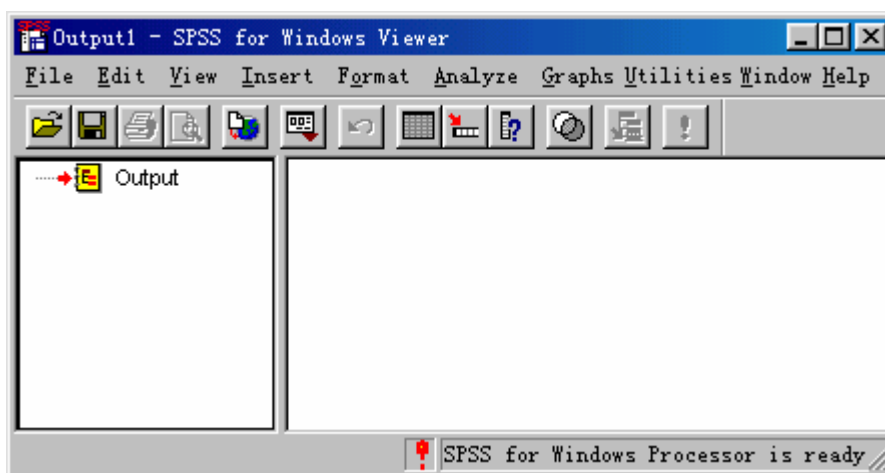


图 2-24 输出窗口

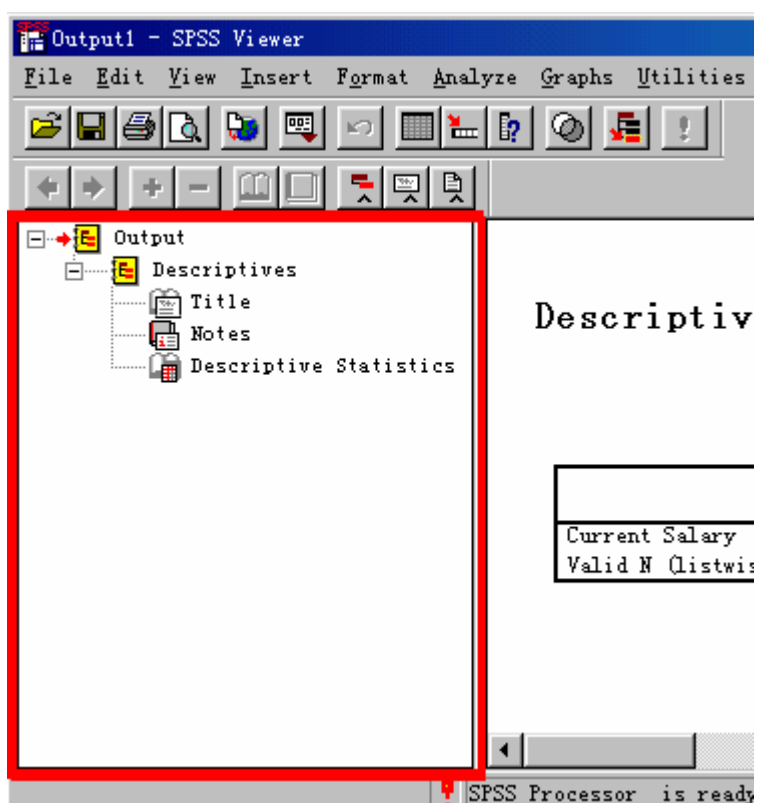
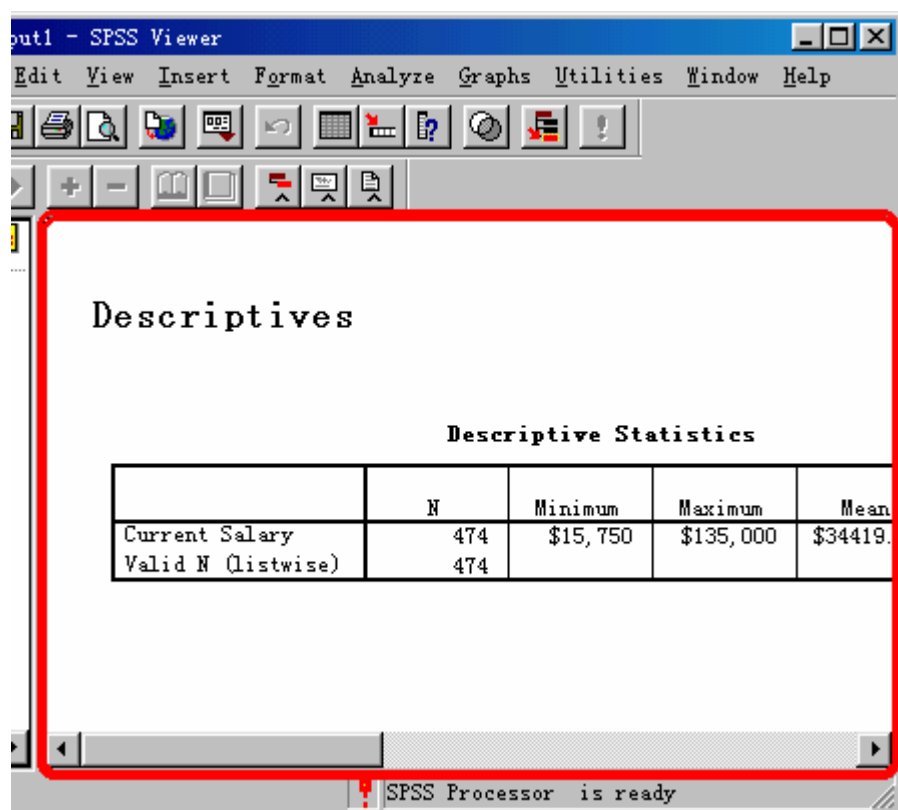


图 2-25 输出大纲视图



SPSS Viewer

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean
Current Salary	474	\$15,750	\$135,000	\$34419.
Valid N (listwise)	474			

SPSS Processor is ready

图 2-26 输出内容

第三章 统计数据的收集、整理与描述

第一节 统计数据的来源

统计数据的来源渠道很多,不同的统计数据可通过不同渠道获得。在进行一项研究时,可以查阅报纸书刊、查阅统计年鉴、也可以通过 Internet 查阅联机数据库。如果这些数据仍不能满足研究的需要,还可以委托调查公司或者自己组织调查,以获得必要的统计数据。我们把来源于直接的调查和科学实验的统计数据称为第一手统计数据;把来源于别人调查和科学实验的数据称为第二手统计数据。第二手统计数据主要是公开出版的统计数据。当然,我们有时也通过一些渠道设法使用一些尚未公开的数据。对于第二手统计数据,作为使用者来说,我们要清楚从哪里可以获得有关数据,并要了解这些数据的来源、指标口径和数据的质量。

一、统计数据的直接来源

统计数据的直接来源主要是通过专门组织的直接调查和科学试验这两个渠道获得的。

(一) 来源于管理和研究需要而专门组织的调查

在进行管理决策和科学研究时,如果能利用现成的数据当然是省时、省钱、省力的好办法。但为了国民经济宏观管理的需要,就必须掌握最新的人口、农作物产量、国内生产总值、主要工业产品产量和产值,以及人民生活的变化情况,这就需要经常组织专门的调查以获得国民经济管理的基本数据。国家统计局系统和国务院各部、委、局的统计系统就承担首这些调查任务。

另一方面,在社会主义市场经济条件下,大量的市场调查和民意测验也都需要组织专门的统计调查,以搜集特殊的数据,满足管理和研究的要求。例如家用电器质量调查,化妆品品牌的调查,广播电视收视率调查,居民闲暇时间使用情况的调查,白领阶层调查等等。为适应社会和市场的需求,现在国内大中城市的市场调查业正在兴起,越来越多的调查公司或调查研究所承担起专门组织调查的任务,为特殊的管理和研究服务。

(二) 来源于科学试验的数据

在社会科学研究和经济管理中,我们用调查的方法搜集必要的统计数据。在自然科学和工程的各个研究领域,如物理、化学、生物、医学、农业和工业等领域是通过科学试验的方法获得统计研究的数据。例如某化工厂生产某种化工产品,为了在不同的影响因素(原料配方)的不同水平中选取最优的水平,就要通过试验的方法获得必要的的数据,通过对数据的统计分析来确定生产的最优方案;又例如农业科研中要通过试验的方法选取最优品种和最佳的种植方式。在医学中通过临床试验的数据分析某种药物或治疗方案的疗效。

二、统计数据的间接来源

对于社会上绝大多数的研究工作者和实际工作者来说,亲自去做直接的调查往往是不可能的。这时,可以通过各种渠道获取别人调查或科学试验的第二手数据。

(一) 来源于公开出版物的数据

第二手统计数据主要是公开出版或公开报道的数据。在我国,公开出版物或报道的社会经济统计数据主要来自国家和地方的统计部门以及各种报刊传媒。现在,随着计算机网络技术的发展,各国的报刊、杂志、图书及各种音像制品都可以从 Internet 上获得,这也为我们获得各种统计数据提供了方便。

（二）来源于内部调查的数据

对于我们的科学研究和经济管理来说,除了利用已经公开发表的数据,还要充分利用已搜集到的但未公开发表的数据。因为统计调查的大量信息,特别是调查的原始数据,或者是由于公开发表篇幅的限制,或者由于数据保密的原因等未公开发表、也未充分利用。例如,城市和农村居民家庭的收入支出调查搜集了大量的统计数据和信息,是我们进行经济管理等大量研究的宝贵数据资源,应该充分挖掘和利用。当然,在使用这些内部数据时既要考虑数据的保密问题,又要考虑与原调查单位的合作问题。因为只有解决了内部数据的合使用问题后,才能发挥这些数据的作用,才能最大限度地发挥二手资料的作用。

利用间接来源的统计数据对使用者来说既省时又省钱、省力。但使用时应注意统计数据指标的含义、计算口径和计算方法,以避免误用或滥用。同时,在引用统计数据时,一定要注明数据的来源。这样,既尊重别人的劳动成果,也便于读者查找核对。

第二节 统计数据的收集

统计数据的收集就是统计调查,它按研究的目的和要求,有组织地向调查对象收集相关的各种资料。为了保证统计数据资料的完整性、准确性和及时性,必须熟悉各种收集方法及各自的特点。

一、问卷调查

问卷是调查者向被调查者了解情况或征询意见时所运用的统一设计的调查表。绝大多数旨在收集定量数据的调查都要采用某种形式的问卷。问卷的质量高低对调查成功与否起决定作用,只有研究者设计出高水平、高质量的问卷,才会使调查得以顺利完成,并获得令人满意的数据。

问卷按传递方式不同,可分为报刊问卷、邮政问卷、送发问卷和访问问卷。问卷还可以按调查方式分类,如按问卷的填答者不同,可分为自填问卷和代填问卷。

从问卷的基本结构来看,应包括封面信、指导语、调查内容及编码四个基本内容。

第一,封面信。给调查者的一封信,一般内容不宜过长,以二三百字为宜。在封面信中写清单位地址、电话号码、邮政编码、联系人姓名等,并说明大致的调查内容和进行这项调查的目的,调查对象的选取和调查结果保密的措施。在信的结尾处,要真诚的感谢调查者的合作和支持。

第二,指导语。指导语是用来指导被调查者填写问卷的一组说明。指导语的形式有两种。一种是写在封面信之后,另一种是分别放在某些较复杂的问题后,用括号括起来,其作用主要是指导被调查者准确理解与填写该问题。比如:(可选择多个答案)(请按重要顺序排列)等,通常两种形式结合使用。

第三,调查内容。调查内容是问卷的主体,调查项目的多少由调查目的和经费决定,每个项目包括问题和答案两个部分。从形式上分,问题可归为开放式问题和封闭式问题;从内容上分,可分为有关个人背景资料问题、行为问题和态度问题及知识问题等。

开放式问题不为被调查者提供答案,而由回答者自由回答。比如:

- 您喜欢什么品牌的啤酒?

回答者可不受限制,回答自己喜欢的啤酒。开放式问题的优点是,被调查者可以充分自由地按自己的方式发表意见,因为所得资料比较生动、丰富。但开放式问题所花费的时间和敬礼较多,编码工作复杂繁琐,且开放式问题难于进行定量分析处理。

封闭式问题是在提出问题的同时,给出若干个答案,要求被调查者选择一个或多个答案作为回答。如:

- 您喜欢哪种国产牌子的啤酒?

青岛	☐
北京	☐
五星	☐
三星	☐
其它_____（请注明）	☐

封闭式问题的优点是被调查者填写十分方便，花的时间较少，易于编码与处理分析，所得资料适于定量分析，缺点是不够丰富生动，在给定选择答案以外的信息难以收集。

第四，编码。市场调查，一般多采用封闭式问题为主的问卷。为了便于计算机进行处理和定量分析，往往要对回答结果进行编码。编码是质对于每一个问题答案赋予一个数字作为它的代码。在问卷设计时就设计好的编码，称为预编码，而在调查完后再编码，称为后编码。编码一般放在问卷每一页的最右边，有时用一条线将它与问题及答案分开。如下例：

		编码
您的年龄：_____岁	☐ ☐	1-2_____
您的性别：男	☐	3_____
女	☐	
您的文化程度：		
小学及以下	☐	4_____
初中	☐	
高中或中专	☐	
大专及以上	☐	
您的月收入为_____元？ ☐ ☐ ☐ ☐		5-8_____

除编码外，访问问卷一般要求在封面印上调查员姓名、访问日期、审核员姓名、被调查者住址等有关资料。

（三）问卷设计技巧

一份问卷包括许多内容，合理安排其顺序、排列方式，并注意问题的提问方式，既能使被调查者乐于回答，又能提高问卷质量。

从问卷的顺序看，应注意如下几点：

先个人背景问题，然后行为资料、态度资料、知识资料问题。

先封闭性问题，后开放性问题。回答开放性问题一般需要考虑的时间要长一些，如先问开放性问题，被调查者会认为此调查表花费精力较多，而拒绝回答。

先一般问题，后敏感性问题。敏感性问题，如收入问题，放在前面会引起回答者的反感，容易被拒绝或产生偏差。

为检验回答的正确性和可靠性，常在一张问卷中对同一问题从正反两方面加以提问。对于这样的问题不宜放在一起，应予隔开。如：“该洗发水去头屑吗？”“该洗发水不去头屑吗？”，若将这两个问题放在一起，回答者为使前后回答一致，又可能选择同一答案，这就不能起到检验可靠性的目的。

先易后难。如将生疏或难回答的问题放在前面，容易使被调查者产生抵触情绪而拒绝回答。

按逻辑顺序排列时应注意：时间要从近及远（或从远及近）排列；具体内容分门别类，不能交叉排列，要由浅入深地按逻辑思维顺序排列。

对后续性问题的设置方式。有些问题只适用于某种类型的人，这时，可考虑设置后续性问题，回答者是否应该回答后续性问题，则取决于前一个问题的答案。例如：“你用过××牌洗发水吗？”“用了多久了？”后一个问题即为后续性问题，是否回答该问题取决于前一

个问题的回答。后续性问题可采用不同的格式，例如：

框架式：

“你用过××牌洗发水吗？”_____

用过,若用过,请回答：用了多久了？_____

没用过

说明式：

“你用过××牌洗发水吗？”_____

用过,(请回答下一问题)

没用过(请跳过下一问题)

●你用了多久了？_____

设计提问规则。提问时应做到：

用语准确，含义清楚；

文字简明扼要；

避免诱导性提问，如提问：“你不赞成吸烟吧？”，回答者容易偏向“不赞成”；

不能超越调查对象的知识水平，如果调查对象不懂词的含义，调查很难继续；

合理使用俚语。俚语的使用可增加亲密感，但俚语的使用具有一定的范围，超出该范围则易引起误解；

尽量少涉及个人隐私及不受欢迎的问题；

对敏感性问题的回答应事先说明，以求真实的回答。

二、普查法

普查，是按照一定标准时间对普查对象的全部单位无一例外地逐个进行的调查。普查按门类划分，可分为人口普查、工业普查、商业普查、农业普查、第三产业普查等。普查按区域划分，有宏观、中观和微观之分。一般而言，我们经常提起的普查为宏观普查。

普查地域广阔，调查对象多，参加调查人员多，且时间短，因此工作十分复杂，组织普查必须注意下列问题：第一，普查时点统一；第二，正确选择调查时间；第三，普查指标不宜过多；第四，对各项调查指标必须有统一的规定和解释，有明确的操作定义，统一的计算公式，一经规定，不得任意改变和增减，以免降低资料质量与破坏一致性；第五，普查应一定周期进行。

普查的实施是一项巨大的系统工程，必须遵循科学规律科学规律有序进行，才能保证普查工作顺利完成。第一，要建立高效、统一的领导机构；第二，精心设计普查方案；第三，培训普查员；第四，普查试点；第五，做好普查登记；第六，提高资料处理水平；第七，普查环节的质量控制。

三、抽样调查

普查的覆盖面宽，但其耗费的人力、物力、财力太大，在统计调查中抽样调查更为常用。抽样调查是从调查对象的总体中，按照一定的抽样原则抽取一部分单位作为样本，并以对样本进行调查的结果来推断总体的方法。

根据抽样方法是否随机，可将抽样调查分为随机抽样和非随机抽样两大类。我们将在第五章讨论随机抽样问题。

四、典型调查

典型调查是从调查对象的总体中选取一个或几个有代表性的单位进行全面、深入的调查。调查单位可依不同调查目的的选取企业、学校、个人、家庭等。

典型调查的目的就是通过对某个典型的深入分析来概括和反映全面。因此，典型调查要

求典型对总体推断有一定的代表性，这也是典型调查的关键。典型的代表性可以从动态、静态两个方面来衡量。从动态上来讲，是指事物的发展趋势；从静态上来讲，是指事物的共同属性与差异。

典型调查常用于统计学，是对市场中的典型单位及消费者进行深入调查的一种方法。通过市场典型调查，可以了解市场的一般规律，如：了解市场商品供需状况和变动趋势，在缺乏全面资料的情况下，可以根据典型调查的资料推算总体状况。

五、观察法

观察法是观察者深入现场或进入一定环境，观察调查对象，获取第一手资料的方法。调查人员直接到调查现场，耳闻目睹顾客对市场的反映和公开言行，或者利用照相机、监测器等现代化器械间接地进行观察来收集资料等，都属于观察法。

观察法的特点就是从侧面观察被观察者的言行和反映，一般不直接向被调查人提出问题，所以，被调查者往往是在不知情的状况下被调查的。

六、实验法

实验法是研究者根据一定的研究目的，控制某种市场条件，或在人工环境中使一定的现象产生，通过观察、记录收集资料，以揭示其发生原因或规律的方法，是一种复杂、高级调查方法。

七、集体访谈法

集体访谈法是访问调查法的延伸和扩展，是调查者邀请若干被调查者，通过集体访谈的方式了解有关情况或研究实用统计学有关问题的方法。

集体访谈法的特点在于，它所访问的不是一个一个的被调查者，而是同时访问若干个被调查者，它不是通过与个别被调查者的个别交谈来了解有关情况，而是通过若干个被调查者集体访谈来了解有关情况。因此，集体访谈过程，不仅是调查者与被调查者之间互相影响互相作用的过程，而且是若干个被调查者之间互相影响、互相作用的过程。在集体访谈法中，对调查者的素质要求较高，调查者不仅要具备熟练的访谈技巧，也要有驾驭调查成功的能力。

目前，在统计学方面，集体访谈法受到的重视程度越来越高，一般以了解产品特性、产品促销、产品质量、广告效果评价、新产品的开发上市、对市场进行预测等方面。

第三节 统计数据的整理

收集统计数据之后，要对获取的数据进行系统化、条理化地整理，以提取有用的信息。

一、统计分组

根据统计研究的目的和客观现象的内在特点，按某个标志（或几个标志）把被研究的总体划分为若干个不同性质的组，称为统计分组。统计分组的对象是总体。从分组的性质来看，分组兼有分和双重含义。对于总体而言，是“分”，即把总体分为性质相异的若干部分；而对于单位而言，又是“合”，即把性质相同的许多单位结合为一组。例如，要对某某班学生的性别进行调查，可将学生人数分成男、女两个组，分组结果如下表所示：

表 3-6 某班学生按性别分组

按性别分组	人数	频率(%)
男生	30	60

女生	20	40
合计	50	100

该班分组的标志是性别，属于定类变量。又如，对该班学生人数按考试成绩分绩，分组结果所下表所示：

表 3-7 某班学生按考试成绩分组

按考试成绩分组	人数	频率(%)
优	5	10
良	10	20
中	20	40
及格	10	20
不及格	5	10
合计	50	100

这一分组标志属于定序变量，因为组间是可以比较大小的，即优>良>中>及格>不及格。如果对该班学生按年龄进行分组，则有

表 3-8 某班学生按年龄分组

按年龄分组	人数	频率(%)
17	6	12
18	14	28
19	18	36
20	9	18
21	3	6
合计	50	100

这一分组标志属于定比变量。

二、频数分布与频率分布

将数据按其分组标志进行分组的过程，就是频数分布和频率分布形成的过程。表示各组的单位的次数称为频数，各组次数与总次数之比称为频率。频数分布就是观察值按其分组标志分配在各组内的次数，由分组标志序列和各组相对应的分布次数两个要素构成。由分组标志序列和各组相应的频率构成频率分布。

[例 3-1]某车间 30 名工人按每天加工某种零件件数如下表所示：

表 3-9 某车间工人每天加工某种零件件数

工人编号	加工零件数	工人编号	加工零件数
1	106	16	97
2	84	17	103
3	110	18	106
4	91	19	95
5	109	20	106
6	91	21	85
7	111	22	106
8	107	23	101
9	121	24	105

10	105	25	96
11	99	26	105
12	94	27	107
13	119	28	128
14	88	29	111
15	118	30	101

要对以上 30 名工人日加工零件数进行分组，先要决定分成多少组，每一组的范围（即上下组限）是多少，即确定组数和组距。组数是分组的个数，组距是每一组最大值与最小值之差。要确定这两个数值，一般是先找出全部数据的最大值和最小值。在本例中，日加工最多的是 128 件，最少是 84 件。如果采用简单的组距，即每 10 件为一组，则该例可分为 5 组，即 80~89 件、90~99 件、100~109 件、110~119 件、120~129 件。在一般情况下，组数不应少于 5 组，但也不应多于 15 组。因为分组的目的是找出数据分布的数量规律性。如果组数太少，数据都分在一、二组或三、四组中，其规律反映不出来；如果组数太多，特别是数据又太少的话，反映出来的都是偶然性差异，也不便于探索出分布的规律。在确定了组数之后，接下来的问题就是组距和组限了，即要确定每组是否相等的组距及每组的上下组限。在本例中，我们以 10 件相等的组距进行分组，则各组的组限就随之确定了。接下来，就将每名工人的加工零件数分配到应落入的组内。按我们的习惯，一般是用划“正”字进行计数，结果见下表所示：

表 3-10 频数（频率）分布表

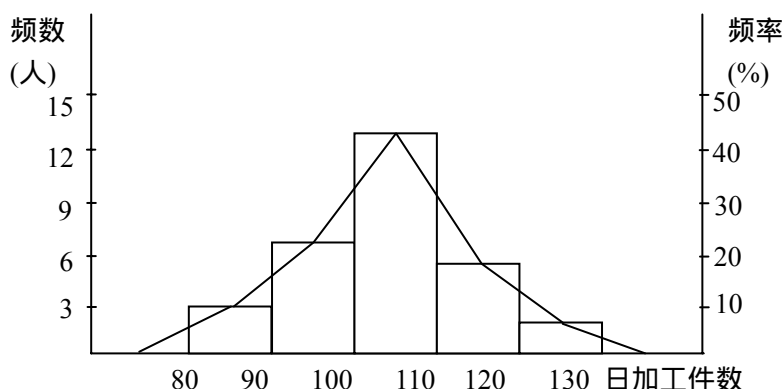
日加工零件数	频数	频率
80~89 件	3 人	10%
90~99 件	7 人	23%
100~109 件	13 人	43%
110~119 件	5 人	17%
120~129 件	2 人	7%

在分组时，要遵循“不重不漏”的原则。“不重”就是任一个单位数值只能分在其中某一组中，不能同时分在两组中；“不漏”就是任一数值必须能够分布在某一组内，不能遗漏。上面的分组是以 10 件为组距的相同组距的分组，也称为等距分组，必要时也可采取不等距分组。

将统计数据整理成频数（频率）分布形式后，已经可以初步看出数据的一些规律。如例 3-1 整理成表 3-10 的频数（频率）分布表后，就可以大致看出该车间工人日加工零件数多数 100~109 件之间，这个加工能力属于中等水平。低于中等水平的有 10 人，高于中等水平的有 7 人，因而是一种非对称的分布。对于这个频数（频率）分布结果，可以用直方图更直观、更形象地表示出来。

在平面直角坐标系上，将分组标志作为横轴并将各组频数（频率）作为纵轴，给出各组的长方形图即直方图。与直方图相似作用的图示是折线图，它以各组标志值中点位置作为该组标志的代表值，然后用折线将各组频数连接起来，开成了折线图。由表 3-10 的频数（频率）分布直方图和折线图如图 3-27 所示。在图 3-27 中，直方图与折线图的面积是相等的。折线图的折线将直方图的直角切下，正好补在旁边较低的直方图上边。这样，直方图与折线图所表示的分布规律是相同的，是两种面积相同，表现形式不同的频数（频率）分布图。

图 3-27 某车间工人日加工零件数分布图



当所观察的次数很多，组距很小并且组数很多时，所绘出的折线图就会越来越光滑，逐渐形成一条光滑的曲线，这种曲线即频数分布曲线，反映了数据的分布规律。统计曲线在统计学中很重要，是描绘各种分布规律的有效方法。常见的频数分布曲线有正态分布曲线、偏态分布曲线、J型分布曲线和U型分布曲线等。

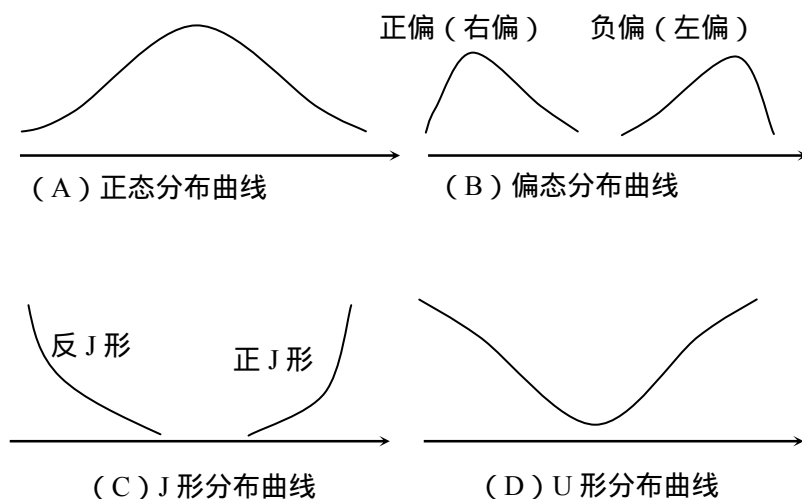


图 3-28 常见的频数分布曲线

正态分布曲线（如图 3-28 (A)所示）形为左右对称的倒挂的大钟，这是客观事物数量特征表现最多的一种频数分布曲线，如人的身高、体重、智商等等，其所有的测量和观测误差等都是服从正态分布。

偏态分布曲线（如图 3-28(B)所示）根据长尾拖向哪一方又可分为正偏（或右偏）分布曲线和负偏（或左偏）分布曲线。例如，人均收入分配的曲线就是正偏曲线，即低收入的人数较多，而高收入的人数较少，二者的收入水平差距较大。

J型分布曲线（如图 3-28(C)所示）又分为正J形分布曲线和反J型分布曲线。例如，经济学中的供给曲线是正J形曲线，需求曲线是反J形曲线。

U形分布曲线（如图 3-28(D)所示）又称为生命曲线。人和动物的死亡率近似服从U形曲线分布。婴儿和动物的幼仔由于抵抗力弱，死亡率很高，随着年龄的增长死亡率逐渐降低。到了中年时期死亡率最低同时也相对稳定，进入老年期后又逐渐增高，形成了一个U形曲

线。产品的故障和报损情况也有类似的分布规律。

三、累积频数分布与频率分布

为了统计分析的需要,有时要观察某一数值以上或某一数值以下频数或频率之和,这就需要在表 3-10 基本分组的基础上绘出累积频数或累计频率。由表的上方向表的下方的频数或频率相加就称为“向下累积”,反之称为“向上累积”。

表 3-11 累积频数(频率)分布表

日加工零件数	频数	频率	向下累积		向上累积	
			频数	频率	频数	频率
80~89 件	3 人	10%	3 人	10%	30 人	100%
90~99 件	7 人	23%	10 人	33%	27 人	90%
100~109 件	13 人	43%	23 人	76%	20 人	67%
110~119 件	5 人	17%	28 人	93%	7 人	24%
120~129 件	2 人	7%	30 人	100%	2 人	7%

例如,我们要了解日加工在 100 件及以上有多少人时,就可以从向上累积的第三组数字中直接读出 20 人;如要了解日加工在 110 件以下的人数,就可以从向下累积的第三组中直接读出 23 人。

累计频数和累计频率不仅可以用上述的表格形式来表示,而且也可以用图形来表示。

累计频数(频率)分布图分为向上累计频数(频率)分布图和向下累计频数(频率)分布图。不论是向上累计或向下累计,均以分组变量为横轴,以累计频数(频率)为纵轴。在直角坐标点系上将各组组距的上限与其相应的累计频数(频率)构成坐标点,依次用直线(或光滑曲线)相连,即是向上累计。对于向下累计频数分布图,在直角坐标系上将各组组距下限与其相应累计频数(频率)构成坐标点,依次用直线(或光滑曲线)相连,即是向下累计分布图。

累计频数(或频率)分布曲线,可用以研究财富、土地和工资收入的分配是否公平。这种累计分布曲线图最早由美国洛伦茨博七(Dr. M.O.Lorenz)提出的,故又称洛伦茨曲线图。其绘制方法如下:

- (1) 将分配的对象和接受分配者的数量均化成结构相对数并进行向上累计;
- (2) 纵轴和横轴均为百分比尺度,纵轴自下而上,用以测定分配的对象(如一国的财富、土地或收入等),横轴由左向右用以测定接受分配者(如一国的人口)。
- (3) 根据计算所得的分配对象和接受分配者的累计百分数,在图中标出相应的绘示点,连接各点并使这平滑化,所得曲线即所要求的洛伦茨曲线。

现以某国某年家庭收入资料为例(见表 3-12)说明洛伦茨曲线的绘制。

表 3-12 某国家收入所得的分配情况

按收入所得水平分组	人口			收入		累计收入的(%)		
	人口数 (万人)	结构 (%)	累计的 (%)	月收入 额(亿美元)	结构 (%)	实际情 况	绝对平 等	绝对不 平等
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

最 低	128.5	12.85	12.85	1.57	5	5	12.85	0
中 下 等	348.0	34.80	47.65	4.08	13	18	47.65	0
中 等	466.9	46.69	94.34	16.33	52	70	94.34	0
较 高	45.6	4.56	98.90	7.54	24	94	98.90	0
最 高	11.0	1.10	100.0	1.88	6	100	100.0	100
合 计	1000.0	100.0	—	31.40	100	—	—	—

在绘制分配曲线图时,先将人口收入的数量(第(1)~(4)栏)计算成为结构相对数(第(2)~(5)栏),再求出累计百分比(第(3)~(6)栏),然后在制好的比率曲线图格上依累计百分比标出绘示点,连接各绘示点即为分配曲线见图 3-29。

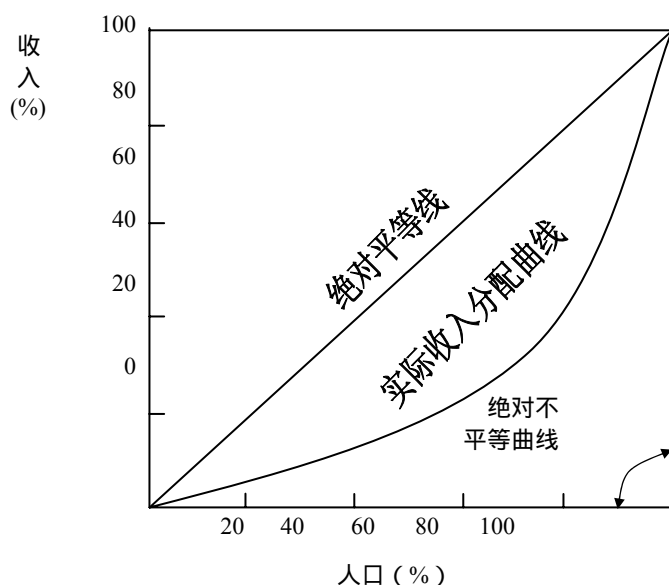


图 3-29 洛伦茨曲线示意图

图中的曲线为实际收入分配曲线,对角线为绝对平等线。根据实际收入分配线与绝对平等线或绝对不平等线进行对比,可衡量其不平等程度。离绝对平等线越远分配越不平等;反之,越靠近绝对平等线分配越平等。

四、SPSS 操作

在 SPSS 中对例 3-1 数据进行频数(率)分析的步骤为:

(一) 定义工人编号和加工零件数的变量名分别为 NO 和 X,然后输入变量 NO 和 X 的原始数据。

(二) 选择[Analyze]>[Descriptive Statistics]>[Frequencies...],弹出[Frequencies]主对话框(如图 3-30所示)。现欲 X 进行频数分析,在对话框左侧的变量列表中选 X,单击按钮使之进入[Variable(s)]列表框,并选择[Display Frequency Tables]显示频数分布表。

(三) 可单击[Format...]按钮弹出[Frequencies: Format]子对话框,在[Order by]栏中有四个选项:

[Ascending values]为根据数值大小按升序从小到大作频数分布;

[Descending values]为根据数值大小按降序从大到小作频数分布;

[Ascending counts]为根据频数多少按升序从少到多作频数分布;

[Descending counts]为根据频数多少按降序从多到少作频数分布。

这里选[Ascending values]项后点击 Continue 钮返回[Frequencies]主对话框。

(四) 可单击[Statistics...]按钮,弹出[Frequencies:Statistics]子对话框,并单击相应项目,在作频数表分析的基础上,附带作各种统计指标的描述,特别是可进行任何水平的百分位数计算。这里不选。

(五) 可单击[Charts...]钮,弹出[Frequencies:Charts]子对话框,用户可选三种图形:直条图(Bar chart)、饼图(Pie Charts)和直方图(Histogram)。这里选择[Histogram]项,并选择[With Normal Curve]要求绘制正态曲线。单击[Continue]按钮返回[Frequencies]主对话框,再单击[OK]钮即可得到(累计)频数(频率)分布表(如表 3-13所示)和直方图(如图 3-31所示)。

应该注意的是,SPSS 在未特别指定的情形下,直方图或频数分布表是按照原始数值逐一作频数分布的,这与日常需要的等距分组、且组数保持在一定数目的要求不符。因此,在调用[Frequencies]统计过程命令之前,可先对原始数据进行预处理:已知最小值为 84,最大值为 128,全距为 44,故可要求分成 5 组,起点为 80,组距为 10。选择[Transform]>[Recode]>[Into Different Variable...],在弹出的[Recode Into Different Variable]对话框中选定 X,单击按钮使之进入[Numeric Variable→Output Variable]列表框,在[Output Variable]栏的[Name]文本框中输入 x1,单击[Change]按钮表示新生成的变量名为 x1。单击[Old and New Values]按钮弹出[Recode Into Different Variable:Old and New Values]子对话框,在[Old Value]选项中单击[Range]项,输入第一个分组的数值范围:80~89,在[New value]栏内输入新值:80,单击[Add]按钮,依此将各组的范围及对应的新值逐一输入,最后单击[Continue]按钮返回,再单击[OK]按钮即完成。系统在原数据库中生成一新变量为 x1,这时再调用[Frequencies]统计过程将输出等距分组且组数为 5 的频数分布表。



图 3-30 频数分析对话框

表3-13 频数(率)分布表

	Frequency	Percent	Valid Percent	Cumulative Percent
84	1	3.3	3.3	3.3
85	1	3.3	3.3	6.7
88	1	3.3	3.3	10.0
91	2	6.7	6.7	16.7
94	1	3.3	3.3	20.0
95	1	3.3	3.3	23.3
96	1	3.3	3.3	26.7

97	1	3.3	3.3	30.0
99	1	3.3	3.3	33.3
101	2	6.7	6.7	40.0
103	1	3.3	3.3	43.3
105	3	10.0	10.0	53.3
106	4	13.3	13.3	66.7
107	2	6.7	6.7	73.3
109	1	3.3	3.3	76.7
110	1	3.3	3.3	80.0
111	2	6.7	6.7	86.7
118	1	3.3	3.3	90.0
119	1	3.3	3.3	93.3
121	1	3.3	3.3	96.7
128	1	3.3	3.3	100.0
Total	30	100.0	100.0	

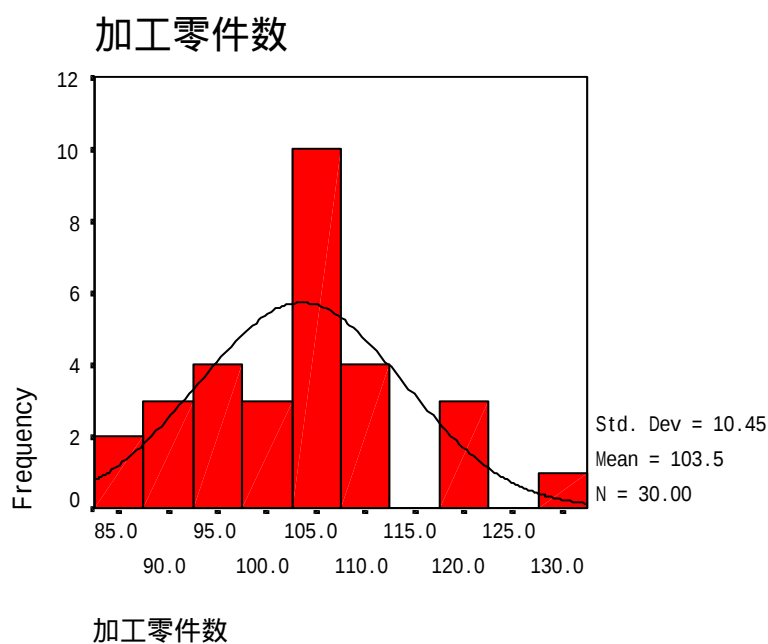


图 3-31 频数分析直方图

第四节 统计数据的描述

将数据整理成频率（频数）分布后，数据的数量规律性就可以大致地呈现在分布的类型和特点上。但频数分布给予我们的中是一个大致的分布形状，还缺少代表性的数量特征值精确地描述出不同的统计数据分布。作为统计数据的代表值，一个是分布的中心，反映分布的集中趋势，另一个是分布的形状，反映分布的离散程度。

一、 分布的中心

定义分布的中心有许多不同的方式。这里介绍三种最常用的,即众数、中位数和平均数。

(一) 众数(mode)

众数表示流行、时兴之意,有众多的意思。因而一个分布的众数就定义为频数出现最多的变量值。在正态分布和一般的偏态分布中,分布曲线最高点所对应的数值即是众数。如果没有明显的最高点,众数可以不存在。当然,如果有两个最高点,也可以有两个众数。众数很容易求得,一般只要看一眼即可。它特别适用于描述定类变量和定序变量的数据。定距变量的数据分组后也可近似地用某个组的组中值来表示众数的大小。但众数并不是一个描述中心的很好的代表值,它常常依赖于数据的分组情况,即分组数改变的话众数可能就会有较大的变化。而且众数也可能不唯一。

(二) 中位数(median)与分位数

中位数是数据排序后,位置在最中间的数值。显然,中位数将数据分成两半,一半数据比中位数大,一半数据比中位数小。用中位数来代表总体标志值的一般水平,可以避免代表值受数列中极端值的影响,稳定性比较好,有时更有代表性。

与中位数相似的还有四分位数(quartiles)、十分位数(decile)和百分位数(percentile)。中位数是将统计分布从中间分成相等的两部分,而四分位数就是将数据分布四等分的三个数值,其中中间的四分位数就是中位数。十分位数和百分位数分别是将数据分布十等分和一百等分的数值。

(三) 平均数(均值)(mean)

平均数是数据集中趋势的最主要测度值。如果数据是未经整理的原始数据,一般用下面的公式计算

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} \quad (3-1)$$

对于例 3-1 中的数据,可用(3-1)式计算该车间日加工零件平均数:

$$\bar{X} = \frac{106 + 84 + \cdots + 101}{30} = 103.5 \text{ (件)}$$

这个结果是日加工多的数据和日加工少的数据相互抵消后的结果,反映了该车间工人日加工零件数量的一般水平。从频数分布表上也可以看出平均数 103.5 件是数据分布的中心。

对于已经分组并形成频数分布的数据,可用下面的公式计算平均数:

$$\bar{X} = \frac{X_1 F_1 + X_2 F_2 + \cdots + X_m F_m}{F_1 + F_2 + \cdots + F_m} = \frac{\sum_{i=1}^m X_i F_i}{\sum_{i=1}^m F_i} \quad (3-2)$$

上式中 X_i 是频数分布中变量分组的组中值, F_i 是各组的频数。这里的 m 表示分组的组数,不再如式(3-1)直接表示数据的个数,因而这里用 m 表示以与 n 加以区别。可以用(3-2)计算例 3-1 日加工零件平均数。

平均数是统计学非常重要的基础内容,因为任何统计推断和分析都离不开平均数。

表 3-14 三个中心度量的比较

众数	中位数	平均数
主要适用于定类变量	主要适用于定序变量	适用于定距或定比变量
最不稳定	较平均数的稳定性差	最稳定
可容易计算,但不是永远存在,最不合适作为集中趋势代表值	只需中间的数据	计算时要用到全部数据,数据信息提取得最充分
有时候对个别值的变动也很敏感	对极端值不敏感	受极端值的影响
分组变化时影响较大	分组变化时有些影响	分组变化时影响不大

二、分布的形状

上面介绍了如何描述分布的中心,其中均值是最重要的一种代表值。但是只从均值来看待数据是片面的,我们还必须考虑数据的分布形状。用于描述数据分布形状即分布关于其中心的波动程度的代表值有:极差、内距、方差和标准差等,它们描述了分布的离散程度和差异程度。

(一) 极差(range)

极差也称为全距,是最大值与最小值之间的距离,它是数据离散或差异程度的最简单测度值,即

$$\text{极差 } R = \max(X_i) - \min(X_i)$$

例如,例 3-1 的数据中,极差为 $128-84=44$ (件)。显然,数据的离散程度大,极差就越大。极差虽然很容易计算,但它只告诉我们数据分布范围,至于分布的中间部分是如何变化的则不得而知。而且它受极端值的影响可能是很大的。

(二) 内距(Inter-Quartile Range, IQR)

内距又称为四分位差,是两个四分位数之差,即内距 $IQR = \text{高四分位数} - \text{低四分位数}$ 。与极差类似,内距也是由两个值之差决定的,也是不全面的。但由于这两个值之差代表了中间 50% 部分的长度,所以比极差能较好地描述分布的特性。例如,若内距比较小,则说明数据比较集中在中位数附近;反之则比较分散。内距常和中位数一起来描述一个定距特别是定序测量数据的分布。

(三) 方差(variance)和标准差(standard deviation)

方差是离差平方的平均数,即

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - u)^2}{N} \quad (3-3)$$

或

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (3-4)$$

(3-3) 式是总体方差的计算公式, σ^2 表示总体方差, u 表示总体均值。(3-4) 式是样本方差的计算公式, S^2 表示样本方差, $\bar{x}$ 是样本均值, n 是样本容量, $n-1$ 称为自由度。所谓自由度(Degree of Freedom)就是可以自由取值的变量个数,计算样本方差时, n 个数据在样本均值 $\bar{x}$ 确定后只有 $n-1$ 个数据可以自由取值,而第 n 个一定不能自由取值,所以其自

由度为 $n-1$ 。

对于分组数据的方差，与分组数据的均值一样，还要考虑各组的频数，即要对其离差平方和加权，有

$$\sigma^2 = \frac{\sum_{i=1}^m (X_i - \mu)^2 f_i}{\sum_{i=1}^m f_i} \quad (3-5)$$

$$S^2 = \frac{\sum_{i=1}^m (x_i - \bar{x})^2}{\sum_{i=1}^m f_i - 1} \quad (3-6)$$

标准差是方差的正平方根，即

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad \text{或者} \quad \sigma = \sqrt{\frac{\sum_{i=1}^M (X_i - \mu)^2 f_i}{\sum_{i=1}^M f_i}} \quad (3-7)$$

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad \text{或者} \quad S = \sqrt{\frac{\sum_{i=1}^m (x_i - \bar{x})^2 f_i}{\sum_{i=1}^m f_i - 1}} \quad (3-8)$$

由以上方差和标准差的公式可以看出，方差是以平方的形式使有正有负的离差变成正的离差，由于先对离差平方，扩大了离差，就有必要再开方根而得到标准差。方差在统计中被广泛应用

三、偏度与峰度

前面讨论了分布的集中趋势和离散趋势。要全面了解分布的特点，仅了解分布的集中趋势和离散程度是不够的，还需要了解分布是否对称和集中趋势高低等特征。偏度和峰度就是对分布的进一步描述。

(一) 偏度(skewness)

所谓偏度是指反映频数分布偏态方向和程度的测度。从方向上看，偏度分左偏和右偏两种。在频数分布中，最大集中点以上（频数曲线图横轴上众数的右边）的频数占总频数的一半多，称为右偏或正偏，如图 3-28（B）中的左图；最大集中点以下（频数曲线图横轴上众数的左边）的频数占总频数的一半多，称为左偏或负偏，如图 3-28（B）中的右图。左偏和右偏是最常见的偏态，尤其是右偏多于左偏。从程度上，偏斜度有大小之分，极端的偏斜状态有 J 形分布（如图 3-28(C)所示）和 U 形分布（如图 3-28(D)所示）两种。偏度的计算公式为：

$$\alpha = \frac{\sum (x - \bar{x})^3 / n}{\left[\sqrt{\sum (x - \bar{x})^2 / n} \right]^3} \quad (3-9)$$

当 $\alpha=0$ 时，表示分布是正态的或对称的；当 $\alpha>0$ 时，表示右偏或正偏；当 $\alpha<0$ 时，表示左偏或负偏。 α 越接近于 0，表示分布偏斜程度越小。

(二) 峰度(kurtosis)

所谓峰度，是指频数分布曲线高峰的形态，即反映分布曲线的尖锐程度的测度。在频数分布中，有的频数分布曲线与正态曲线相比是尖顶，有的则是平顶，无论是尖顶或平顶，峰度就是用来衡量频数分布曲线的高耸程度的一个数字特征。峰度的高低一般以正态分布的高

峰作为比较的标准。在单峰（分布曲线只有一个高耸点）的频数分布中，若均值所在组的频数特别大，其上下各组的频数先陡然地下降继而缓和地减少，曲线的高峰必定较正态峰高而狭，称为尖顶高峰（或高狭峰）；若中间约有半数的频数相当均匀，曲线的高峰必定较正态峰低而阔，称为平顶高峰（或低阔峰），如图 3-32 所示。

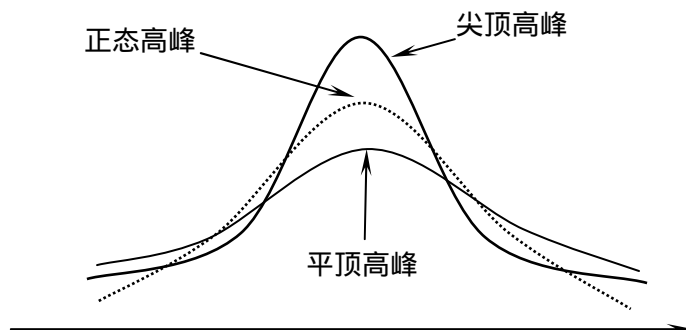


图 3-32 峰度

峰度的计算公式为

$$\beta = \frac{\sum (x - \bar{x})^4 / n}{[\sum (x - \bar{x})^2 / n]^2} - 3 \quad (3-10)$$

当 $\beta = 0$ 时，表示分布的峰度是正态分布的峰度；当 $\beta > 0$ 时，表示分布曲线的高峰是尖顶高峰；当 $\beta < 0$ 时，表示分布曲线的高峰是平顶高峰。

四、SPSS 操作

在 SPSS 中计算例 3-1 各种指标的步骤为：

（一）定义加工零件数的变量名为 X，并输入原始数据。

（二）选择[Analyze]⇒[Descriptive Statistics]⇒[Descriptives...]，打开[Descriptives]主对话框（如图 3-33 所示）。在主对话框左边列表选定变量 X，单击按钮使之进入[Variables(s)]列表框。

（三）单击[Options...]按钮，打开[Descriptives : Options]子对话框。选择均值(Mean)、总和(Sum)、标准差(Std.Deviation)、方差(Variance)、极差(Range)、最小值(Minimum)、最大值(Maximum)、偏度 (Skewness) 和峰度(Kurtosis)，选好后单击[Continue]按钮返回 [Descriptives] 主对话框，再单击[OK]按钮即可得到各种统计量的计算结果（如图 3-34 所示）。

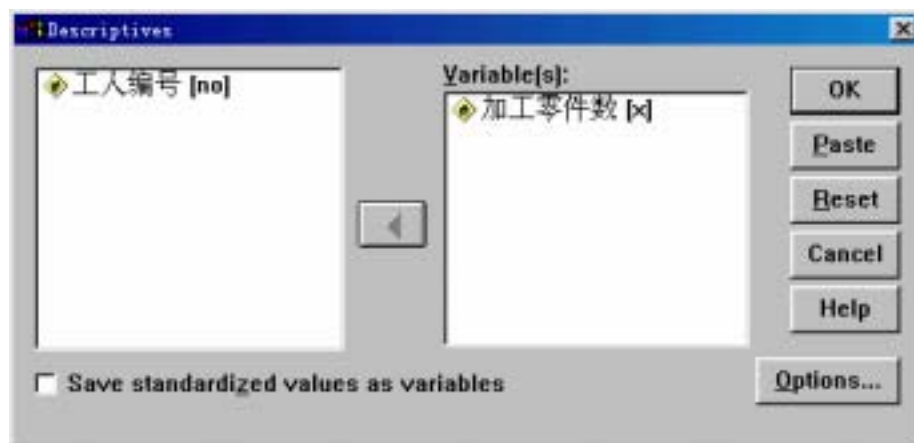


图 3-33 描述性统计主对话框

图3-34 描述统计量

加工零件数	Statistic (统计量)	N	30
		Range (极差)	44
		Minimum (最小值)	84
		Maximum (最大值)	128
		Sum (和)	3105
		Mean (均值)	103.50
		Std. Deviation (标准差)	10.45
		Variance (方差)	109.224
		Skewness (偏度)	.149
		Kurtosis (峰度)	.050
	Std. Error	Skewness (偏度标准误)	.427
		Kurtosis (峰度标准误)	.833
Valid N (listwise)	Statistic	N	30

以上结果没有给出中位数、众数等统计量，可以在频数（率）分析时增加选项计算相应的统计量，具体操作步骤如下：

（一）定义工人编号和加工零件数的变量名分别为 NO 和 X，然后输入变量 NO 和 X 的原始数据。

（二）选择[Analyze]=>[Descriptive Statistics]=>[Frequencies...]，弹出[Frequencies]主对话框（如图 3-30 所示）。现欲 X 进行频数分析，在对话框左侧的变量列表中选 X，单击按钮使之进入[Variable(s)]列表框，并选择[Display Frequency Tables]显示频数分布表。

（三）单击[Statistics...]按钮，弹出[Frequencies:Statistics]子对话框，并单击相应项目。本例中选择均值(Mean)、中位数(Median)、众数(Mode)、总和(Sum)、标准差(Std.Deviation)、方差(Variance)、极差(Range)、最小值(Minimum)、最大值(Maximum)、偏度 (Skewness) 和峰度(Kurtosis)，选好后单击[Continue]按钮返回[Frequencies]主对话框，再单击[OK]按钮即可得到各种统计量的计算结果（如表 3-15 所示）。

表3-15 统计量

加工零件数	
N (有效样本容量)	30
Mean (均值)	103.50
Median (中位数)	105.00
Mode (众数)	106
Std. Deviation (标准差)	10.45
Variance (方差)	109.22
Skewness (偏度)	.149
Std. Error of Skewness (偏度标准误)	.427
Kurtosis (峰度)	.050
Std. Error of Kurtosis (峰度标准误)	.833
Range (极差)	44
Minimum (最小值)	84
Maximum (最大值)	128

前面介绍的统计数据描述方法一般是先将数据分组,然后将分组数据画成直方图或折线图观察数据的分布规律性。这种传统的数据整理方法的局限性表现为整理后就损失了原始数据的信息,用分组数据计算的平均数就只能使用近似公式(3-2)。为了能更简单、直观地描述统计数据的分布特征,并能根据数据的特点选择适当的分析工具探索数据的内在数量规律,国外在二十世纪七十年代末出现了探索性数据分析统计新领域。探索性数据分析在一般描述性统计指标的基础上,增加有关数据其他特征的文字与图形描述,显得更加细致与全面,有助于用户思考对数据进行进一步分析的方案。其中探索性数据分析中最简单的是茎叶图和箱线图。

这种数据整理方法将传统的统计分组与画直方图两步工作一次完成。既保留了数据的原始信息，又为准确计算均值等提供了方便和可能。下面以例 3-1 的数据来画茎叶图。

由表 3-9 可以看出,第 1 号工人日加工零件数 106 件,这个数可分成两部分,树茎的 10 和树叶的 6,就在树茎 10 的右边写上第一个树叶 6。第 2 号工人日加工零件 84 件,就在树茎 8 的右边写上第一个树叶 4。按照这种方法可将 30 个工人加工的零件数全都分成树茎和树叶两部分,按照每一数字十位数和百位数的数值选定树茎并写上树叶。在画图时,要注意树叶竖行要对齐,这样,树叶的个数是各组的频数。当我们将图画好后,不难看出这就是一个放倒了直方图,各树茎上树叶的个数就是各组的频数。在茎叶图画好后,不仅可以一目了然地看出频数分布的形状,而且茎叶图中还保留了原始数据的信息。这使得我们在进一步计算数据平均数和中位数时,就可以计算准确的数值而不必应用近似公式了。利用茎叶图进行分组还有一个好处,就是在连续数据的分组中,不会出现重复分组的可能性。在现代统计分析软件中,绝大多数都有做茎叶图的功能,为我们应用探索性数据方法分析复杂问题提供了方便。

箱线图是由一组数据 5 个特征绘制的一个箱子和两条线段的图形,这种直观的箱线图不仅能反映出一组数据的分布特征,而还可以进行多组数据的分析比较。

图 3-35 简单箱线图



三、SPSS 操作

在 SPSS 中绘制茎叶图和箱线图进行探索性分析的操作步骤如下：

(一) 定义加工零件数的变量名为 X，并输入原始数据。

(二) 选择[Analyze]⇒[Descriptive Statistics]⇒[Explore...]，打开[Explore]主对话框（如图 3-36 所示）。在主对话框左边列表中选定变量 X，单击按钮使之进入[Dependent List]列表框。

(三) 单击[Plot...]按钮打开[Explore:Plot]子对话框（如图 3-37 所示），在[Boxplot]栏内选[Factor levels together]项要求按组别进行箱图绘制；在[Descriptive]栏内选[Stem-and-leaf]项要求作茎叶图描述。然后单击[Continue]按钮返回[Explore]主对话框，再单击[OK]按钮即可得到探索性统计分析结果（如图 3-38 所示），该结果与手工绘制的图形形状有些不同，但它们是一致的。

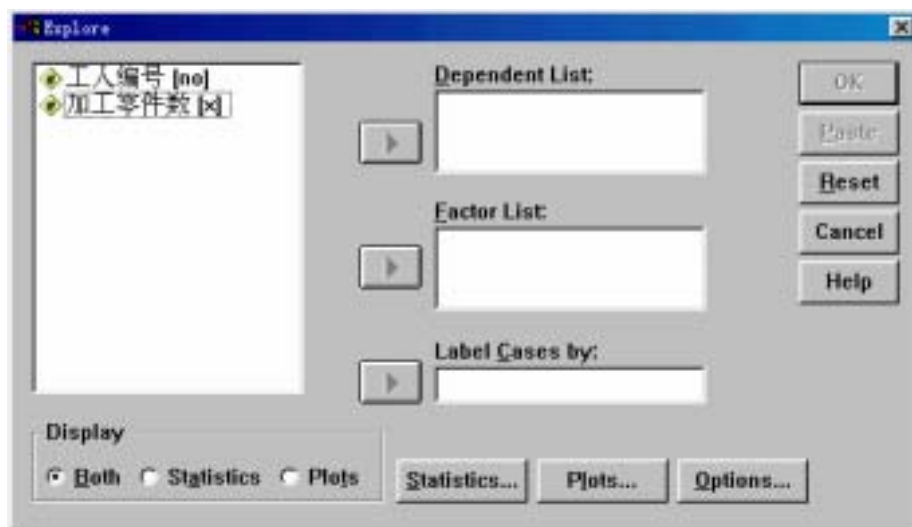


图 3-36 探索性分析对话框

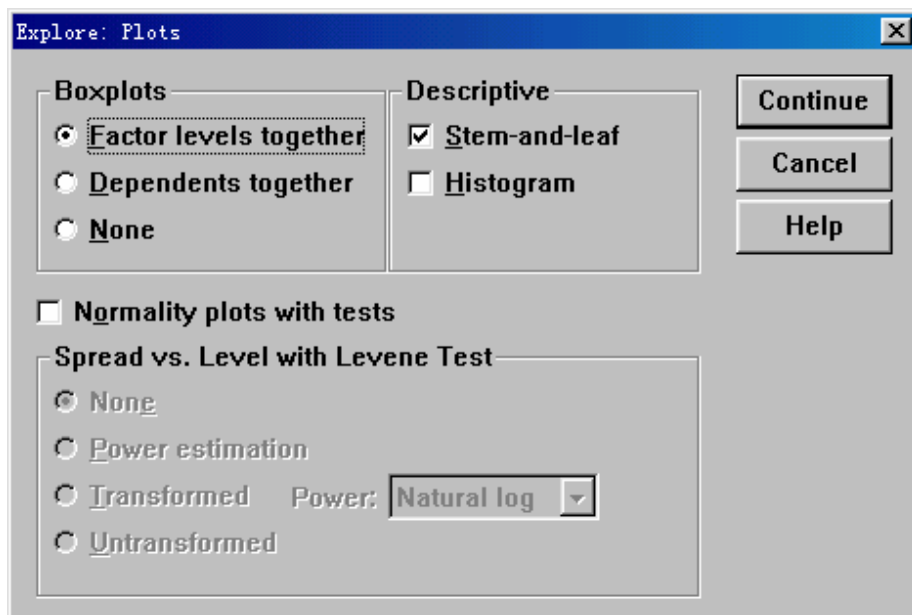


图 3-37 探索性分析绘图子对话框

加工零件数 Stem-and-Leaf Plot

Frequency	Stem & Leaf
1.00	8 . 4
2.00	8 . 58
3.00	9 . 114
4.00	9 . 5679
3.00	10 . 113
10.00	10 . 5556666779
3.00	11 . 011
2.00	11 . 89
1.00	12 . 1
1.00	12 . 8
Stem width: 10	
Each leaf: 1 case(s)	

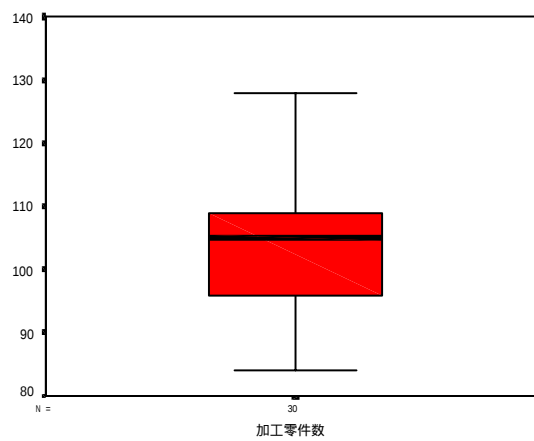


图 3-38 茎叶图与箱线图

第四章 总体与样本的描述

第一节 总体、样本与随机变量

一、总体与样本

我们把所要研究对象的全体称为总体(Population)或母体。组成总体的最小研究单位称为个体。例如,每亩地的粮食产量,可能是从 100 斤到 2000 斤之间的任意数值,从 100 斤到 2000 斤之间的所有值就是每亩粮食产量的总体,其中的每一个值就是个体(单位)。总体所包含的个体数目,称为总体容量。总体容量随着研究对象的不同而异。例如,国内生产总值(GDP)可以在一个地区范围内研究,也可以在全国范围内研究,也可以在世界范围内研究。在第一种情况下总体是由一个地区的 GDP 构成,在第二种情况下总体是由国家所有各地区的 GDP 构成,在第三种情况下,总体是由世界上所有国家的 GDP 构成的。当一个总体是由有限个个体组成时,那么这个总体就称为有限总体;当它是由无限个个体构成时,称为无限总体。

总体中的每一个体,具有相同的观察特征。我们把这种特征作为不同总体的区别指标。在研究问题时,人们对表征总体状况的某一个或某几个数量指标的情况感兴趣,并把总体看作所研究对象的若干数量特征的全体。这些数量特征在总体中往往随概率不同而有不同的取值。

实际上,在大多数情况下,我们是不知道总体的全部个体的,通常只能从总体中抽出若干个体。我们把从总体中抽出若干个体而组成的集体称为样本(Sample),样本中所含个体的个数,称为样本容量。从总体抽取样本的过程称为抽样(见第四章)。按照随机原则进行重复抽样所得的样本,称为简单随机样本(Simple Random Sample)。

二、随机变量

在研究实际问题时,要对客观事物进行观察。观察的过程称为试验,试验的结果称为事件。事件可以分为确定性事件和随机性事件两种。在给定条件下,一定发生或一定不发生的事件分别称为必然事件和不可能事件,它们是确定性事件;在给定条件下的每一次试验中可能发生也可能不发生,而在大量试验中具有某种规律性的事件称为随机事件。有的随机事件可采用数量标识表示。如检验一批零件,可能出现不同尺寸长度或直径;抛掷一颗骰子,可能出现的点数为 1, 2, 3, 4, 5, 6。显然,这些随机事件都是采用数量标识表示的。有的不采用数量标识表示,如检验一批产品,没检验一件可能出现合格的或不合格的;抛掷一枚钱币,每抛一次可能出现正面或反面。显然这些随机事件都是不采用数量标识表示的。为了把随机事件数量化,以便于数学上的处理,有必要把不采用数量标识的化为采用数量标识的。如把每检验一件产品可能出现合格的指定为 0,可能出现不合格的指定为 1;或把每次抛掷一枚钱币可能出现正面的指定为 0,可能出现反面的指定为 1。这样把指定的 0、1 与合格、不合格一一对应,或把 0、1 与正面、反面一一对应,就可以把随机事件完全数量化了。为了研究随机事件的数量规律性,把表征随机事件的变量称为随机变量(Random Variable),记为 X 、 Y 等。

假设 X 是一个随机变量,随着观察数目的增加,可以发现其中任何值 X_i 的频率趋于某一稳定值,此值即为频率的极限值,或称之为随机变量值 X_i 的概率。换句话说,一个随机变量值 X_i 的概率就是当变量的观察值总数趋于无穷时,其频率的极限值,即

$$P(X = X_i) = \lim_{n \rightarrow \infty} \frac{f_i}{\sum f_i}$$

这里用频率的极限给出了概率的定义,不仅可以使初学者易于理解,同时也便于根据经验数据予以估算。但是,应当注意的是频率并不等于概率,它们是两个不同的概念,只有当观察值的数目趋于无穷时,频率才趋于概率。任何事件或任一随机变量值的概率的取值是从 0 到 1 之间的任何值,即

$$0 \leq P(X_i) \leq 1$$

如果随机变量 X 取特定值 X_i 的概率为零,这就意味着 X 不能取 X_i 值,也就是说, X_i 的概率为零表示 X_i 值不可能出现。如果随机变量 X 取一个特定值 X_i 的概率为 1,则意味着其任何时候任何情况下都必定出现,也就是说 X_i 是这个变量唯一可能取的值。在这种情况下,该变量实际上是一个常数,即为 X_i 。随机变量 X 取任何特定值 X_i 的概率在 0 到 1 之间,则表示变量 X 的这个特定值的出现具有不确定性,随机变量的取值与概率有关。因此,也可以把随机变量定义为根据概率不同而取不同数值的变量。

三、总体、样本与随机变量之间的关系

由上面我们已经了解到,表示总体状况的某一个或某几个数量特征在总体中往往随概率不同而取不同的值。显然,对于这样的数量特征,用一般的变量是无法加以描述的,能够对之给予描述的是一类特征的变量,即随机变量。

我们对于总体中某一个体所具有的特殊属性往往并不关心,真正感兴趣的是表示总体特征的数量指标。例如,全天生产的 10000 个灯泡中寿命在 1000 小时到 1200 小时之间、1000 小时以下以及 1200 小时以上的灯泡所占的百分比等。就总体的某一数量特征 X 而言,如灯泡的使用寿命,每个个体的取值不一定完全相同,但它是按照一定规律分布的,如 10000 个灯泡中各种寿命灯泡所占的比例是基本稳定的。因此,对于一个总体来说,其第一个数量特征完全随机变量的定义,是根据概率的不同而取不同值的变量,即总体中每一个数量特征就是一个随机变量。由于我们主要是研究总体的数量特征,所以把总体看成是具有若干数量特征的研究对象的全体,可直接用一个随机变量来表示。

因此,所谓总体就是一个随机变量,所谓样本就是 n 个(样本容量为 n)相互独立且与总体有相同分布的随机变量 $x_1, x_2, \dots, x_n$ 。每一次具体抽样所得的数据,就是 n 元随机变量的一个观察值(样本值),记为 $(x_1, x_2, \dots, x_n)$ 。

样本是总体的一小部分,是对总体进行随机抽取后所得到的集合。对于观察者来说,整个总体的状况是不了解的,观察者所能了解的只是总体的一部分——样本的具体状况,我们所要做的就是通过对这些样本具体状况的研究,来推知整个总体的状况。下图的箭头表示了研究方向。既然要通过研究样本来推知总体的状况特性,那么通过什么能把样本和总体联系起来,使我们的研究在二者之间得以沟通呢?答案是通过总体及随机变量函数的定义、分类和分布。

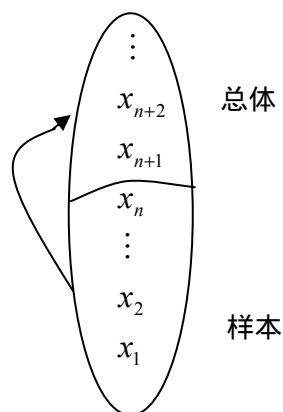


图 4-39 由样本推断总体

第二节 总体与随机变量的描述

总体就是一个随机变量。因而，对总体的描述就是对随机变量的描述。随机变量可以分为两种类型。如果随机变量的所有可能取值都可以按一定顺序一一列举出来，则称为离散型的随机变量，如人口数都属于有限的离散型随机变量。如果随机变量的所有可能取值是充满某一区间，无法按顺序一一列举，则称为连续型的随机变量。例如人的身長体重等等，都是连续型的随机变量。

一、随机变量（总体）的概率分布

（一）离散型随机变量的概率分布

虽然我们无法知道一次试验可能出现的结果，但是所有可能结果则是已知的，而且其相应的出现概率也已确定。就是说：随机变量 X 的可能取值及其概率都是已知的，我们就可将变量 X 的取值以及相应概率按顺序排列起来，以显示 X 的概率分布情况。

1、概率函数

概率分布可用概率函数来描述。设随机变量 X 的可能取值为： $X_1, X_2, \dots, X_N$ ，其相应的概率为 $p_1, p_2, \dots, p_N$ ，则概率函数为

$$P(X=X_i)=p_i (i=1,2,\dots,N)$$

因此，随机变量 X 的概率分布（简称分布）就是 X 在各个可能取值上出现的概率大小情况。分布具有如下性质：

（1）随机变量 X 取值的概率都是非负的。即

$$p_i \geq 0 \quad (i=1,2,\dots,N)$$

（2）随机变量 X 所有取值的概率总和等于 1。即

$$\sum_{i=1}^N p_i = 1 \quad (i=1,2,\dots,N)$$

例：连续投掷两次硬币，求正面向上的次数的概率分布。

连续投掷两次硬币，正面向上的次数可能为 0 次、1 次、2 次三种，其概率分布为：

$$P(X=0)=\frac{1}{2} \cdot \frac{1}{2}=\frac{1}{4}$$

$$P(X=1)=\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}=\frac{2}{4}$$

$$P(X=2)=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$$

2、分布表

设随机变量 X 的可能取值为 $X_1, X_2, \dots, X_N$, 其相应的概率为 $P(X=X_i)=p_i (i=1,2,\dots,N)$ 。将这些结果列表如下：

X	X_1	X_2	...	X_i	...	X_N
p	p_1	p_2	...	p_i	...	p_N

该表称为 X 的(概率)分布表, 它具体而完整地描述了随机变量 X 取值的概率分布情况。

例如上例连续两次投掷硬币正面向上的次数 X 的分布表为：

X	0	1	2
p	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

3、分布图

将这些资料表示在直角坐标系上, 以 X 为横轴, 以 p 为纵轴, 以坐标 $(X_1, p_1), (X_2, p_2), \dots, (X_N, p_N)$ 构成平面上各点。联结各点就能形象地表明概率 P_i 的分布情况 (如图 4-40 所示)。

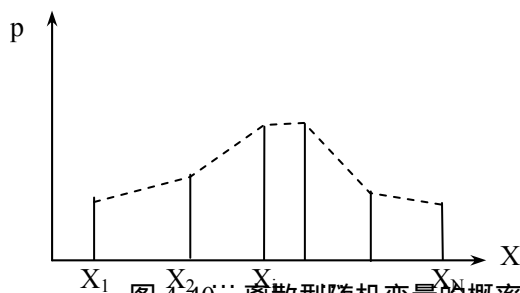


图 4-40 离散型随机变量的概率分布图

例如上例连续两次投掷硬币正面向上的次数 X 的分布图为：

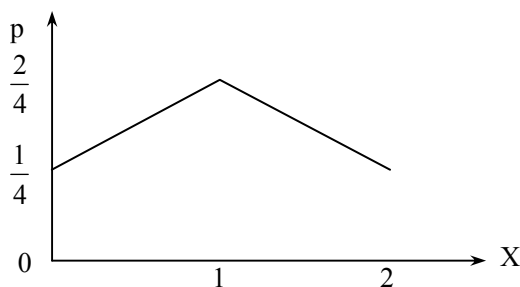


图 4-41

4、分布函数

我们还可以用另一种方法来描述离散型随机变量的概率分布, 即以 X 取值小于实数 x 的概率来描述概率分布的情况。 X 取值小于 x 的概率为

$$P(X < x) = \sum_{X_i < x} P(X = X_i) = \sum_{X_i < x} p_i \quad (4-1)$$

对于任意两个实数 $x_1 < x_2$, $P(x_1 \leq X < x_2) = P(X < x_2) - P(X < x_1)$ 。这说明, 如果以任何给定的实数 x , 概率 $P(X < x)$ 确定的话, 则概率 $P(x_1 \leq X < x_2)$ 也就确定了。所以掌握了分布函数 $F(x) = P(X < x)$ 那么这一随机变量任何取值的概率都可以由此给出。

例如上例连续两次投掷硬币正面向上的次数 X 的分布函数可以表达为：

$$F(x) = \begin{cases} 0 & \text{当 } -\infty < x < 0 \\ \frac{1}{4} & \text{当 } 0 \leq x < 1 \\ \frac{3}{4} & \text{当 } 1 \leq x < 2 \\ 1 & \text{当 } 2 \leq x < \infty \end{cases} \quad (4-2)$$

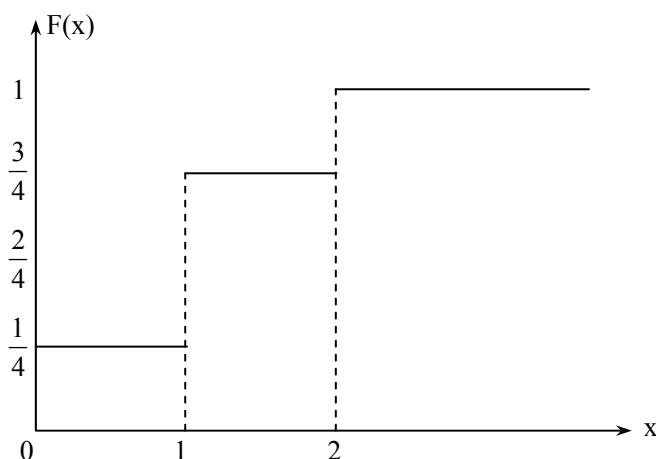


图 4-42

如图 4-42 所示，从图形中可以直观地看到 $P(X \leq x)$ 的数值随着 x 的增加而递增。在点 $x=k(k=0,1,2)$ 处分布函数的增加值等于这一点的概率 $P(X=k)$ ，当 $x < 0$ 时， $P(X \leq x)=0$ ，当 $x > 2$ 时， $P(X \leq x)=1$ 。

5、常见的离散型随机变量分布

要掌握每一种研究总体的概率分布是比较麻烦的，但经过大量观察，常用的离散型随机变量理论分布有 0-1 分布、二项分布、泊松分布、几何分布等等。下面简单介绍二项分布和泊松分布。

(1) 二项分布(Binomial Distribution)

二项分布又称为贝努里 (Bernoulli) 分布，是一种具有广泛应用的离散型随机变量的概率分布。二项分布研究的是试验仅有两种结果的分布（这种试验称为贝努里试验），如某产品质量合格与不合格等。其定义为：

设有 n 次试验，各次试验是相互独立的，每次试验某事件出现的概率都是 p ，某事件不出现的概率都是 $1-p$ ，记为 q ，则对于某事件出现 k ($k=0,1,2,\dots,n$) 次的概率分布为：

$$P(X = X_i) = P(X = k) = C_n^k p^k q^{n-k} \quad (4-3)$$

式 (4-3) 即为二项分布。

例如，按照规定，某种型号电子管的使用寿命超过 5000 小时的为一级品。已知某一大批产品的一级品率为 0.2 现在从中随机地抽查 20 只。问 20 只管子中恰有 k 只 ($k=0,1,2,\dots,n$) 管子为一级品的概率是多少？

这里抽查的管子数量相对于管子的总数来说很少，可以当作重复抽样来处理，将产品是否是一级品看成是一次试验的结果，检查 20 只管子相当于做 20 次贝努里试验。以 X 记 20 只管子中一级品的只数，那么， X 是一个随机变量，它服从 $n=20, p=0.2$ 的二项分布，由 (4-3) 式即可得所求的概率为：

$$P(X = k) = C_{20}^k (0.2)^k (0.8)^{20-k} \quad k = 0, 1, \dots, 20$$

(2) 泊松分布(Poisson Distribution)

泊松分布是由法国人 S·D·泊松提出来的,也是一种重要的离散型随机变量的概率分布。当试验次数 n 相对地增多,且每次试验中某事件出现的概率很小,而 $np = \lambda$ 的值为大小适中的常数时,这时某事件恰好发生 k 次的概率分布即为泊松分布。用公式表示为:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (4-4)$$

其中, $P(X = k)$ 表示恰好是 k 出现的概率, $\lambda = np$ 表示每一时间间隔内事件出现的平均数。

泊松分布适用于描述某些稀有事件的现状或出现机会非常小的一些事件,通常称为“稀有事件法则”。例如,对于一个企业来说,地震、火灾等突发性灾害是稀有事件,电器产品爆炸、电话询问台呼叫占线、高速公路汽车相撞等都可以视为稀有事件,这些事件在一定时间内发生的概率均可以应用泊松分布来计算。

例如,某电话询问台在某段时间内呼叫占线的概率为 0.005,在该段时间内有呼叫电话 260 个,问呼叫占线有 4 次的概率是多少?

该例明显服从泊松分布,因有: $n=260$, $p=0.005$, $\lambda = np = 260 \times 0.005 = 1.3$, $k=4$, 则

$P(X = 4) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1.3^4 \times e^{-1.3}}{4!} = 0.0324$, 表明该电话询问台在某段时间内呼叫占线有 4 次的概率为 0.0324。

(二) 连续型随机变量的概率分布

1、概率函数

连续型随机变量的概率函数为

$$P(a \leq X < b) = \int_a^b f(x)dx \quad -\infty < X < +\infty \quad (4-5)$$

其中 $f(x)$ 为密度函数, 满足:

(1) 密度函数 $f(x)$ 是非负函数, 即 $f(x) \geq 0$;

(2) $\int_{-\infty}^{+\infty} f(x)dx = 1$ 。

2、分布函数

由于连续型随机变量 X 的取值充满着一个区间, 不能一一列出, 所以无法用分布表描述。通常用分布函数 $F(x) = P(X \leq x)$ 来描述概率的分布情况。通过密度函数 $f(x)$ 把它表示成积分的形式:

$$F(x) = \int_{-\infty}^x f(x)dx \quad (4-6)$$

由于 x 是一个具有连续分布的随机变量, 它的分布函数 $F(x)$ 存在导数, 而且 $F'(x) = f(x)$ 。根据导数的定义有:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x)}{\Delta x}$$

由此可见, 密度函数 $f(x)$ 表示随机变量 X 在点 x 上的概率密度。通常把密度函数的图形称为分布曲线。分布曲线 $y=f(x)$ 和 x 轴所包围的全体面积等于 1。

因此, 随机变量 X 落在区间 (x_1, x_2) 内的概率即概率函数值等于它的密度函数在该区间上的定积分。即

$$P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} f(x)dx \quad (4-7)$$

其几何意义就是概率 $P(x_1 \leq x < x_2)$ 等于区间 (x_1, x_2) 上分布曲线 $y=f(x)$ 和 x 轴围成的面积, 如图 4-43 中的阴影部分所示。

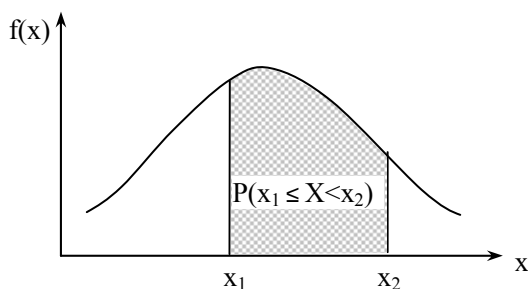


图 4-43

分布函数 $F(x) = \int_{-\infty}^x f(x)dx$ 具有 $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ 及 $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$, 介于 $F(x)=0$ 与 $F(x)=1$ 之间处处左连续, 而且单调上升的性质, 如图 4-44 所示。

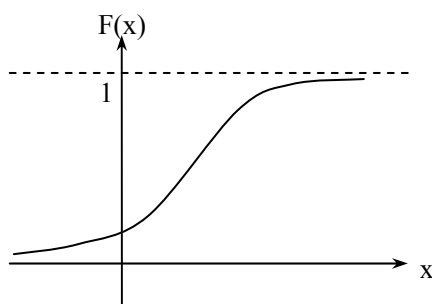


图 4-44 分布函数

3、最常用的连续型随机变量分布——正态分布

在统计中, 许多重要的分布都是连续型分布, 其中一种特别重要的连续型随机变量的概率分布就是正态分布(Normal Distribution)。正态分布最初为 De Moivre 于 1773 年发现, 其后, 拉普拉斯(Laplace)和高斯(Gauss)对它作出了很大的贡献, 尤其是高斯的贡献最为突出, 所以正态分布又称为高斯分布。我们将在第四节中作较详细的讨论。

二、随机变量(总体)的数学特征

随机变量的分布是对随机变量的一种最完整的描述, 然而要求出随机变量的分布往往不是容易的事。同时, 在很多情况下, 我们并不需要全面地考察随机变量的变化情况, 而只需了解它的一些综合指标就足够了。随机变量的数字特征(又称特征数、分布参数)就是指能集中反映随机变量概率分布基本特点的综合指标, 通过它我们就可以对随机变量和总体有一个粗略的认识。常见的离散型随机变量的数字特征有: 数学期望(均值)、方差(或标准差)、协方差、相关系数等。

(一) 离散型随机变量的数字特征

1、数学期望(Expected Value)(简称期望)

设离散型随机变量 X 的概率分布表为:

X	X_1	X_2	$\dots$	X_i	$\dots$	X_N
P	p_1	p_2	$\dots$	p_i	$\dots$	p_N

定义随机变量 X 的数学期望 $E(X)$ 为:

$$E(X) = \mu_x = X_1 p_1 + X_2 p_2 + \dots + X_i p_i + \dots + X_N p_N$$

可见, 期望实质上是随机变量的加权平均数, 是用于描述随机变量(或总体)的一般水平的。因此平均数的性质也完全适用于数学期望。

期望具有如下性质：

- (1) 若 c 为常数，则 $E(c)=c$ ；
- (2) 若 a 、 b 为常数，则 $E(aX+b)=aE(X)+b$ ；
- (3) 若 X 、 Y 为两个随机变量，则 $E(X+Y)=E(X)+E(Y)$ ；
- (4) 若 X 、 Y 为两个独立的随机变量，则 $E(XY)=E(X)E(Y)$ ；
- (5) 设 n 个随机变量 $X_1, X_2, \dots, X_n$ 其数学期望分别为 $E(X_1), E(X_2), \dots, E(X_n)$ ，则有： $E(X_1+X_2+\dots+X_n)=E(X_1)+E(X_2)+\dots+E(X_n)$ ；
- (6) 设 n 个随机变量 $X_1, X_2, \dots, X_n$ ，则 $E(X_1X_2\dots X_n)=E(X_1)E(X_2)\dots E(X_n)$ 。

2、方差(Variance)与标准差 (S.D. : Standard Deviation)

设离散型随机变量 X 的概率函数为 $P(X=X_i)=p_i, (i=1, 2, \dots, n)$ 。其数学期望 $E(X)$ ，简记为 EX 。定义随机变量 X 的方差 $Var(X)$ 为：

$$Var(X)=\sigma_X^2=\sum_{i=1}^n(X_i-EX)^2 p_i \quad (4-8)$$

方差还可以表示为：

$$Var(X)=E(X^2)-[E(X)]^2 \quad (4-9)$$

方差的平方根称为标准差或均方差，记为 σ ：

$$\sigma=\sqrt{Var(X)}=\sqrt{\sum_{i=1}^n(X_i-EX)^2 p_i} \quad (4-10)$$

方差和标准差用于描述随机变量的离散程度，即描述随机变量相对于它的期望值的偏离程度，这种偏离越大，说明随机变量的取值越分散。

方差具有如下性质：

- (1) 若 c 为常数，则 $Var(c)=0$ ；
- (2) 若 c 为常数， X 为随机变量，则 $Var(c+X)=Var(X)$ ；
- (3) 若 c 为常数， X 为随机变量，则 $Var(cX)=Var(-cX)=c^2Var(X)$ ；
- (4) 若 a 、 b 为常数， X 为随机变量，则 $Var(a+bX)=b^2Var(X)$ ；
- (5) 若 X 、 Y 为两个独立的随机变量，则

$$Var(X+Y)=Var(X-Y)=Var(X)+Var(Y)。$$

(6) 设 n 个独立随机变量 $X_1, X_2, \dots, X_n$ 其方差分别为 $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ ，则 $X_1+X_2+\dots+X_n$ 的方差 $Var(X_1+X_2+\dots+X_n)$ 有：

$$Var(X_1+X_2+\dots+X_n)=\sigma_1^2+\sigma_2^2+\dots+\sigma_n^2$$

但要注意，随机变量和的标准差通常不等于各变量标准差之和，而小于它们。

- (7) n 个独立随机变量平均的方差等于各变量方差平均数的 $1/n$ 。

设 n 个独立变量 $X_1, X_2, \dots, X_n$ ，其方差分别为 $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ 。设各变量的平均数为：

$$\bar{X}=\frac{X_1+X_2+\dots+X_n}{n}$$

则 $\bar{X}$ 的方差 $Var(\bar{X})$ 为：

$$Var(\bar{X})=\frac{\sigma_1^2+\sigma_2^2+\dots+\sigma_n^2}{n^2}=\frac{1}{n}\overline{\sigma_i^2}$$

$$\text{式中 } \overline{\sigma_i^2} = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}{n}$$

特别地，当每个随机变量的方差相等，即 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$ 时，有：

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \\ \sigma_{\bar{X}} &= \sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}} \end{aligned}$$

n 个独立变量平均数的方差，仅为各变量平均方差的 $1/n$ ，这一事实说明变量平均数的分布比各变量的分布更集中于总平均数的周围，因而以变量平均数来估计总平均数将更加接近真实。

3、协方差(Covariance)

设离散型随机变量 X 和 Y 的数学期望分别为 $E(X)$ 、 $E(Y)$ ，并简记为 EX 、 EY 。定义随机变量 X 与 Y 的协方差 $\text{Cov}(X, Y)$ 为：

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] \quad (4-11)$$

可以简记为：

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) \quad (4-12)$$

协方差用于描述两个变量之间线性相关的密切程度，其值大小与两个变量的量纲有关，不适于比较。

4、相关系数(Correlation Coefficient)

如果连续型随机变量 X 、 Y 的方差都不为零，则 X 与 Y 的相关系数（记为 ρ ）定义为：

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \quad (4-13)$$

ρ 是描述 X 与 Y 之间线性相关程度的一个数字特征，相当于是协方差的“标准化”，消除了量纲的影响。可证明 $|\rho| \leq 1$ ，适于比较。如果 $\rho = 1$ ，称 X 与 Y 完全线性相关；如果 $\rho = 0$ ，称 X 与 Y 无线性关系；如果 $\rho > 0$ ，称 X 与 Y 正相关；如果 $\rho < 0$ ，称 X 与 Y 负相关。如果两个随机变量相互独立，则它们的相关系数为 0；如果两个随机变量的相关系数为 0，这两个随机变量却未必独立。

（二）连续型随机变量的数字特征

1、数学期望

设连续型随机变量 X 的密度函数为 $f(x)$ ，定义 X 的数学期望 $E(X)$ 为：

$$E(X) = \mu_X = \int_{-\infty}^{\infty} xf(x)dx \quad (4-14)$$

可以把这一数学期望理解为随机变量无限和的加权平均。

离散型随机变量数学期望的性质也适用于连续型随机变量。

2、方差与标准差

设连续型随机变量 X 的密度函数为 $f(x)$ ，其数学期望为 $E(X)$ ，简记为 EX 。定义随机变量 X 的方差 $\text{Var}(X)$ 为：

$$\text{Var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - EX)^2 f(x)dx \quad (4-15)$$

标准差为

$$\sigma_x = \sqrt{\text{Var}(X)} = \sqrt{\int_{-\infty}^{\infty} (x - EX)^2 f(x) dx} \quad (4-16)$$

离散型随机变量方差的性质对于连续型随机变量仍然成立。

3、协方差

设连续型随机变量 X 和 Y 的数学期望分别为 $E(X)$ 、 $E(Y)$ ，并简记为 EX 、 EY 。定义随机变量 X 与 Y 的协方差 $\text{Cov}(X, Y)$ 为：

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] \quad (4-17)$$

可以简记为：

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) \quad (4-18)$$

4、相关系数

如果连续型随机变量 X 、 Y 的方差都不为零，则 X 与 Y 的相关系数（记为 ρ ）定义为：

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \quad (4-19)$$

ρ 是描述 X 与 Y 之间线性相关程度的一个数字特征，其性质同离散型随机变量间的相关系数。

第三节 样本的描述

一、样本分布

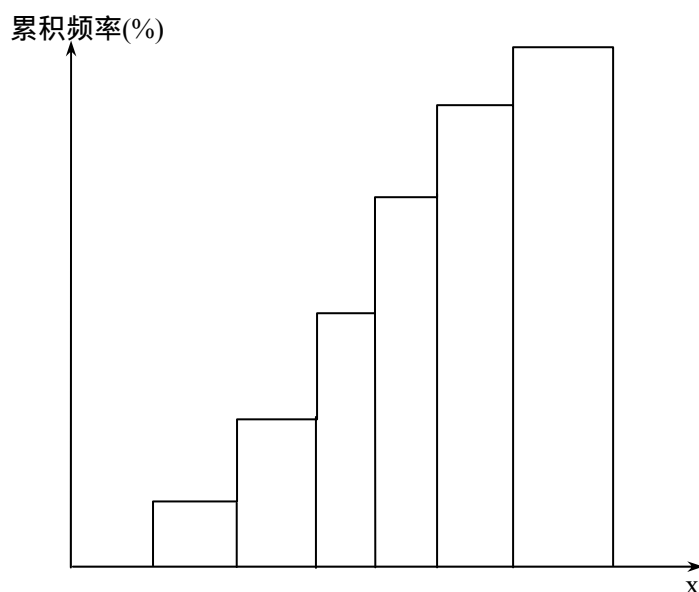
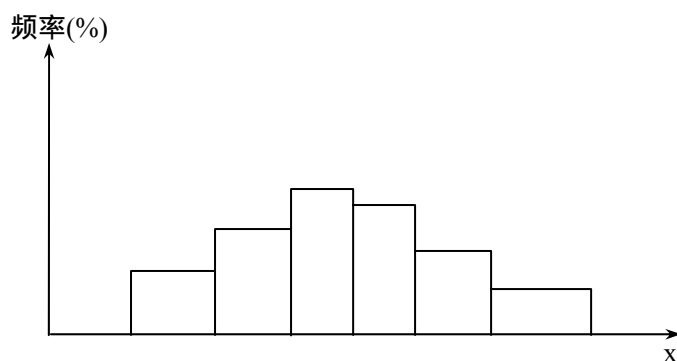
第一节讲到，总体就是一个随机变量，而样本就是 n 个相互独立且与总体有相同分布的随机变量 $x_1, x_2, \dots, x_n$ 。每一次具体抽样所得的数据，就是 n 元随机变量的一个观察值（样本值），记为 $(x_1, x_2, \dots, x_n)$ 。每当提到一个容量为 n 的样本时，常有双重涵义：一是指一个 n 元随机变量；二是指一次具体抽样的可能结果。

当样本指一个 n 元随机变量时，样本的分布就是 n 元随机变量的分布，该 n 个随机变量相互独立且其分布均与总体的分布相同。当样本指一次抽样的可能结果时，样本的分布可定义如下：

设 $(x_1, x_2, \dots, x_n)$ 是总体 X 的一个样本观测值，我们可以把这些观测数据按第三章的方法绘成频率直方图。频率直方图能大致地描述出总体 X 的概率分布情况，每个长方形面积正好近似地代表了 X 的取值落入相应一组的概率。结合连续型随机变量密度函数的直观意义，可以看出，只要有了频率直方图，就可以大致画出概率密度函数曲线。因而可以通过增加观测数据，把频率直方图作为概率密度函数的一种近似。但是，它只适用于连续型随机变量。累积频率曲线所代表的函数 $F_n(x)$ ，无论对于连续型或离散型随机变量都可以用，它是总体分布函数 $F(x)$ 的良好近似。将 $(x_1, x_2, \dots, x_n)$ 按大小排列为： $x_1^* \leq x_2^* \leq \dots \leq x_n^*$ ，

令

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_1^* \\ \frac{1}{n}, & \text{当 } x_1^* \leq x < x_2^* \\ \vdots & \\ \frac{k}{n}, & \text{当 } x_k^* \leq x < x_{k+1}^* \\ \vdots & \\ 1, & \text{当 } x \leq x_n^* \end{cases} \quad (4-20)$$



$F_n(x)$ 的图形就是累积频率直方图。它是跳跃上升的一条阶梯型曲线。若观测值不重复，则每一跃度为 $1/n$ ；若有重复情形，则按 $1/n$ 的倍数跳跃上长。对于任何实数 x ， $F_n(x)$ 等于样本的 n 个观测值中不超过 x 的个数除以样本容量 n 。由频率与概率的关系知道， $F_n(x)$ 可以作为未知分布函数 $F(x)$ 的一个近似， n 越大，近似得越好。我们称它为样本分布函数或经验分布函数。

二、样本数字特征

样本的数字特征,是显示一个样本分布某些特征的数字。人们经常用它们来估计总体相应的数字特征。常用的样本数字特征有:样本平均数、样本方差、样本协方差、样本相关系数和其它样本统计量等。

(一) 样本平均数

对于样本 $(x_1, x_2, \dots, x_n)$, 称

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4-21)$$

为样本的平均数。若样本 $(x_1, x_2, \dots, x_n)$ 指 n 元随机变量, 则样本平均数为 n 元随机变量的函数; 若样本 $(x_1, x_2, \dots, x_n)$ 指一具体观测值, 则样本平均数为一具体值。

(二) 样本方差

对于样本 $(x_1, x_2, \dots, x_n)$, 称

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4-22)$$

为样本的方差。称样本方差的平方根

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4-23)$$

为样本标准差。

我们注意到, 样本方差的分子也是 n 项离差之和, 为什么分母只除以 $n-1$, 而不是除以 n ? 因为这里定义的样本方差主要用于估计总体方差。当估计总体均值时, 它是第一次估计, 我们可以在总体中任意取 n 个变量值来计算样本均值, 这时选择变量值的自由度和样本容量是相同的, 刚好有 n 个自由度。但方差估计却是在样本均值估计基础上的第二次估计, 当我们确定了样本均值之后, 就不再有 n 个自由度, 最多只有 $n-1$ 个。所以样本方差应该以 $n-1$ 为分母。

(三) 样本协方差

对于样本 $(x_1, x_2, \dots, x_n)$, 称

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (4-24)$$

为样本的协方差。

(四) 样本相关系数

对于样本 $(x_1, x_2, \dots, x_n)$, 称

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4-25)$$

为样本的相关系数。

(五) 样本统计量(Statistics)

样本统计量是指样本数字特征, 它定义为不含未知数的样本随机变量的函数。一个样本可以构造出许多统计量, 如上面的样本平均数、样本方差等等。根据统计推断的需要而构造不同的统计量。而且统计量的观测值是建立在随机抽样的基础上, 随着抽到的样本单位不同, 其观察值也会有变化, 统计量的取值也随之变化, 所以统计量本身也是随机变量。从同一总体中抽出样本容量相同的所有可能样本后, 计算每个样本统计量的取值和相应的概率, 就组成样本统计量的概率分布, 简称抽样分布。

第四节 抽样分布——总体与样本的连接点

抽样分布就是样本统计量的分布,是总体与样本的连接点。本节首先讨论几种重要分布,然后讨论总体和样本是如何通过抽样分布被联系在一起的。

一、几种重要分布

(一) 正态分布

在统计推断中正态分布居于特别重要的地位,它是连续型随机变量的分布,其密度函数为:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4-26)$$

式中 μ 为正态分布的均值, $\sigma > 0$ 是它的标准差。这两个参数就唯一决定了正态分布密度函数的形状。所以正态分布可以简记为 $N(\mu, \sigma^2)$, 其图形如图 4-45 所示。

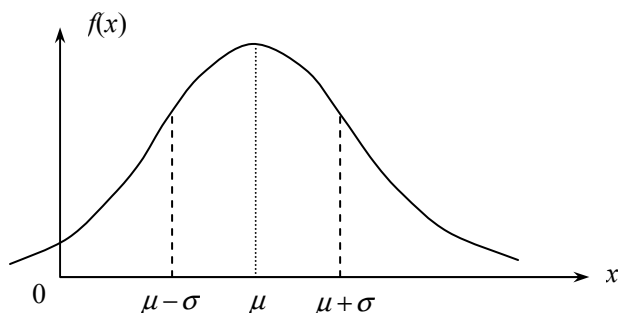


图 4-45 正态分布密度函数曲线

正态分布密度函数有如下性质:

- (1) 对称性。即以 $x = \mu$ 为对称轴, 曲线完全对称地向两边延伸。
- (2) 非负性。密度函数 $f(x)$ 都处于 OX 轴的上方。
- (3) 当 $x = \mu$ 时 $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$ 为最大值。 $f(x)$ 的值随 $|x|$ 递增而递减。

变动均值 μ 而 σ 不变, 则并不改变正态分布的形状, 而只改变正态分布的中心位置, 如图 4-46 所示。

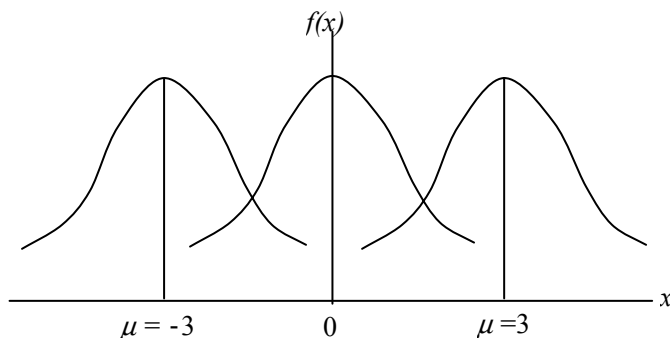


图 4-46 μ 变而 σ 不变

参见黄良文:《社会经济统计学原理》,中国统计出版社,1996年。

(4)在 $\mu \pm \sigma$ 处为密度函数 $f(x)$ 的拐点，即在 $\mu - \sigma < x < \mu + \sigma$ 的区间里，曲线凸向上，此外曲线凹向下。如图 4-45 所示。

变动标准差 σ 而 μ 不变，则并不改变正态分布的中心位置，而只改变分布曲线的尖锐程度，如。当 σ 变小时，密度函数曲线的中心部分纵坐标升高，曲线两侧迅速趋于 μ ，表示变量分布比较集中。反之，当 σ 变大时，则曲线呈现扁平，表示变量分布比较分散。

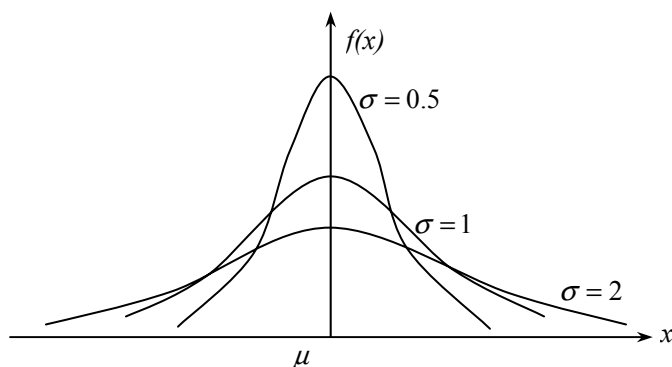


图 4-47 σ 变而 μ 不变

(5)当 $x \rightarrow \pm\infty$ 时，密度函数 $f(x) \rightarrow 0$ ，即曲线向两边下垂，伸向无穷远处。

根据正态分布的密度函数，可以推导出正态分布的分布函数 $F(x)$ 为：

$$F(x) = \int_{-\infty}^x f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-(x-\mu)^2 / 2\sigma^2} dx \quad (4-27)$$

可以证明作为分布函数的两个基本性质：

(1 对于任何 x ，有

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \geq 0$$

$$(2) \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2 / 2\sigma^2} dx = 1$$

将正态分布的密度函数和分布函数的图形对比见图 4-48 和图 4-49。

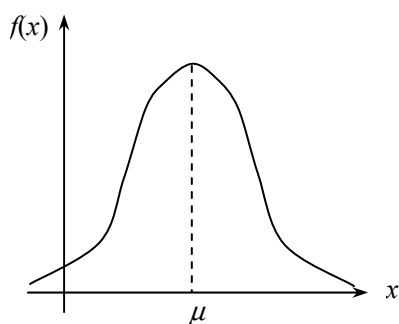


图 4-48 正态分布密度函数曲线图

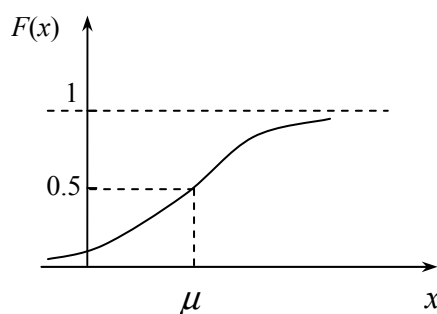


图 4-49 正态分布分布函数曲线图

利用正态分布函数可以计算 x 落在区间 $(\mu - a, \mu + a)$ 之间的概率，即：

$$\begin{aligned} P(\mu - a \leq x < \mu + a) &= P(|x - \mu| \leq a) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-a}^{\mu+a} e^{-(x-\mu)^2 / 2\sigma^2} dx \quad (4-28) \end{aligned}$$

不同现象的随机变量就有不同的均值和方差，不同的正态分布参数也就有不同的正态分布形式，要利用上述分布函数 $F(x)$ 对各类不同的正态分布求某点或某区间的概率是很困难的。为此我们需要对各种正态分布加以标准化，使不同的正态分布变换为具有相同参数的标准正态分布。标准正态分布要求：第一，分布的均值为 0；第二，分布的标准差为 1。现在我们对随机变量 X 作下列变换使新的随机变量 Z 等于：

$$Z = \frac{X - \mu}{\sigma} \quad (4-29)$$

$$\text{则 } E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0$$

$$\begin{aligned} \text{Var}(Z) &= \sigma_z^2 = E\left[\frac{X - \mu}{\sigma} - E\left(\frac{X - \mu}{\sigma}\right)\right]^2 = E\left(\frac{X - \mu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2} E(X - \mu)^2 = \frac{\sigma^2}{\sigma^2} = 1 \end{aligned}$$

所以，标准正态分布函数 $F(Z)$ 为：

$$F(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2} \quad (4-30)$$

标准正态分布函数 $F(z)$ 为：

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-Z^2/2} dZ \quad (4-31)$$

并简记为 $N(0,1)$ 。

标准正态分布的几何意义是将分布曲线的中心移到原点，使 $\mu = 0$ ，并对 $x - \mu$ 的离差化为以 σ 为单位的相对离差，即 σ 作为新变量 Z 的计量单位。将标准正态密度函数和分布函数图形比较如图 4-50 和图 4-51 所示。

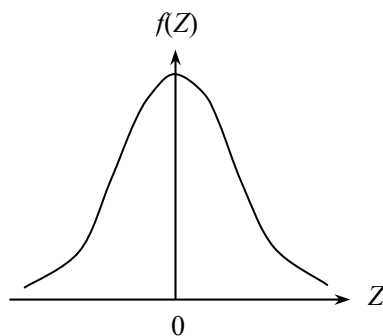


图 4-50 标准正态分布的密度函数图

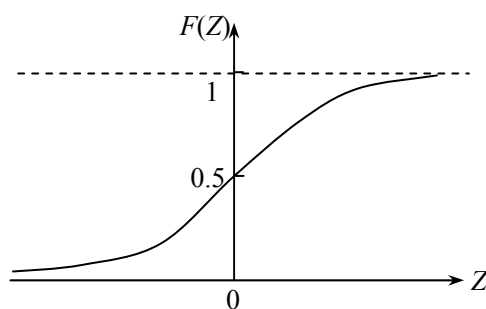


图 4-51 标准正态分布的分布函数图

在统计推断中，常常需要解释变量离中心 $\pm z$ 间的概率，即变量落在 $(-z, z)$ 区间的概率，并且考虑到正态分布的对称性，则所要求的概率积分可以给出如下形式：

$$P(-z < Z < z) = P(|Z| < z) = \frac{2}{\sqrt{2\pi}} \int_{-z}^z e^{-Z^2/2} dZ \quad (4-32)$$

这就是标准正态分布概率积分的标准式。由此可知，给定 z 值就有相应的 $P(|Z| < z)$ 。为了应用上的方便，把 Z 从 0-3 相应的概率编成正态分布概率表，列于本书的附录中，实际工作中可以直接套用，不必计算概率积分。

如果所研究的随机变量服从于标准正态分布 $N(0,1)$ ，则可以直接查用概率表，从给定的 z 值查所需的概率，或从给定的概率反查相应的 z 值。

1、求 z 距中心的绝对值不超过 a 的概率，如图 4-52 所示的阴影部分。就可以从概率表中查出当 $z=a$ 时，对应的概率值。例如：

当 $z=0.5$, $P(|Z| < 0.5) = 0.3829$

$z=1.0$, $P(|Z| < 1.0) = 0.6827$

$z=2.0$, $P(|Z| < 2.0) = 0.9545$ 等。

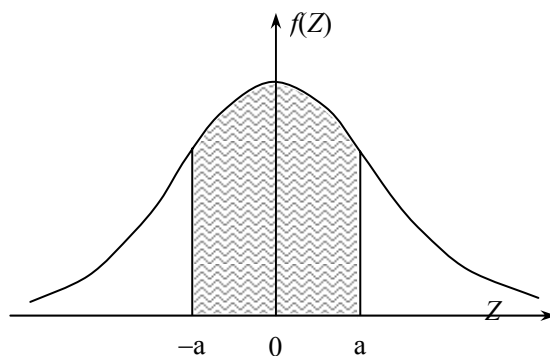


图 4-52

2、给定 $P(|Z| < z)$ ，求 z 距中心的绝对值 a 。例如

给定 $P(|Z| < z) = 0.1585, z = 0.2$

$P(|Z| < z) = 0.8030, z = 1.2$

$P(|Z| < z) = 0.9973, z = 3.0$ 等。

如果所研究的随机变量服从于一般正态分布 $N(\mu, \sigma^2)$ ，要估计变量 X 与平均数 μ 的离差绝对值不大于某数 a 的概率，或变量 x 落于 $(\mu - a, \mu + a)$ 区间的概率。根据正态分布标准化的要求，第一步将 X 变换为新变量 Z ，使 $Z = \frac{X - \mu}{\sigma}$ ；第二步将区间 $(\mu - a, \mu + a)$ 相应变换为 $(-\frac{a}{\sigma}, \frac{a}{\sigma})$ 即 $(-z, z)$ ，然后根据标准正态分布函数计算新区间的概率。

[例 4-2] 某农场的小麦产量服从正态分布，已知平均亩产为 550 公斤，标准差为 50 公斤，求亩产在 525—575 公斤间所占的比例。

根据正态分布标准化的要求，令 $X = \frac{X - \mu}{\sigma} = \frac{X - 550}{50}$ ，按题意要求 X 落在 $(\mu - a, \mu + a)$ 区间的

的概率，这里 $a = 25$ 公斤，所以新变量 z 的区间相应为 $(-\frac{a}{\sigma}, \frac{a}{\sigma}) = (-0.5, 0.5)$ 。当 $t = 0.5$ ，查概率表得：

$$\begin{aligned} P(525 \leq x < 575) &= P(|x - 550| < 25) \\ &= 0.3829 \end{aligned}$$

即约有 38.29% 的亩产量在 525—575 公斤之间。

[例 4-3] 解放军战士的身高是按正态分布的，经抽查平均身高 175 公分，标准差 4 公分，现在军服厂要裁制 100000 套军服，问身高在 171—179 公分之间应载几套？

根据正态分布标准化的要求 $Z = \frac{|X - \mu|}{\sigma} < \frac{4}{4} = 1$ ，查概率表则有：

$$P(171 \leq x < 179) = P(|x - 175| < 4) = 0.6827$$

即身高在 171—179 公分之间需载制 $100000 \times 0.6827 = 68270$ 套。

(二) 卡方(χ^2)分布

设 $x_1, x_2, \dots, x_n$ 相互独立, 且都服从标准正态分布, 即 $x_i \sim N(0,1)$ ($i=1,2,\dots,n$), 则

$$\chi^2 = x_1^2 + x_2^2 + \dots + x_n^2 = \sum_{i=1}^n x_i^2 \quad (4-33)$$

服从具有 n 个自由度的卡方分布, 记为 $\chi^2 \sim \chi^2(n)$ 。其密度函数 (具体函数式略) 的图形如图 4-53 所示。

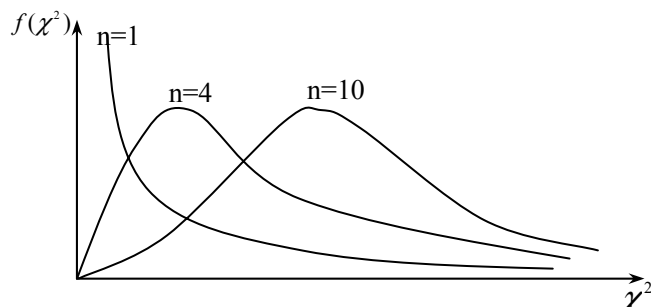


图 4-53 卡方分布密度函数

其中的 n 称为自由度。所谓自由度, 是指自由变量的个数。

(三) t (student) 分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则随机变量

$$t = \frac{X}{\sqrt{Y/n}} \quad (4-34)$$

服从自由度为 n 的 t 分布, 记为 $t \sim t(n)$ 。其密度函数 (具体函数式略) 的图形如图 4-54 所示。

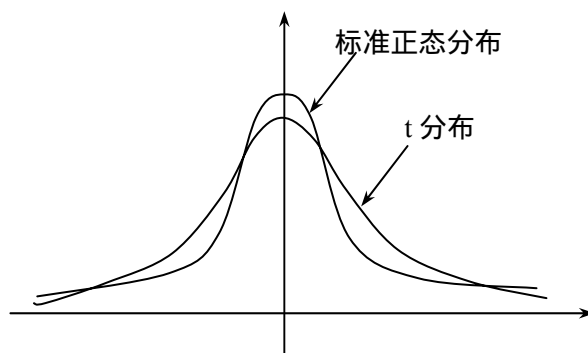


图 4-54 t 分布密度函数

(四) F 分布

设 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且 X 与 Y 相互独立, 则随机变量

$$F = \frac{X/n_1}{Y/n_2} \quad (4-35)$$

服从第一自由度为 n_1 , 第二自由度为 n_2 的 F 分布, 记为 $F \sim F(n_1, n_2)$ 。其密度函数 (具体函数式略) 的图形如图 4-55 所示。

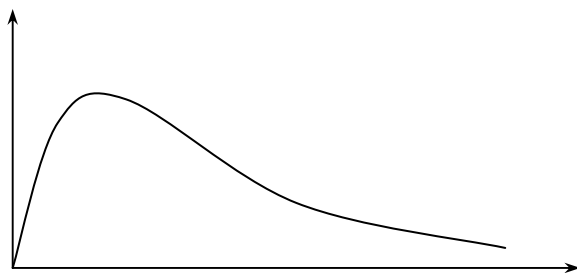


图 4-55 F 分布密度函数

二、抽样分布——总体与样本的连接点

(一) 样本均值 $\bar{x}$ 的抽样分布

样本均值的分布是由总体中全部样本均值的可能取值和与之相应的概率组成。下面举例说明。

例如，某班组 A、B、C、D、E 五人的日工资分别为 34、38、42、46、50 元，则总体工人日平均工资

$$\mu = \bar{X} = \frac{\sum x}{N} = \frac{34+38+42+46+50}{5} = 42 \text{元}$$

总体日工资方差

$$\sigma_x^2 = \text{Var}(X) = \frac{(34-42)^2 + (38-42)^2 + (46-42)^2 + (50-42)^2}{5} = 32.2 \text{元}$$

现在用重置抽样（即有放回抽样）的方法从五人中间随机抽 2 个构成样本，并求样本平均工资来推断总体的平均工资水平。由于是重置抽样，所以第 1 个单位是从总的 5 种工资中取种，第 2 个单位也是从同一总体的 5 种中取一种，共有 25 个样本，各样本的日平均工资可以列表如下：

表 4-16 样本日工资平均数（单位：元）

样本变量	34	38	42	46	50
34	34	36	38	40	42
38	36	38	40	42	44
42	38	40	42	44	46
46	40	42	44	46	48
50	42	44	46	48	50

从上表容易看出样本的平均数及其次数，可以整理列出样本平均数据的分布表以及图示如下：

表 4-17 样本日平均工资分布

样本日平均工资（元）	频数	频率
34	1	1/25
36	2	2/25
38	3	3/25

参见黄良文：《社会经济统计学原理》，中国统计出版社，1996 年。

40	4	4/25
42	5	5/25
44	4	4/25
46	3	3/25
48	2	2/25
50	1	1/25
合 计	25	1

根据以上资料，可以计算样本日工资平均数的平均数 $E(\bar{x})$ 和样本日工资平均数的方差 $Var(\bar{x})$ 。

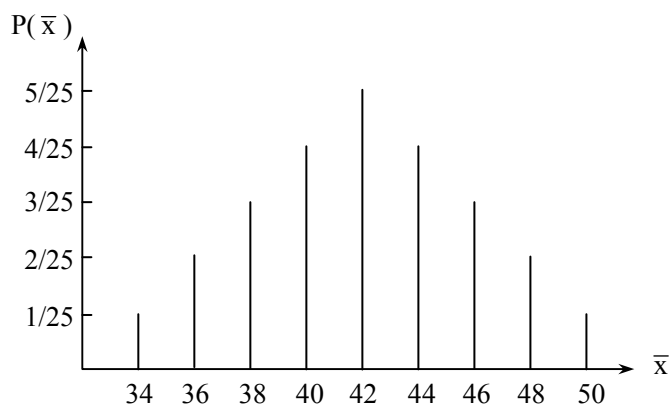


图 4 - 56 样本日平均工资分布图

$$E(\bar{x}) = \frac{\sum \bar{x}f}{\sum f} = \frac{1}{25} (34 \times 1 + 36 \times 2 + 38 \times 3 + 40 \times 4 + 42 \times 5 +$$

$$44 \times 4 + 46 \times 3 + 48 \times 2 + 50 \times 1)$$

$$= 16 \text{ 元}$$

$$Var(\bar{x}) = \frac{\sum [\bar{x} - E(\bar{x})]^2 f}{\sum f}$$

$$= \frac{1}{25} [(34 - 42)^2 + (36 - 42)^2 \times 2 + (38 - 42)^2 \times 3 + (40 - 42)^2 \times 4 + (44 - 42)^2 \times 4 + (46 - 42)^2 \times 3 + (48 - 42)^2 \times 2 + (50 - 42)^2]$$

$$= 16 \text{ 元}^2$$

$$\sigma_{\bar{x}} = \sqrt{Var(\bar{x})} = \sqrt{16} = 4 \text{ 元}$$

从以上计算，可以得到两个重要的结论：

(1) 重置抽样的样本均值 $\bar{x}$ 的平均数等于总体均值，即：

$$E(\bar{x}) = \bar{X} = \mu$$

上例两者都等于 42 元。这说明虽然每个样本平均数的取值可能与总体平均数有一定离差，但从总体来看，所有样本均值说来和总体平均数是没有离差的。

(2) 抽样平均数的标准差 $\sigma_{\bar{x}}$ 反映样本平均数与总体平均数的平均误差程度，这是因为：

$$\sqrt{E[\bar{x} - E(\bar{x})]^2} = \sqrt{E(\bar{x} - \mu)^2}$$

所以，称之为抽样平均误差，或抽样标准误差(SE: Standard Error)，以 SE 表示。重置抽样的抽样平均误差等于总体标准差除以样本单位数的平方根。即：

$$SE = \sigma_{\bar{x}} = \frac{\sigma_{\bar{x}}}{\sqrt{n}}$$

在本例中，直接以总体标准差 $\sigma_{\bar{x}}$ 和样本单位数 n 代入上式得：

$$SE = \sigma_{\bar{x}} = \sqrt{\frac{\sigma_{\bar{x}}^2}{n}} = \sqrt{\frac{32}{2}} = 4 \text{元}$$

所得结果和上面计算的结果完全一致。它表明所有样本日平均工资和总体日平均工资的平均离差为 4 元。

可以看出：首先，抽样标准误差比总体标准差小得多，仅为总体标准差的 $\frac{1}{\sqrt{n}}$ 。例如一个县的粮食亩产高低悬殊，亩产标准差 σ 为 80 公斤，如果随机取 100 亩求平均亩产，那么样本平均亩产量的差异就显著缩小，标准误差只是总体亩产标准差的 $\frac{1}{\sqrt{n}} = \frac{1}{10}$ ，即

$SE = \frac{80}{\sqrt{100}} = 8$ 斤。所以用样本平均亩产来代表总体平均亩产是更有效的。其次，抽样平均误差和总体标准差成正比变化，而和样本容量 n 的平方根成反比变化。例如在同一总体中，如果抽样单位数即样本容量扩大为原来的 4 倍，则抽样标准误差就缩小一半，如果抽样平均误差增加一倍，则样本单位数只需原来的 1/4 等等。

一般地，设 $x_1, x_2, \dots, x_n$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的随机样本，则样本均值

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \sim N(\mu, \frac{\sigma^2}{n}) \quad (4-36)$$

即 $E(\bar{x}) = \mu$, $Var(\bar{x}) = \frac{\sigma^2}{n}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ 。那么样本均值的标准化变量

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)。 \quad (4-37)$$

以下是关于正态分布的两个定理：

1、正态分布再生定理。如果变量 X 服从于其总体平均数为 μ 、总体标准差为 σ 的正态分布，即总体变量 X 服从正态分布 $N(\mu, \sigma^2)$ ，则从这个总体中抽取容量为 n 的样本平均数 $\bar{x}$ 也服从于正态分布，其平均数 $E(\bar{x})$ 仍为 μ ，其标准差 $\sigma_{\bar{x}} = SE$ ，即样本平均数 $\bar{x}$ 服从于正态分布 $N(\mu, SE^2)$ 。而标准随机变量 $z = \frac{\bar{x} - \mu}{SE}$ 则服从于标准正态分布 $N(0, 1)$ 。

这条定理表明，只要总体分布是正态的，则不管样本单位数 n 是多少，样本平均数都服从正态分布，分布的中心不变，而标准差即抽样标准误差则视重置抽样为 $\frac{\sigma}{\sqrt{n}}$ ，它们比总体

标准差都大大缩小了,因而样本平均数的分布是更加集中于总体平均数周围。

2、中心极限定理。如果变量 X 的分布具有有限的平均数 μ 和标准差 σ ,则从这个总体所抽取的容量为 n 的样本,样本平均数 $\bar{x}$ 的分布随着 n 的增大而趋近于平均数 μ 、标准差为 $\sigma_{\bar{x}} = SE$ 的正态分布,即样本平均数 $\bar{x}$ 趋近于正态分布 $N(\mu, SE^2)$ 。而样本变量 $z = \frac{\bar{x} - \mu}{SE}$ 则趋近于标准正态分布 $N(0,1)$ 。

这条定理并不要求总体分布是正态的,甚至可以是不知道的。客观上存在着总体平均数和标准差,只要样本的单位增多,则样本平均数 $\bar{x}$ 就越趋近于正态分布。这和正态分布再生定理限制总体为正态,而对样本单位数不加限制的情况是不同的。这条定理是中心极限定理的推广。

在实际工作中,总体变量的分布通常是不知道,样本平均数分布是否接近于正态,或接近到什么程度,起决定作用的因素是样本容量 n 。样本容量 n 越大,样本平均数的分布也越接近正态。一般认为样本单位数不少于 30 的是大样本,抽样分布就接近于正态分布。

(二) 样本方差 S^2 的抽样分布

设 $x_1, x_2, \dots, x_n$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的随机样本,则样本方差

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

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1) \quad (4-38)$$

(三) t 统计量的分布

设 $x_1, x_2, \dots, x_n$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的随机样本, $\bar{x}$ 为样本均值, S 为样本标准差,则统计量

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t(n-1) \quad (4-39)$$

(四) F 统计量的分布

设 $x_1, x_2, \dots, x_{n_1}$ 和 $y_1, y_2, \dots, y_{n_2}$ 是来自正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 和正态总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的随机样本, S_1 、 S_2 分别为相应的样本标准差,则统计量

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1, n_2) \quad (4-40)$$

式 (4-36) (4-38) (4-39) 和 (4-40) 实际上就是总体与样本相联系的纽带。我们已经知道,所谓总体就是一个随机变量,所谓样本就是 n 个相互独立且与总体具有相同分布的随机变量,即一个 n 元随机变量 $(x_1, x_2, \dots, x_n)$ 。总体与样本之间的联系在于具有相同的分布。本节则使这一点具体化。从而为下一章进一步通过样本的特征来估计和代替总体的特征铺平道路。这里我们以式 (4-39) 为例进行分析。当我们研究某一个问题时,要利用某一个样本均值 $\bar{x} \sim N(\mu, \sigma^2/n)$ 这一条件,我们根据一般正态分布与标准正态分布之间的关系,首先把

$\bar{x}$ 标准化,即 $(\bar{x} - \mu) / (\sigma / \sqrt{n}) \sim N(0,1)$,然后查表或计算就可以得到我们所需要的信息。然而,

假设我们对样本所抽自的总体的分布了解得并不完备,只知道它属于某一个均值为 μ 为正态分布,而不知道其方差 σ^2 的具体值。在这种情况下,我们显然无法再利用 $(\bar{x}-\mu)/(\sigma/\sqrt{n}) \sim N(0,1)$ 这一关系式。自然地,我们想到了一种解决办法,即计算这个样本的方差 S^2 ,并用 S^2 代替未知的总体方差 σ^2 来估算 $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ 。这里,我们就是在用样本的数量特征 S^2 来代替总体的数量特征 σ^2 进行估计了,而这样对 $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ 估计所得的结果 $(\bar{x}-\mu)/(S/\sqrt{n})$ 是否仍服从标准正态分布呢?由 (4-39) 式可知, $\frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t(n-1)$

第五章 由样本推断总体

第一节 抽样

一、抽样的意义和原则

在日常生活中,做菜时想要知道咸淡如何、烧得烂不烂,只需取一勺尝尝就知道了,并不需要把整锅菜都吃完。这实际上就是抽样,用部分来代表总体。不过有一个前提,在品尝之前要将汤搅拌均匀。否则如果盐没有完全融化,这一勺菜就没有代表性了;而且品尝时,还要将锅里的肉、菜叶等都挑些尝尝,这样一小勺菜就可以代表一大锅菜了。

抽样就是从总体中抽取能代表总体的一部分,即样本,然后根据样本数据中所包含的信息按一定的逻辑推理对总体进行推断,样本数据是判断的基础。而样本数据的准确、有效、充分又要依赖于对抽样的科学组织形式。抽样组织形式,不仅关系到抽样组织工作的好坏优劣,甚至决定了对总体进行推断的成效,影响全局。因此,如何科学地设计抽样,保证随机条件的实验,并且取得最佳的抽样效果,便是一个至关重要的问题。

设计抽样时,首先要保证随机原则的实现。按随机原则抽样是推断的前提,失去这个前提,推断的理论和方法也就失去存在的意义。从理论上说,随机原则就是要保证总体每一单位(个体)都有同等的中选机会,或样本的抽选的概率是已知的。但在实践上,如何保证这个原则的实现,需要考虑许多问题。一是要有合适的抽样框。抽样框固然要具备可实施的条件,可以从中抽取样本单位。仅仅这样是很不够的,一个合适的抽样框必须考虑它是不是能覆盖总体的所有单位。例如,某城市进行民意调查,如果以该市的电话号码簿名单为抽样框显然是不合适的,因为并不是所有居民户都安装电话,而且安装电话的居民户又多数是经济条件较好的人物,从这里取得的样本资料是很难说具有全市的代表性。抽样框还要考虑抽样单位与总体单位的对应问题。在实践中发生不一致的问题也不是少见的。有的是多个抽样单位对应一个总体单位,例如调查学校学生家庭情况,以学生名单为抽样框,在学生名单中可能有两个或更多的学生属于同一家庭。也有是一个抽样单位对应几个总体单位,例如人口调查中以住户列表为抽样框,每一住户就包括许多人口。像这类抽样很可能造成总体单位中选机会不均等,应该注意加以调整。二是取样的实施问题。当总体中单位数很大甚至无限的情况下,要保证总体每单位中选的机会均等绝非是简单的工作。在设计中要考虑将总体各单位加以分类、排队或分阶段等措施,尽量保证随机原则的实现。

其次,要考虑样本容量和结构问题。样本的容量究竟要多大才算是适应的?例如在民意测验中,要调查多少人才能反映全国几亿人口的意见呢?调查单位多了会增加组织抽样的负担,甚至造成不必要的浪费;但调查单位太少又不能够有效地反映情况,直接影响着推断的效果。样本的容量取决于对抽样推断准确性、可靠性的要求,而后者又因所研究问题的性质和抽样结果的用途而不同,很难给出一个绝对的标准。但在抽样设计时应该重视研究现象的差异、误差的要求和样本容量之间的关系,作出适当的选择。对相同的样本容量,还有容量的结构问题,例如一个县要求抽取 500 亩播种面积,它可以是先抽 5 个村,然后每村抽 100 亩,也可以是先抽 10 个村,然后每村抽 50 亩等等,样本容量的结构不同,所产生的效果也不同。抽样设计应该善于评价而且有效利用由于调整样本结构而产生的效果。

再次,关于抽样的组织形式问题。要认识到不同的抽样组织形式,会有不同的抽样误差,因而就有不同的效果。一种科学的组织形式往往有可能以更少的样本单位数,取得更好的抽

样效果。在抽样设计时必须充分利用已经掌握的辅助信息,对总体单位加以预处理,并采取合适的组织形式取样。例如粮食生产按地理条件分类,并分类取样。或按历史单产资料、当年估产资料,将各单位顺序排队,并等距取样等等,都能收到更好的抽样效果。还应该指出,即使是同一种抽样组织形式,由于采用的分类标志不同,群体的划分不同等等原因,仍然会产生不同的效果。因此应该认真细致地估计不同组织形式和不同抽样方法的抽样误差,并进行对比分析,从中选择有效和切实可行的抽样方案。

在抽样设计中还必须重视调查费用这个基本因素。实际上任何一项抽样调查都是在一定费用的限制条件下进行的,抽样设计应该力求调查费用节省的方案。调查费用可以分为可变费用和不变费用。可变费用随着调查单位的多少、远近、难易而变化,如搜集数据费、数据处理和制表费等等。不变费用是指不随工作量大小而变化的固定费用,如工作机关管理费、出版费等等。节约调查费用往往集中于可变费用的开支上。在设计方案中我们还要注意,提高精确度的要求和节省费用的要求并非一致,有时是相互矛盾的。抽样误差要求愈小,则调查费用往往需要愈大,因此并非抽样误差愈小的方案便是愈好的方案,许多情况是允许一定范围的误差就能够满足分析研究的要求。我们任务就在于在一定误差的要求选择费用最少的方案;或在一定的费用开支条件下,选择误差最小的方案。

二、抽样的组织形式

常用的抽样组织形式有简单随机抽样、分层抽样、等距抽样、整群抽样、阶段抽样、多阶段抽样等等。这里仅介绍简单随机抽样、分层抽样。

(一) 简单随机抽样

简单随机抽样(Simple Random Sampling)是按随机原则直接从总体 N 个单位中取 n 个单位作为样本。不论是重置抽样或不重置抽样,都要保证每个单位在抽选中有相等的中选机会。由于这种抽样组织形式对于总体除了抽样框的名单外,不需要利用任何其他信息,所以也称为单纯随机抽样。简单随机抽样是抽样中最基本也是最简单的方式,它适用于均匀总体,即具有某种特征的单位均匀地分布于总体的各个部分。在抽样之前要求对总体各单位加以编号,然后用抽签的方式或根据《随机数表》来抽选必要的单位数。未特别说明,本书所提到的抽样方法都是就简单随机抽样而言的。

组织抽样调查的一项重要工作是要确定合适的样本容量。在设计的时候,通常是先根据研究问题的性质确定允许的误差范围 Δ 和必要的概率保证程度(或概率度 t),然后根据历史资料或其他试点资料确定总体的标准差 σ ,通过抽样平均误差公式来推算必要的样本单位数 n 。

在重置抽样下,样本均值的误差公式:

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$\Delta_x = t \cdot SE = \frac{t\sigma}{\sqrt{n}}$$

所以必要的样本单位数 $n = \frac{t^2 \sigma^2}{\Delta_x^2}$

可以证明,在不重置抽样下,样本均值的误差公式:

$$SE = \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}$$

$$\Delta_x = t \cdot SE = \sqrt{\frac{t^2 \sigma^2}{n} \left(1 - \frac{n}{N}\right)}$$

$$\text{所以必要的样本单位数 } n = \frac{Nt^2\sigma^2}{N\Delta_x^2 + t^2\sigma^2}$$

从上式可以看出，必要的样本单位数受允许的误差范围 Δ 的制约， Δ 要求愈小则样本单位数 n 就需要愈多，但两者并不保持反比例的变化。以重置抽样来说在其他条件不变情况下，误差范围 Δ 缩小一半，则样本单位数必须增至四倍，而 Δ 扩大一倍，则样本单位数只需原来的 1/4。所以在抽样组织中对抽样误差可以允许范围要十分慎重地考虑。

简单随机的样在实践上受到许多限制，当总体很大时对每个单位编号、抽签等都会遇到难以克服的困难。但这种抽样方式在理论上说最符合随机原则，它的抽样误差容易得到数学上的论证，所以可以作为发展其他更复杂的抽样形式的基础，同时也是衡量其他抽样组织形式抽样效果的比较标准。

（二）分层抽样

分层抽样(Stratified Sampling)又称类型抽样，它先按一定标志对总体各单位进行分类，然后分别从每一类按随机原则抽取一定单位构成样本。分层抽样的前提是对总体事先有一定的认识，有辅助信息可资利用，这种信息和所研究的标志值大小有密切关系，可以作为分类的标志。通过分类把总体中标志值比较接近的单位归为一组，减少各组内部的差异程度，再从各组抽取样本单位就有更大的代表性，因而抽样误差也就相对减小了。在实际工作中广泛应用分层抽样方式。例如农产量抽样按地区分组，家计调查按国民经济部门分组，产品质量抽查按加工车床型号分组等等，都收到明显的效果。设总体由 N 个单位组成，把总体分为 k 组，使 $N=N_1+N_2+\dots+N_k$ ，然后从每组的 N_i 中取 n_i 单位构成总容量为 n 的样本，即 $n=n_1+n_2+\dots+n_k$ 。由于 k 组是根据一定标志划分的，各组单位数一般是不同的，怎样从 N_i 中取 n_i 呢？通常是按比例取样的，即按各组单位数占总体单位数的比例来分配各级应抽样本单位数，单位数较多的组应该多取样，单位数较少的组则少取样，保持各组样本单位数与各组单位数之比都等于总容量与总体单位数之比。即

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n}{N}$$

所以各组的样本单位数应为：

$$n_i = \frac{nN_i}{N}$$

采用按比例抽样这是因为保持结构和总体结构相同，避免样本均值由于各组比重差异而引起的误差。

分层抽样的样本均值计算是：首先，由各组分别取样，可以计算各组抽样均值：

$$x_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i=1,2,\dots,k)$$

然后将各组抽样均值 $\bar{x}_i$ 以各组单位数 N_i 或样本单位 n_i 为权数计算加权均值，即为所求的样本均值：

$$\bar{x} = \frac{\sum_{i=1}^k N_i \bar{x}_i}{N} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{n}$$

分层抽样的抽样平均误差可以这样考虑：由于分层抽样是对每一组抽样，所以不存在组间误差，抽样平均误差取决于各组内方差平均水平。首先计算各组内方差：

$$\sigma_i^2 = \frac{\sum (X_i - \bar{X}_i)^2}{N_i} \approx \frac{\sum (x_i - \bar{x}_i)^2}{n}$$

再以各组样本单位数 n_i 为权数，计算各组内方差的均值：

$$\overline{\sigma_i^2} = \frac{\sum n_i \sigma_i^2}{n}$$

样本均值的抽样平均误差可以按下列公式计算：

在重置抽样条件下：

$$SE = \sqrt{\frac{\overline{\sigma_i^2}}{n}}$$

在不重置抽样条件下：

$$SE = \sqrt{\frac{\overline{\sigma_i^2}}{n} \left(1 - \frac{n}{N}\right)}$$

例如某乡粮食播种面积 20000 亩，现在按平原和山区面积比例抽取其中 2%，计算各组平均亩产 $\bar{x}_i$ 和各组标准差 σ_i 如下表，求样本平均亩产 $\bar{x}$ 和抽样平均误差 μ_x 。

	全部面积 (亩) N_i	样本面积 (亩) n_i	样本平均亩产 (公斤) $\bar{X}_i$	亩产标准差 (公斤) σ_i
平原	14000	280	560	80
山区	6000	120	350	150
合计	20000	400	497	106

$$\bar{X} = \frac{\sum n_i \bar{x}_i}{n} = \frac{560 \times 280 + 350 \times 120}{400} = 497 \text{ 公斤}$$

$$\overline{\sigma_i^2} = \frac{\sum n_i \sigma_i^2}{n} = \frac{80^2 \times 280 + 150^2 \times 120}{400} = 11230 \text{ 公斤}$$

在重置抽样条件下：

$$SE = \sqrt{\frac{\overline{\sigma_i^2}}{n}} = \sqrt{\frac{11230}{400}} = 5.3 \text{ 公斤}$$

在不重置抽样条件下：

$$SE = \sqrt{\frac{\overline{\sigma_i^2}}{n} \left(1 - \frac{n}{N}\right)} = \sqrt{\frac{11230}{400} \left(1 - \frac{400}{2000}\right)} = 5.25 \text{ 公斤}$$

第二节 估计

一、总体参数估计概述

统计推断(Statistical inference)就是根据样本的实际数据,对总体的数量特征作出具有一定可靠程度的估计和判断。统计推断的基本内容有参数估计和假设检验两方面。概括地说,研究一个随机变量,推断它具有什么样的数量特征,按什么样的模式来变动,这属于估计理论的内容,而推测这些随机变量的数量特征和变动模式是否符合我们事先所作的假设,这属于检验理论的内容。参数估计和假设检验的共同点是它们都对总体无知或不很了解,都是利用部分观察值所提供的信息,对总体的数量特征作出估计和判断,但两者所要解决问题的着重点的所有方法有所不同。本节先研究总体参数估计的问题。

总体参数估计是以样本统计量(即样本数字特征)作为未知总体参数(即总体数字特征)的估计量,并通过对样本单位的实际观察取得样本数据,计算样本统计量的取值作为被估计参数的估计值。

不论社会经济活动还是科学试验,人们作出某种决策之前总是要对许多情况进行估计。例如商品推销人员要估计新式时装可能为消费者所喜欢的程度,自选商场经理要估计附近居民的购买能力,民意调查机构要估计竞选者的得票率,医药生产部门要推广某种药品的新配方,必须估计新药疗效的提高程度等等。这些估计通常是在信息不完全、结果不确定的情况下作出。参数估计为我们提供一套在满足一定精确度要求下根据部分信息来估计总体参数的真值,并作出同这个估计相适应的误差说明的科学方法。

科学的抽样估计方法要具备三个基本条件。

首先是要有合适的统计量作为估计量。我们知道统计量是样本随机变量的函数,根据样本随机变量可以构造许多统计量,但不是所有的统计量都能够充当良好的估计量。例如,从一个样本可以计算均值、中位数、众数等等,现在要用来估计总体均值,究竟以哪个样本统计量作为估计量更合适,如果采用样本均值作为估计量,这就需要回答样本均值和总体均值存在什么样的内在联系,以样本均值作为良好估计量的标准是什么等等。只有这些问题解决了,才能通过样本的实际观察确定估计值,而估计值是参数估计的基础。

其次,要有合理的允许误差范围。允许误差范围又称抽样极限误差,指样本统计量与被估计总体参数离差的绝对值可允许变动的上限或下限。离差的绝对值愈小表明抽样估计的准确度愈高,反之,就表明准确度愈差了。由于统计量本身也是随机变量,所以要使所做的估计完全没有误差是难以实现的,但估计误差也不能太大,估计误差如果超过了一定限度参数估计本身也就会失去价值。当然也不见得误差愈小就是愈好的估计,因为减少误差势必增加费用、时间,增加人力、物力、财力的负担,这样甚至会失去组织抽样调查的意义。所以在做估计的时候应该根据所研究对象的变异程度和分析任务的要求确定一个合理的允许误差范围,凡估计值与被估计值之间的离差不超过允许范围,这种估计都算是有效的。例如估计粮食亩产 600 公斤,允许误差范围 6 公斤,这意味着如果实际的粮食亩产在 594—606 公斤之间都应该认为估计是有效的。我们把允许误差的区间 594—606 公斤称为估计区间,允许误差与估计值之比称为误差率, $(1 - \text{误差率})$ 称为估计精度,上例误差率为 $6/600=1\%$,估计精度为 $1 - 1\%=99\%$ 。

再次,要有一个可接受的置信度。估计置信度又称估计推断的概率保证程度,这是估计的可靠性问题。由于抽样是随机抽样,统计量是随机变量,估计值所确定的估计区间也是随机的,在实际抽样中并不能做主被估计的参数真值都落在允许误差的范围内。这就产生要冒多大风险相信所作的估计。如果一种估计可信度很低,这就意味着所冒的风险很大,这种估计也就没有什么价值。例如我们愿意冒 10% 的风险,这表示如果进行多次重复估计,则平均每 100

次估计将 10 次是错误, 90 次估计正确。90% 就称为置信度或称概率保证程度。在抽样估计中要求达到 100% 的置信度是难以做到的, 但置信度小了, 估计结论的可靠性太低, 又会影响估计本身的价值, 所以在做估计的时候, 也应该根据所研究问题的性质和工作的需要确定一个可接受的估计置信度。当然估计置信度的要求和准确度的要求应该结合起来考虑, 估计的准确度很高而置信度很低或准确很低而置信度很高都是不合适的。

二、总体参数的点估计

点估计是直接以样本统计量作为相应总体参数的估计量。例如 $\bar{x} = \hat{\mu}$, 表示以样本均值 $\bar{x}$ 作为总体均值 μ 的估计量, 并根据实际抽样调查资料计算样本平均值作为总体均值参数的估计值。例如根据某地区样本资料计算粮食亩产 600 公斤, 就以这个数字作为全地区粮食亩产水平的估计值。

点估计的优点在于它能够提供总体参数的具体估计值, 可以作为行动决策的数量依据。例如推销部门对某产品估计出全所推销额数值, 并分出每月销售额, 便可传递给生产部门作为制订生产计划的依据, 而生产部门又可将各月产量计划传递给采购部门作为制订原材料采购计划的依据等等。点估计也有不足之处, 任何点估计不是对就是错, 并不能提供误差情况如何, 误差程度有多大的信息。

估计总体参数, 未必只能用一个统计量, 也可以用其他统计量。例如估计总体均值, 可以用样本均值, 也可以用样本中位数、众数等等。应当以哪一种统计量作为总体参数估计量才是最优的, 这就有评价统计量的优良估计标准问题。所谓优良估计总是从总体上来说的, 作为优良的估计量应该符合以下三个标准:

1、无偏性。即以样本统计量作为总体参数的估计量要求样本统计量的期望值 (均值) 等于被估计的总体参数。

用符号表示, 如果 θ 是被估计的参数, $\hat{\theta}$ 是有估计 θ 的样本统计量, 则当 $E(\hat{\theta}) = \theta$ 时, 就称

$\hat{\theta}$ 为 θ 的无偏估计量。就是说, 虽然每一次抽样, 所决定的统计量取值和总体参数的真值可能有误差, 误差可正可负, 可大可小, 但在多次反复的估计中, 所有样本统计量取值的均值应该等于总体参数本身。亦即说样本统计量的估计平均说来是没有偏误的。前已证明, 样本均值作为总体均值的估计量是符合无偏性要求的。即:

$$E(\bar{x}) = \mu$$

2、一致性。以样本统计量估计总体参数, 要求当样本的单位数充分大时, 样本统计量也充分靠近总体参数。一般地说, 如果样本容量 n 增大时, 估计量 $\hat{\theta}$ 更紧密地趋近于参数 θ ,

我们就称 $\hat{\theta}$ 为 θ 的一致估计量。就是说随着样本容量 n 的无限增加, 样本统计量和被估计的总体参数之差的绝对值小于任意小数, 它的可能性也趋近于必然性, 或者说这一事实几乎是肯定的。可以证明, 以样本均值估计总体均值, 也符合一致性的要求, 即存在一系列关系式:

$$\lim_{n \rightarrow \infty} P(|\bar{x} - \mu| < \varepsilon) = 1$$

式中, ε 为任意小数。

3、有效性。以样本统计量估计总体参数, 要求作为优良估计量的方差应该比其他估计量的方差小。一般地说, 如果 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计量 (对于给定的样本容量而言), 而

$\hat{\theta}_1$ 的方差 $\sigma^2(\hat{\theta}_1)$ 小于 $\hat{\theta}_2$ 的方差 $\sigma^2(\hat{\theta}_2)$, 我们可以说 θ_1 相对来说是更有效的估计量。

例如用样本均值或用总体任一变量来估计总体均值,虽然两者估计量都是无偏的,而且在每次估计中,两种估计值与总体均值都可能离差,但样本均值更集中在总体均值的周围,样本均值的方差只及总体变量方差的 $1/n$,就是说,平均说来样本均值的偏差更小,相对而言样本均值是更为有效的估计量。即

$$\text{Var}(\bar{x}) < \text{Var}(x)$$

不是所有估计量都符合以上标准。可以说符合以上标准的估计量要比不符合或不完全符合以上标准的估计量更为优良。例如在正态分布的情况下,总体均值和中位数是相重合的,样本均值是总体中位数的无偏估计量和一致估计量,而且样本均值比样本中位数作为总体中位数的估计量也是更有效的,因为样本均值的方差比样本中位数的方差更小。在下态分布的情况下,样本中位数是总体均值的无偏估计量和一致估计量。但对比样本均值却不是更有效的估计量,因为它的方差比样本均值的方差大,当然样本中位数也不是总体中位数的有效估计量。

三、总体参数的区间估计

总体参数的点估计事实上几乎不可能做到完全准确,更谈不上有多大的置信度。如果换一种思路,估计总体参数落在某一区间内,这就有把握多了。区间估计是根据给定的置信度要求,指出总体参数被估计的上限和下限。一般地说,对于总体被估计参数 θ ,找出样本的两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ (其中 $\hat{\theta}_2 > \hat{\theta}_1$) 使被估计参数落在区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 内的概率为 $1 - \alpha$, 其中 α 为介于 0—1 之间的已知数,即

$$P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha$$

称区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 为总体参数的估计区间, $\hat{\theta}_1$ 为估计下限, $\hat{\theta}_2$ 为估计上限, $1 - \alpha$ 为估计置信度, α 为显著性水平,如图 5-57 所示。

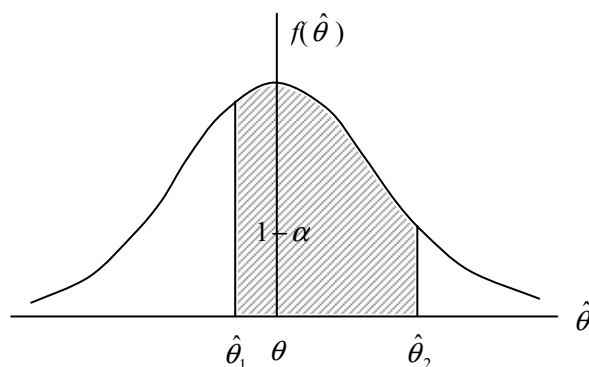


图 5-57

区间估计的特点是它不是指出被估计参数的确定数值,而是指出被估计参数的可能范围,同时对参数落在这一范围内给定相应的概率保证程度。正如上面已经指出的那样,参数的可能范围是估计的准确性问题,而相应的概率保证程度(置信度)是估计的可靠性问题。出于好意,在作估计时常常希望准确性尽可能提高,而且可靠性也不能小,但是这两个要求是矛盾的。在样本容量不变的条件下,要缩小估计区间,提高估计准确性,势必减少置信度,降低估计的可靠性。

[例 5-4] 根据第四章第三节的例子。从总体 5 个工人的日工资(总体平均日平均工资为 42 元、总体方差为 32 元²) 中用重置抽样的方法抽取样本容量为 2 人的样本平均工资的抽样分布如

下：

表 5 - 18

样本日 平均工 资 $\bar{x}$	34	36	38	40	42	44	46	48	50
频率 (概率)	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25

根据以上分布资料可以写出样本日平均工资落在各种区间的概率 P ，例如

$$P(40 \leq \bar{x} \leq 44) = (4/25) + (5/25) + (4/25) = 13/25$$

$$P(38 \leq \bar{x} \leq 46) = (3/25) + (4/25) + (5/25) + (4/25) + (3/25) = 19/25$$

$$P(34 \leq \bar{x} \leq 50) = (1/25) + (2/25) + (19/25) + (2/25) + (1/25) = 1$$

很容易将上述概率形式转换为样本均值与总体均值误差不超过一定范围的概率的形式即：

$$P(|\bar{x} - \bar{X}| \leq 2) = \frac{13}{25}$$

$$P(|\bar{x} - \bar{X}| \leq 4) = \frac{9}{25}$$

$$P(|\bar{x} - \bar{X}| \leq 8) = 1$$

这说明在重置抽样中，样本日平均工资与总体日平均工资绝对离差不超过 2 元的概率为 13/25，即有 52% 的概率保证总体日平均工资落在 40—44 元之间。同理，抽样误差不超过 4 元的概率为 9/25=36%，抽样误差不超过 8 元的概率为 100% 等等。由此可见，抽样误差范围和估计置信度是密不可分。估计置信度是抽样误差范围的函数，抽样误差范围愈小，估计准确度愈高，但置信度愈小。因此，在区间估计的时候，我们不可能对抽样误差范围和估计置信度都提出要求，只能根据给定的置信度（概率保证程度）来推算抽样误差范围的上下限，或根据给定的允许范围，来推算相应的置信度（概率保证程度）。

根据抽样分布定理，在样本单位数足够多 ($n \geq 30$) 的情况下，样本均值接近于正态分布，因而我们可以根据正态分布逼近的原理直接利用《正态分布概率表》查找确定所需要的概率或估计区间。但有一点必须说明，根据正态根据正态分布理论，抽样误差 $|\bar{x} - \bar{X}| = \Delta$ 的概率，

是指样本均值 $\bar{x}$ 落在 $(\bar{X} - \Delta, \bar{X} + \Delta)$ 区间的概率，即 $\bar{X} - \Delta \leq \bar{x} \leq \bar{X} + \Delta$ 的概率。然而实际上

总体均值 $\bar{X}$ 是未知的，而样本均值 $\bar{x}$ 在这里却已求知，也不需要再去估计，需要估计的是用已知的样本均值 $\bar{x}$ 去估计未知总体均值 $\bar{X}$ 所在的区间。我们所求的应该是总体均值 $\bar{X}$ 落在 $(\bar{x} - \Delta, \bar{x} + \Delta)$ 区间内的概率，即 $\bar{x} - \Delta \leq \bar{X} \leq \bar{x} + \Delta$ 的概率。那么我们可以不可以从概率

表所求的 $P(\bar{X} - \Delta \leq \bar{x} \leq \bar{X} + \Delta)$ 来代替 $P(\bar{x} - \Delta \leq \bar{X} \leq \bar{x} + \Delta)$ 呢？回答是肯定的。因为不等式 $\bar{X} - \Delta \leq \bar{x} \leq \bar{X} + \Delta$ 和 $\bar{x} - \Delta \leq \bar{X} \leq \bar{x} + \Delta$ 是等价的，所以 $P(\bar{X} - \Delta \leq \bar{x} \leq \bar{X} + \Delta) = P(\bar{x} - \Delta \leq \bar{X} \leq \bar{x} + \Delta)$ 。我们可以从概率表查询的 $P(\bar{X} - \Delta \leq \bar{x} \leq \bar{X} + \Delta)$ 概率来作为 $P(\bar{x} - \Delta \leq \bar{X} \leq \bar{x} + \Delta)$ 概率使用。

对于估计置信度 $1 - \alpha$ (例如 90%)，也可以这样来理解：虽然总体参数区间 $(\bar{X} - \Delta, \bar{X} + \Delta)$ 是固定的，而样本估计区间 $(\bar{x} - \Delta, \bar{x} + \Delta)$ 则是可变的，但如果反复抽样的结果将有 $1 - \alpha$ (即 90%) 的估计区间 $(\bar{x} - \Delta, \bar{x} + \Delta)$ 包含着总体参数 $\bar{X}$ 在内，如图 5 - 58 的(a)(b)，而其余 10% 的

估计区间不包含总体参数 $\bar{X}$ (见图 5 - 58)。因此在一次抽样估计中我们认为 $\bar{X}$ 落在 $(\bar{x} - \Delta, \bar{x} + \Delta)$ 区间的判断只有 $1 - \alpha$ (即 90%) 的可信度。

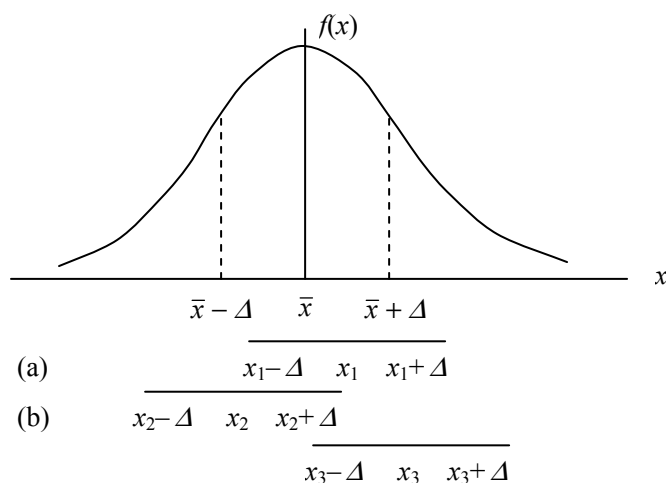


图 5 - 58

在进行区间估计时，需要先将允许极限误差 Δ 加以标准化，即以抽样平均误差 μ 除以极限误差 Δ 求得 z 值，表示以 μ 为单位的相对误差。

$$z = \frac{\Delta}{\mu} = \frac{\bar{x} - \bar{X}}{\mu}; \quad \Delta = z\mu$$

z 称为概率度。求 z 值过程也就是样本变量 $\bar{x}$ 的标准化过程。标准变量 z 服从标准正态分布， z 值大小是确定正态分布函数 $F(z)$ 的决定因子，由 z 值从正态分布表查到总体参数（总体均值）落在估计区间的概率。

例如经抽样调查计算样本亩产粮食 600 公斤，并求得抽样平均误差为 3 公斤，现在给定允许误差极限误差为 6 公斤，求总体平均亩产落在估计区间的概率。

已知： $\bar{x}$ = 600 公斤， μ = 3 公斤，并给定 Δ = 6 公斤

则估计区间是为 $(600 - 6, 600 + 6) = (594, 606)$

$$z = \frac{\Delta}{\mu} = \frac{\bar{x} - \bar{X}}{\mu} = \frac{6}{3} = 2$$

查正态概率表得，落在估计区间内的概率为

$$F(z) = F(2) = 95.45\%$$

这一结果表明，如果多次重复抽样，每组样本值都可以确定一个估计区间 (x_1, x_2) ，每个区间或者包含总体参数的真值，或者不包含总体参数真值，包含真值的样本区间占 $F(z)$ ，即每 1000 次抽样，就有 9545 个样本估计区间包括总体平均亩产，其余 55 个样本区间不包括总体均值，如果接受估计区间的判断也将要冒 4.55% 的机会犯错误的风险。

四、总体均值的估计

(一) 总体方差 σ^2 已知时，总体均值 μ 的估计

设总体 $X \sim N(\mu, \sigma^2)$ ，随机抽取一个容量为 n 的样本 $(x_1, x_2, \dots, x_n)$ 。计算样本均值

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ 作为总体均值的点估计。由样本均值的抽样分布}$$

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

有

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

对于给定的显著性水平 α , 令

$$P\left(-Z_{\frac{\alpha}{2}} \leq Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

即

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

从而

$$P\left(\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu < \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

即在给定 α 条件下, 总体均值在 $1 - \alpha$ 的置信水平下的置信区间为:

$$\left(\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

[例 5-5]某零件长度 X 服从均值为 μ m, 标准差为 0.15m 的正态分布, 现从中抽取 9 个零件, 测得其平均长度为 21.4。求在显著性水平为 0.05 时, 这种零件平均长度的置信区间。

解: 由题意, $X \sim N(\mu, 0.15^2)$, $\bar{x} = 21.4$, $n = 9$ 。查表得 $Z_{0.025} = 1.96$ 该零件平均长度的置信区间为:

$$\begin{aligned} & \left(\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) \\ &= \left(21.4 - 1.96 \cdot \frac{0.15}{\sqrt{9}}, 21.4 + 1.96 \cdot \frac{0.15}{\sqrt{9}}\right) \\ &= (21.302, 21.498) \end{aligned}$$

(二) 总体方差 σ^2 未知时, 总体均值 μ 的估计

总体方差 σ^2 未知时通常用样本方差 S^2 来估计。用 S^2 代替 σ^2 建立置信区间, 这时新的统计量

$$\frac{\bar{x} - \mu}{S / \sqrt{n}}$$

不再服从标准正态分布, 而是服从自由度为 $n-1$ 的 t 分布, 记为

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t(n-1)$$

因此, 对于给定的显著性水平 α , 令

$$P\left(-t_{\frac{\alpha}{2}} \leq t < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

即

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

从而

$$P\left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \mu < \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

即在给定 α 条件下，总体均值在 $1 - \alpha$ 的置信水平下的置信区间为：

$$\left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$$

例如，在[例 5-5]中，假设总体方差未知，通过样本求得样本方差为 0.17^2 ，那么零件平均长度的置信区间为：

$$\begin{aligned} & \left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) \\ &= \left(21.4 - 2.306 \frac{0.17}{\sqrt{9}}, 21.4 + 2.306 \frac{0.17}{\sqrt{9}}\right) \\ &= (21.269, 21.531) \end{aligned}$$

第三节 检验

一、总体参数假设检验概述

总体参数假设检验是利用样本的实际资料计算统计量的取值，并以引来检验事先对总体某些数量特征的假设是否可信作为决策取舍依据的一种统计分析方法。假设检验是统计推断的一项重要内容。

在现实生活中，由于我们通常难以完全知道所关心的总体的某些数量特征及其变化情况，因此对总体进行比较研究时，常常需要对目前总体的状况作出某种假设。例如工厂生产某种产品，经过工艺改革，使用新材料、新配方，企业管理者十分关心产品质量是否有所提高，因此可以假设经过改革以后产品质量可能提高或并没有提高。又如我们考虑目前股票市场上价格指数的走势是否正常，我们可以根据过去长期观察的平均水平和变异情况，作出当前股票价格水平可能正常或不正常的假设。

假设检验的基本思路是：首先对研究的命题提出一种假设，称为零假设或原假设，即从原来总体没有变化出发，假设所有措施都是无效的，这样就有一个总体参数，而且它的分布也是知道的，例如某轴承厂生产某型号轴承，按规定轴承标准承载压力 4000 公斤，标准差为 200 公斤，承载压力按正态分布。我们就可按 4000 公斤压力作为总体参数建立比较标准。现在从实际总体中抽取样本，并根据实际观察的资料计算统计量的取值。当然我们不知道总

体是否已经发生了变化,即不知道样本是来自新的总体,还是仍然从原来的总体抽取的,我们通过样本统计量取值与假设的总体参数比较来判断。要求两者完全一致的可能性是极少的,那么差异要达到多大才算是显著呢?所谓显著性是指差异程度而言的,程度不同说明引起差异的原因也有不同,存在着两种不同性质的差异,一种是条件差异,即由于工艺或试验条件的改变所引起的差异;一种是随机差异,即由于生产或试验过程中受偶然因素的影响,所引起结果的差异。这两种原因的共同作用导致各种各样的误差,如果样本统计量与假设总体参数之间的差异超过了通常偶然因素起作用的程度,它说明所发生的差异,除了随机因素之外还存在条件差异的因素,因此我们可以据此否定总体的变动纯粹由随机原因引起,没有显著差异的零假设。换句话说,如果我们能证明统计量和假设的总体参数实际发生的差异超过给定的标准的可能性很小,那么我们就有理由用反证法认为零假设是错误的,从而拒绝接受这个假设。否则,我们就没有理由拒绝零假设,而称零假设是可容的。

其次,确定显著性水平。如上所说,我们所以拒绝零假设,并不是因为它存在逻辑的绝对矛盾,或实际上不可能存在这种假设,而仅仅因为它存在的可能性很小。根据小概率事件原理,概率很小的事件在一次试验中几乎是不会发生的,如果根据零假设的条件正确计算出某一结果发生的概率很小,理应在一次试验中不至于发生,然而在一次试验中事实又发生了,则我们认为零假设不正确,而拒绝接受。

这里关键的问题是概率要小到如何程度才足以否定原来所作的假设。在进行假设检验时应该事先规定一个小概率的标准,作为判断的界限,这个小概率标准称为显著性水平。由于零假设的分布是已知的,因而样本统计量和总体参数的离差在一定范围内的概率也可以知道,离差超过这个范围的概率也同样知道,如统计量与参数的差异过大,以至发生这种事件的概率很小,而且小低于给的标准,我们就拒绝零假设,如果计算出的统计量与参数并异的相应概率大于给定标准,我们就接受零假设,这样,我们把概率分布分为两个区间:离差的绝对值大于给定的标准的概率分布区间称为拒绝区间,离差的绝对值小于这个标准则为接受区间。例如给定小概率标准 $\alpha=0.05$,凡概率小于5%的差异都是小概率事件,属于拒绝区间,如图中分布两端的阴影部分,而 $1-\alpha=0.95$,则是对立事件的概率,其概率在95%以内的,为接受区间,如图5-59中央部分所示。

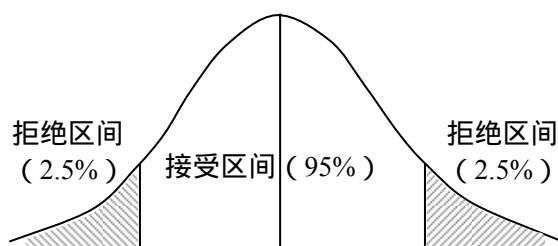


图 5 - 59

事件属于接受区间,零假设成立,判断总体无显著差异,事件属于拒绝区间,推翻零假设,认为总体有显著差异,其区间以小概率标准 $\alpha=0.05$ 为界限,所以称 α 为显著性水平, α 所对应的概率度称显著性水平 α 的临界值。例如 $\alpha=0.05$ 时,在正态分布的情况下,则临界值 $z_{0.05}=1.96$ 。我们以概率小于0.05的事件作为小概率事件,也就等于说大于临界值 $z_{0.05}=1.96$ 的事件作为小概率事件,这样我们可以直接利用概率表查找临界值作为判断的依据。

显著性水平主要视拒绝区间所可能承担的风险来决定,应该根据研究问题的性质和对结

论准确性的要求而有所不同。通常多采用 0.1、0.05、0.01、0.001 等显著性水平。例如民意测验采用显著性水平 $\alpha=0.1$ ，其他社会经济现象的检验取 $\alpha=0.05$ ，产品质量检验取 $\alpha=0.01$ ，工程技术检验取 $\alpha=0.001$ ，甚至取 $\alpha=0.0001$ 等等。取显著性水平愈大，则冒无显著性差异而被错判断为有显著性差异的风险也愈大。

现在我们可以把总体参数检验的步骤归纳如下：

(一) 提出假设。

首先提出零假设，记为 H_0 ，设立零假设的目的在于检验中要予以拒绝或接受的假设，零假设总是假定总体没有显著性差异，所有差异都是由随机原因引起的。所以这种假设又称无效假设。其次提出备择假设，记为 H_1 ，如果零假设被拒绝等于接受了备择假设，所以备择假设也就是所假设的对立事件。

(二) 决定检验的显著性水平 α 。

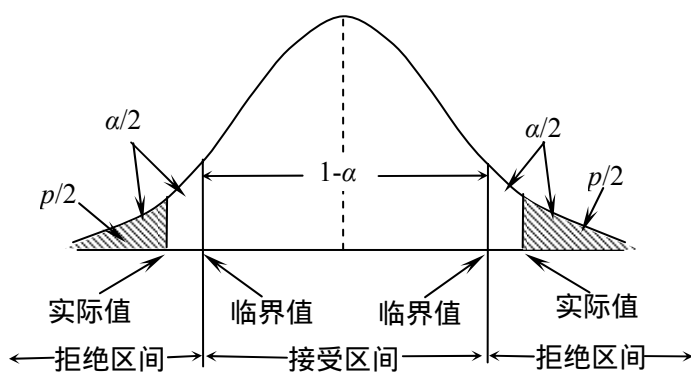
在零假设成立的条件下，由被检验的统计量分布求出相应的临界值，该临界值即为零假设的拒绝域和接受的分界线。

(三) 构造检验统计量，并依据样本信息计算检验统计量的实际值。

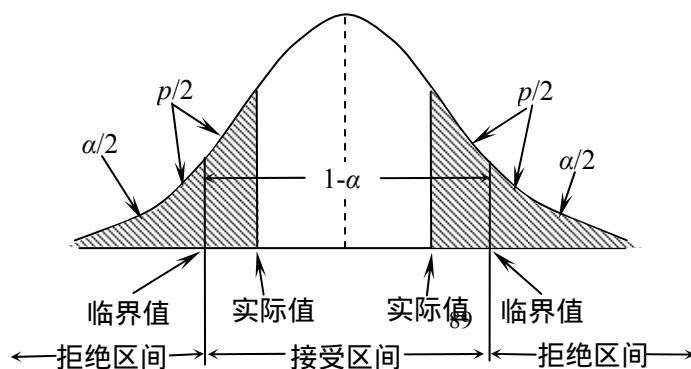
(四) 将实际求得的检验统计量取值与临界值进行比较，做出拒绝或接受零假设的决策。

如果检验统计量取值超过临界值，说明零假设落入拒绝域中，我们就选择拒绝接受零假设；若检验统计量的取值小于临界值，零假设落入接受域中，我们就不能拒绝零假设，而必须接受零假设或作进一步的检验。

SPSS 的输出结果中给出了相应检验统计量的实际取值，但由于显著性水平根据不同要求而有所不同，SPSS 并不给出临界值。如果不查概率表，就无法直接采用上面的步骤进行检验。SPSS 给出了检验统计量的概率值即文献中常见的 p 值 (p-value)，利用 p 值就可以直接进行检验。 p 值是在零假设成立的情况下，检验统计量的取值等于或超过检验统计量的实际值的概率，从而 p 值即为否定零假设的最低显著性水平。 p 值经常被称为实际显著性水平，以区别于给定的显著性水平 α 。



(A)



(B)

图 5-60

p 值可用式子表示为：

$$p = P(\text{检验统计量值} \geq |\text{检验统计量实际值}|)$$

当检验统计量的实际值超过临界值时（如图 5-60（A）所示），检验统计量的 p 值将小于给定的显著性水平 α ，零假设落入拒绝域中，我们就拒绝接受零假设；当检验统计量的取值小于临界值时（如图 5-60（B）所示），检验统计量的 p 值将大于给定的显著性水平 α ，零假设落入接受域中，我们就不能拒绝零假设，而必须接受零假设或作进一步的检验。

当 $p < \alpha$ 时，意味着如果给定一个真实的零假设，那么检验统计量的取值等于或超过实际观察到的极端值的概率为 α 。大多数学者都把这一结果解释为支持你否定零假设而接受替代假设的证据。有学者称 p 值为“实验使零假设相信者感到吃惊的程度的度量”。 p 值越小，零假设相信者吃惊的程度越高。

为了便于记忆，我们可以把 p 值理解为零假设的支持率或可信程度。 p 值越小，零假设越不可信。

利用 SPSS 输出的 p 值，我们可以直接用它来替代检验统计量实际值进行检验，而不必去查有关统计表并比较临界值了。在 SPSS 中进行总体参数检验的步骤如下：

- （一）提出零假设（ H_0 ）和备择假设（ H_1 ）。
- （二）给定检验的显著性水平 α 。
- （三）构造检验统计量，并依据样本信息由 SPSS 计算检验统计量的 p 值。
- （四）将实际求得的检验统计量的 p 值与给定的显著性水平 α 进行比较，做出拒绝或接受零假设的决策。

如果检验统计量的 p 值小于显著性水平 α ，说明零假设落入拒绝域中，我们就选择拒绝接受零假设；若检验统计量的 p 值大于显著性水平 α ，零假设落入接受域中，我们就不能拒绝零假设，而必须接受零假设或作进一步的检验。

二、双侧检验与单侧检验

根据我们所研究的问题性质不同，以及我们所关心的统计量与总体参数的显著性差异的方向不同，对统计检验方法的设计有双侧检验与单侧检验两种类型。

（一）双侧检验。当我们所关心的问题是检验样本均值与总体均值有没有显著性差异，而不问差异的方向是正差或负差，应该采用双侧检验。在双侧检验中，在零假设取等式，而备择假设取不等式，如：

$$H_0: \bar{X} = \bar{X}_0 \quad ; \quad H_1: \bar{X} \neq \bar{X}_0$$

同时，由于双侧检验不问差距的正负，所以给定的显著性水平 α ，须按正态对称分布的原理平均分配到左右两侧，每方各为 $\alpha/2$ ，相应得到下临界值为 $-Z_{\alpha/2}$ ，上临界值为 $Z_{\alpha/2}$ 。如图

5-61。

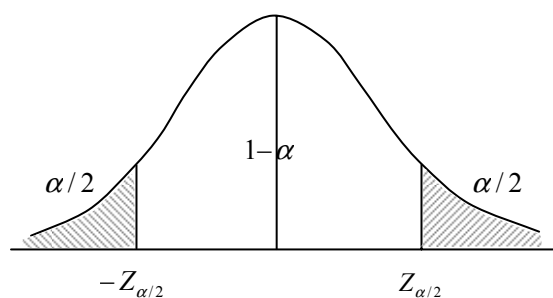


图 5 - 61 双侧检验

用检验统计量的实际值与临界值比较法进行检验：由样本信息计算的统计量 Z 实际值并与事先给定的临界值 $Z_{\alpha/2}$ 作比较。在双侧检验中，如果 $Z > Z_{\alpha/2}$ 或 $Z < -Z_{\alpha/2}$ ，就拒绝零假设 H_0 ，而接受备择假设 H_1 ；如果 $-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$ ，就不能否定零假设，而接受零假设是真实的。

用检验统计量的 p 值与显著性水平 α 比较法进行检验：由样本信息计算的统计量 p 值并与事先给定的显著性水平 α 作比较。如果 $p < \alpha$ ，就拒绝零假设 H_0 ，而接受备择假设 H_1 ；如果 $p \geq \alpha$ ，就不能否定零假设，而接受零假设是真实的。

（二）单侧检验。当我们所关心的问题不仅仅要检验样本均值和总体均值之间有没有显著的差异，而且追究是否发生预先指定方向的差异，应该采用单侧检验。而且根据关心的是正差异或负差异，单侧检验又有左单侧检验和右单侧检验。均值的单侧检验，零假设和备择假设都是以不等式的形式表示的。

1. 左单侧检验

当我们所关心的是总体均值是否低于预先假设，应该采用左单侧检验，零假设与备择假设为：

$$H_0: \bar{X} \geq \bar{X}_0 \quad ; \quad H_1: \bar{X} < \bar{X}_0$$

检验的显著性水平 α ，以及相应的临界值左侧临界值 Z_α ，如图 5 - 62 所示。

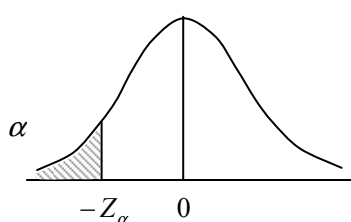


图 5 - 62 左单侧检验

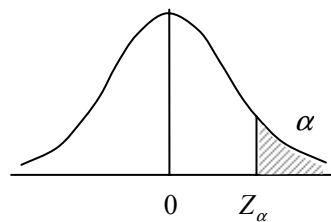


图 5 - 63 右单侧检验

用检验统计量的实际值与临界值比较法进行检验：将实际求得的 Z 值与事先给定的 Z_α 或 $-Z_\alpha$ 作比较，如果 Z 值等于或小于 $-Z_\alpha$ ，即 $Z \leq -Z_\alpha$ ，则拒绝零假设，如果 Z 值大于 $-Z_\alpha$ ，即 $Z > -Z_\alpha$ ，则接受零假设。

用检验统计量的 p 值与显著性水平 α 比较法进行检验 :由样本信息计算的统计量 p 值并与事先给定的显著性水平 α 作比较。如果 $p < \alpha$,就拒绝零假设 H_0 ,而接受备择假设 H_1 ;如果 $p \geq \alpha$,就不能否定零假设 ,而接受零假设是真实的。

2. 右单侧检验

当我们所关心的是总体均值是否高于预先假设 ,应该采用右单侧检验 ,零假设与备择假设为 :

$$H_0: \bar{X} \leq \bar{X}_0 \quad ; \quad H_1: \bar{X} > \bar{X}_0$$

检验的显著性水平 α , 以及相应的右临界值 Z_α , 如图 5 - 63所示。

用检验统计量的实际值与临界值比较法进行检验 :在右单侧检验中 , 如果 Z 值等于或大于 Z_α , 即 $Z \geq Z_\alpha$, 则拒绝零假设 , 接受备假设 , 如果 Z 值小于 Z_α , 即 $Z < Z_\alpha$, 则接受零假设。

用检验统计量的 p 值与显著性水平 α 比较法进行检验 :由样本信息计算的统计量 p 值并与事先给定的显著性水平 α 作比较。如果 $p < \alpha$,就拒绝零假设 H_0 ,而接受备择假设 H_1 ;如果 $p \geq \alpha$,就不能否定零假设 ,而接受零假设是真实的。

从双侧检验、左单侧检验和右单侧检验的检验方法中可以发现 , 不管是哪一种检验 , 用检验统计量的 p 值与显著性水平 α 进行比较的方法是统一的。如果用 SPSS 等统计软件进行检验 , 应采用 p 值与显著性水平比较进行检验 ; 如果通过查概率表进行检验 , 一般采用检验统计量实际值与临界值比较进行检验。

三、Z 检验与 t 检验

在假设检验中 , 由于样本容量和样本资料的限制 , 而使样本统计量有不同的概率分布 , 并据此形成 Z 检验和 t 检验两种方法。

(一) Z 检验

Z 检验又称正态分布检验。我们知道 , 从正态总体中随机抽取容量为 n 的样本 , 不论 n 的大小 , 样本均值 $\bar{x}$ 都服从正态分布 $N(\bar{X}, \mu^2)$, 而统计量 $Z = \frac{\bar{x} - \bar{X}}{\mu}$ 服从标准正态分布 $N(0,1)$ 。从一般非正态总体中抽取容量为 n 的样本 , 当容量 n 很大时 , 样本均值也趋近于正态分布 $N(\bar{X}, \mu^2)$, 而统计量 $Z = \frac{\bar{x} - \bar{X}}{\mu}$ 趋近于标准正态分布。这里抽样平均误差 $\mu = \frac{\sigma}{\sqrt{n}}$ 中 , σ 是已知的 , 因而 μ 也是确定的。但是常常总体的标准差 σ 不知道 , 必须用样本标准差

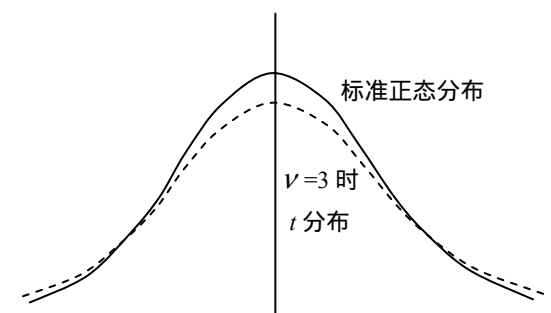
$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ 来代替 σ , 即用 $S(\bar{x}) = \frac{S}{\sqrt{n}}$ 来代替 μ 。这时统计量 $t = \frac{\bar{x} - \bar{X}}{S(\bar{x})}$ 不再是标准正

态统计量 $Z = \frac{\bar{x} - \bar{X}}{\mu}$ 了 , 因为 Z 中唯一变量 $\bar{x}$, 而在 t 中除了变量 $\bar{x}$ 外又加了另一变量 $S(\bar{x})$ 。

数学上已证明当样本容量足够大 ($n > 30$) 时 , 统计量 t 的分布趋近于正态分布。因此在大样本的情况下 , 我们可以利用正态分布来进行统计推断包括总体参数的估计和检验。这就是迄今为止我们都是用正态分布的统计量 Z 作区间估计和统计检验的原因。

(二) t 检验

在统计假设检验中，当总体的标准差 σ 未知，而需要用样本标准差 $S = \sqrt{\frac{(x - \bar{x})^2}{n-1}}$ 来代替时，则统计量 $t = \frac{\bar{x} - \bar{X}}{s/\sqrt{n}}$ 再不是服从标准正态分布，而服从于另一种概率分布，称为 t 分布。 t 分布是假定样本取自正态总体并且样本均值 $\bar{x}$ 和抽样标准差 $S(\bar{x})$ 相互独立的一种分布。 t 分布类似于标准正态分布，其期望值为 0， $E(t)=0$ ，并以它为中心形成钟型的两边对称分布。但标准正态分布的方差 $\sigma=1$ ，而 t 分布的方差 $\sigma^2(t)$ 则受自由度 $\gamma=n-1$ 这个参数的影响。当自由度很小，即小样本时， $\sigma^2(t)$ 大于 1，当自由度在 30 以上， t 分布和标准正态分布极为相近，以 S 估计 σ 的误差可以忽略不计，但当自由度很小时 t 分布的 S 变异就很明显，因此 t 分布和标准正态分布就有显著的差别。（图）是自由度为 3 时的 t 分布与标准正态分布的比较。 t 分布也是左右对称的，但 t 分布的顶部比标准正态分布低，而两端又比较高些。这个现象的直观解释是， t 分依赖于两个随机变量 $\bar{x}$ 和 $S(\bar{x})$ ，在小样本中， $\bar{x}$ 的极值和 S （因而 $S(\bar{x})$ ）的极值很可能会成对出现，所以统计量 t 势必比 Z 分散些。但是当自由度 ν 增大时， t 的变异性减小，当自由度无限增大，则 t 分布的方差趋近于 1。 t 分布与 Z 分布便重叠在一起。

图 5-64 正态分布与 t 分布

由此可见， t 分布受自由度 $\nu=n-1$ 大小的影响。一个自由度决定一个 t 分布，形成 t 分布族。在自由度 1—30 间可以按不同自由度编制 30 张 t 分布表。但在假设检验中，我们常用的显著性水平 α 只有 0.005、0.01、0.05、0.10 等几种，我们可以选用 t 分布中的部分概率，编制一张综合性的 t 分布表。

四、总体均值的假设检验

对来自正态总体的样本均值进行假设检验，又称为均值比较，通常采用 Z 检验法和 t 检验法。如果总体方差已知，则采用 Z 检验法；如果总体方差未知，则采用 t 检验法。对于多个总体的均值比较，请参见下一章的方差分析。

1、一个正态总体的均值比较

设总体 $X \sim N(\mu, \sigma^2)$ 的一个样本的均值为 $\bar{x}$ ，样本的方差为 S^2 ，样本容量为 n 。

当总体方差 σ^2 已知时，采用 Z 检验法检验总体均值 μ 是否等于某已知常数的步骤如表 5-19 所示。

表 5-19 正态总体方差已知时的均值比较 Z 检验表

步骤	双侧检验	右单侧检验	左单侧检验
1	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu \neq \mu_0$	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu > \mu_0$	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu < \mu_0$
2	构造检验统计量： $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ 若 H_0 成立, $Z \sim N(0,1)$	同左	同左
3	根据显著性水平 α , 查表确定临界值 $Z_{\frac{\alpha}{2}}$ 。 (或计算 Z 对应的 p 值)	根据显著性水平 α ,查 表确定临界值 Z_{α} 。 (或计算 Z 对应的 p 值)	同左
4	决策： 若 $ Z > Z_{\frac{\alpha}{2}}$ (或 $p < \alpha$) , 拒绝 H_0 ; 否则不否定 H_0 。	决策： 若 $Z > Z_{\alpha}$ (或 $p < \alpha$) , 拒 绝 H_0 ; 否则不否定 H_0 。	决策： 若 $Z < -Z_{\alpha}$ (或 $p < \alpha$) , 拒绝 H_0 ; 否则不否定 H_0 。

若总体的方差 σ^2 未知, 用样本标准差 S 代替 $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ 中的 σ , 这时该统计量不再服从正态分布, 而是服从自由度为 n-1 和 t 分布, 因此采用 t 检验法 (如表 5-20 所示)。

表 5-20 正态总体方差未知时的均值比较 t 检验表

步骤	双侧检验	右单侧检验	左单侧检验
1	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu \neq \mu_0$	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu > \mu_0$	提出假设： $H_0: \mu = \mu_0$ (μ_0 为常数) $H_1: \mu < \mu_0$
2	构造检验统计量： $t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$ 若 H_0 成立, $t \sim t(n-1)$	同左	同左
3	根据显著性水平 α 和 自由度 n-1 查表确定 临界值 $t_{\frac{\alpha}{2}}(n-1)$ 。(或 计算 t 对应的 p 值)	根据显著性水平 α 和 自由度 n-1 查表确定临 界值 $t_{\alpha}(n-1)$ 。(或计 算 t 对应的 p 值)	同左
4	决策：若 $ t > t_{\frac{\alpha}{2}}(n-1)$ (或 $p < \alpha$) 拒绝 H_0 ; 否则不否定	决策：若 $t > t_{\alpha}(n-1)$ (或 $p < \alpha$) , 拒绝 H_0 ; 否则不否定	决策：若 $t < -t_{\alpha}(n-1)$ (或 $p < \alpha$) 拒绝 H_0 ; 否则不否定

	拒绝 H_0 ; 否则不否定 H_0 。	H_0 。	H_0 。
--	--------------------------	---------	---------

2、两个正态总体的均值比较

我们还可以用 Z 检验法和 t 检验法对两个正态总体的均值进行比较。设两正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的两组相互独立样本 的均值分别为 $\bar{x}_1$ 和 $\bar{x}_2$, 样本方差分别为 S_1^2 和 S_2^2 , 样本容量分别为 n_1 和 n_2 。

当两个总体的方差 σ_1^2 和 σ_2^2 已知时 , 采用 Z 检验法检验两总体均值差是否等于某已知常数的步骤如表 5-21 所示。

表 5-21 两正态总体方差已知时的均值比较 Z 检验表

步骤	双侧检验	右单侧检验	左单侧检验
1	提出假设 : $H_0 : \mu_1 - \mu_2 = \Delta$ $H_1 : \mu_1 - \mu_2 \neq \Delta$ (Δ 为常数)	提出假设 : $H_0 : \mu_1 - \mu_2 = \Delta$ $H_1 : \mu_1 - \mu_2 > \Delta$ (Δ 为常数)	提出假设 : $H_0 : \mu_1 - \mu_2 = \Delta$ $H_1 : \mu_1 - \mu_2 < \Delta$ (Δ 为常数)
2	构造检验统计量 : $Z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ 若 H_0 成立, $Z \sim N(0,1)$	同左	同左
3	根据显著性水平 α , 查表确定临界值 $Z_{\frac{\alpha}{2}}$ 。 (或计算 Z 对应的 p 值)	根据显著性水平 α , 查 表确定临界值 Z_α 。 (或计算 Z 对应的 p 值)	同左
4	决策 : 若 $ Z > Z_{\frac{\alpha}{2}}$ (或 $p < \alpha$) , 拒绝 H_0 ; 否则不否定 H_0 。	决策 : 若 $Z > Z_\alpha$ (或 $p < \alpha$) , 拒 绝 H_0 ; 否则不否定 H_0 。	决策 : 若 $Z < -Z_\alpha$ (或 $p < \alpha$) , 拒绝 H_0 ; 否则不否定 H_0 。

实际应用中 , 经常检验两总体均值是否相等 , 即两总体均值的差异是否为零 ($\Delta = 0$)。

若两个总体的方差 σ_1^2 和 σ_2^2 未知 , 当 $\sigma_1^2 = \sigma_2^2$ 时 , 记

$$S_w = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} ;$$

当 $\sigma_1^2 \neq \sigma_2^2$ 时 , 记

非独立样本的 t 检验参见有关文献。

$$S_w = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

检验两总体均值差是否等于某已知常数采用 t 检验法（如表 5-22 所示）。

表 5-22 两正态总体方差未知时的均值比较 t 检验表

步骤	双侧检验	右单侧检验	左单侧检验
1	提出假设： $H_0: \mu_1 - \mu_2 = \Delta$ $H_1: \mu_1 - \mu_2 \neq \Delta$ (Δ 为常数)	提出假设： $H_0: \mu_1 - \mu_2 = \Delta$ $H_1: \mu_1 - \mu_2 > \Delta$ (Δ 为常数)	提出假设： $H_0: \mu_1 - \mu_2 = \Delta$ $H_1: \mu_1 - \mu_2 < \Delta$ (Δ 为常数)
2	构造检验统计量： $t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{S_w}$ 若 H_0 成立， $t \sim t(n_1 + n_2 - 2)$	同左	同左
3	根据显著性水平 α 和自由度查表确定临界值 $t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$ 。（或计算 t 对应的 p 值）	根据显著性水平 α 和自由度查表确定临界值 $t_{\alpha}(n_1 + n_2 - 2)$ 。（或计算 t 对应的 p 值）	同左
4	决策：若 $ t > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$ ， （或 $p < \alpha$ ） 拒绝 H_0 ；否则不否定 H_0 。	决策：若 $t > t_{\alpha}(n_1 + n_2 - 2)$ ， （或 $p < \alpha$ ） 拒绝 H_0 ；否则不否定 H_0 。	决策：若 $t < -t_{\alpha}(n_1 + n_2 - 2)$ ， （或 $p < \alpha$ ） 拒绝 H_0 ；否则不否定 H_0 。

如果检验两总体均值是否相等，这时 $\Delta = 0$ 。

两个正态总体方差未知而且不等时使用的计算 t 值的公式是不同的，因此进行 t 检验时，必须先检验两正态总体的方差是否相等，再决定使用哪一个式子计算 t 值。通常采用 Levene 检验法对两总体方差是否存在显著性差异进行检验。步骤如下：

计算每个样品对组平均值的绝对误差，对这些误差做一元方差分析（详见下一章），从而计算出 F 值，并求出 p 值。当 $p < \alpha$ ，拒绝两个总体方差相等的零假设，即两总体方差不等；否则不能拒绝零假设，认为两总体方差相等。

3、SPSS 操作

SPSS 提供了计算指定变量的综合描述统计量的过程和对均值进行比较检验的过程：

（1）用于计算变量的综合统计量的 Means 过程

[Analyze] => [Compare Means] => [Means]

（2）用于单独样本的 t 检验过程

[Analyze] => [Compare Means] => [One-Sample T Test]

（3）用于独立样本的 t 检验过程

[Analyze] => [Compare Means] => [Independent-Samples T Test]

用于检验是否两个不相关的样本来自具有相同均值的总体。

（4）用于配对样本的 t 检验过程

[Analyze]=>[Compare Means]=>[Paired-Samples T Test]

用于检验两个相关的样本是否来自具有相同均值的总体。

[例 5-6] 分别测得 14 例老年慢性支气管炎病人及 11 例健康人的尿中 17 酮类固醇排出量 (mg/dl) 如下, 试比较两组均值有无显著性差别 ($\alpha=0.05$)。

病 人	2.90	5.41	5.48	4.60	4.03	5.10	4.97	4.24	4.36	2.72
	2.37	2.09	7.10	5.92						
健康人	5.18	8.79	3.14	6.46	3.72	6.64	5.60	4.57	7.71	4.99
	4.01									

(1) 定义变量: 把实际观察值定义为 X, 再定义一个变量 G 来区分病人与健康人。输入原始数据, 在变量 G 中, 病人输入 1, 健康人输入 2。

(2) 选择[Analyze]=>[Compare Means]=>[Independent-Samples T Test], 打开[Independent-samples T Test]主对话框。从主对话框左侧的变量列表选中 X, 单击按钮使之进入[Test Variable(s)]列表框, 选 G 单击按钮使之进入[Grouping Variable]框, 单击[Define Groups]按钮弹出[Define Groups]定义框, 在[Group 1]中输入 1, 在[Group 2]中输入 2, 单击[Continue]按钮, 返回[Independent-samples T Test]主对话框 (如图 5-7所示), 单击[OK]按钮即完成。

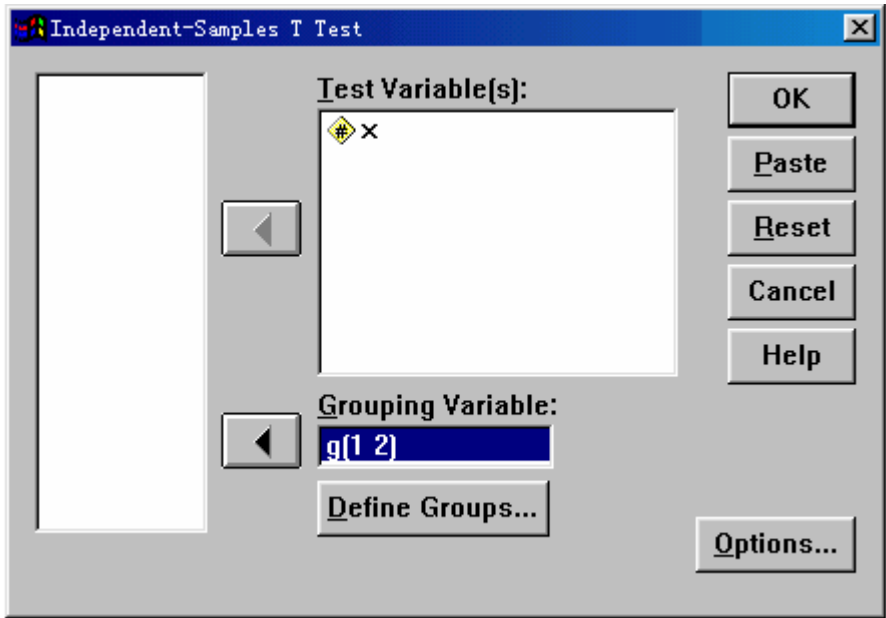


图 5-7 独立样本 T 检验主对话框

表5-23 独立样本T检验结果

(分组统计量)		G	N	Mean (均值)	Std. Deviation (标准差)	Std. Error Mean (均值标准误)
X		1	14	4.3779	1.4499	.3875
		2	11	5.5282	1.7354	.5232

(检验结果)		Levene's Test for Equality of		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
X	Equal variances assumed	.440	.514	-1.807	23	.084

Equal variances not assumed			-1.767	19.472	.093
-----------------------------	--	--	--------	--------	------

表 5-23 检验结果中, 经 Levene 方差齐性检验: $F = .440, p \text{ 值} = .514, p > \alpha$, 两总体方差无显著性差异。第三行表示方差齐性情况下的 t 检验的结果, 第四行表示方差不齐情况下的 t 检验的结果。依次显示 t 值 ($t\text{-value}$)、自由度 (df)、双侧检验 p 值 (Sig 2-Tail) 等。因本例属方差齐性, 故采用第三行 (即 Equal variances assumed) 结果: $t = -1.807, p = 0.084 < 0.1$, 差异显著, 即老年慢性支气管炎病人的尿中 17 酮类固醇排出量低于健康人。

五、假设检验的两类错误分析

零假设究竟是真实还是不真实, 事实上是不知道的。在参数检验中, 我们接受零假设仅仅由于它出现的可能性比较大, 而拒绝零假设也仅仅由于它出现的可能性小。这样按概率大小所作的判断, 并不能保证百分之百的正确, 不论是接受零假设或拒绝零假设都可能犯错误, 总是要承担一定的风险。所作的判断不外以下四种情况:

- (一) 零假设是真实的, 而作出接受零假设的判断, 这是正确的决定。
- (二) 零假设是不真实的, 而作出拒绝接受零假设的判断, 这是正确的决定。
- (三) 零假设是真实的, 而作出拒绝零假设的判断, 这是犯了第一类型的错误。
- (四) 零假设是不真实的, 而作出接受零假设的判断, 这是犯了第二类型的错误。

这四种情况构成如下统计决策表:

表 5 - 24

	接受	拒绝
H_0 真实	正确的决定($1 - \alpha$)	第一类型错误(α)
H_0 不真实	第一类型错误(β)	正确的决定($1 - \beta$)

在作检验决策的时候, 当然希望所有真实的零假设都能得到接受, 尽量避免真实的假设被拒绝, 少犯或不犯第一类型的错误。也希望所有不真实的零假设都被拒绝, 尽量避免不真实的假设被接受, 少犯或不犯第二类型的错误。因此需要对可能犯第一类型或第二类型错误的概率作分析。

假设检验是建立在小概率事件几乎不会发生的原理基础上, 给定显著性水平 α , 如果样本均值与总体均值的差异出现的概率等于或小于 α , 则认为此事件可能性很小, 因此就拒绝零假设。但是这个差异的发生并不是完全不可能, 而是有 α 的可能性存在。这就是说, 有 α 的可能性发生零假设是真实的而被拒绝了, 所以显著性水平 α 实际上就是犯第一类型错误的概率, α 也称为拒真概率。犯第一类型的错误所引起的损失可能很大, 例如实际无效的药物而决定大批量生产等都会造成很大的浪费。因此要根据实际需要, 对显著性水平 α 加以控制。 α 定的越小, 则犯第一类型错误的可能性也越小, 例如 $\alpha = 0.05$, 表示可以保证判断时犯第一类错误的可能性不超过 5%, 而当 $\alpha = 0.01$ 时, 则保证犯第一类型错误的可能性不超过 1% 等等。

但是, 第一类型错误和第二类型错误又是一对矛盾, 在其他条件不变下, 减少犯第一类型错误的可能性, 势必增加犯第二类型错误的可能性。即增加零假设是不真实的, 而被接受的错误, 设犯第二类型错误的概率为 β , 则 β 称为纳伪概率。犯第二类型错误也可能引起很大损失, 例如把有显著效果的新药检验为无效果, 以致不敢投入生产, 使某种疾病蔓延, 贻误不浅, 要比较第一类型错误与第二类型错误的损失哪个更大, 就要对不同情况作具体的分析, 例如新药的成本很低廉, 不妨冒犯第一类型错误的危险, 如果新药成本很昂贵, 又宁肯冒犯

第二类型错误的危险。

如果说 β 表示接受不真实的零假设的概率, 那么 $1 - \beta$ 就是表示拒绝不真实的零假设的概率, $1 - \beta$ 的值接近于 1, 表示不真实的零假设几乎都能够加以拒绝, 反之 $1 - \beta$ 接近于 0, 表示犯第二类型的错误的可能性是很大的, 因此 $1 - \beta$ 是表明检验工作做得好坏的一个指标, 称为检验功效(Test Power)。一般地说检验功效随着备择假设的真值与不真实的零假设距离有关, 离零假设愈远的检验功效也愈高, 但是由于备择假设的真值通常是不知道的, 而且 β 的大小又和显著性水平 α 成反比变化, 因此在假设检验时总是将冒第一类型错误的风险概率固定下来, 对所得的结果进行判断。要同时减少一、二两类错误的概率, 只有增加样本单位数, 但在实际工作中, 不可能无限增大样本容量, 因而选择控制第一类型错误便是更切实际的办法。

第六章 方差分析

第一节 单因素方差分析

一、方差分析概述

在实际问题中,经常使用上一章的方法对两个正态总体进行均值比较,即检验两个样本是否取自同一总体。如果分组样本不止两个,就必须使用方差分析(ANOVA: ANalysis Of VAriance)对它们所取自的总体进行均值比较。也就是说方差分析是检验两个总体或多个总体的均值间差异是否具有统计意义的一种方法。既然方差分析通常用于均值比较,那么把它称为“ANOVA 方差分析”似乎是不合适的,为什么不用“均值分析”(ANOME, ANalysis Of MEans)来代替呢?事实上,用“方差分析”这个名称是很有道理的:虽然经常比较的是均值,但比较时是采用方差的估计量进行分析的。方差分析所使用的检验统计量是 F 统计量,它是方差估计值之比。这里不是根据用途而是根据分析方法来命名的。

方差分析与第八章将讨论的回归分析之间存在一定的关系。对于方差分析,所有的自变量都被视为定类变量;而回归分析中,自变量可以是各种测度的变量(包括定类变量、定序变量、定距变量和定比变量)。事实上,经常把方差分析看作回归分析的一种特例,几乎所有方差分析模型可以由回归模型来表示,可以用回归分析的一般方法估计出相应的参数并进行推断。

为使方差分析更加有效,一般要假定所比较的总体具有相同的方差和正态分布,正如我们在前一章中比较两个均值时那样。不过,方差分析方法在更宽的条件也还是近似有效的,因此称之为是稳健的。

二、方差分析原理

我们通过一个具体的例子来说明方差分析原理。

[例 6-8]在 1990 年秋对“亚运会期间收看电视的时间”调查结果如表 6-25、表 6-26 所示。问:收看电视的时间比平日减少了(第一组)与平日无增减(第二组)比平日增加了(第三组)的三组居民在“对亚运会的总态度得分”上有没有显著的差异?即要检验从“态度”上看,这三组居民的样本是取自同一总体还是取自不同的总体。

表 6-25 三组居民的样本
的态度得分

第一组	第二组	第三组
42	39	43
41	40	44
42	40	43
42	41	45
43	40	45
$\bar{x}_1=42$	$\bar{x}_2=40$	$\bar{x}_3=44$

表 6-26 同一组居民抽取的三个样本
的态度得分

样本 1	样本 2	样本 3
39	40	41
42	44	44
44	43	44
40	40	45
40	43	41
$\bar{x}_1=41$	$\bar{x}_2=42$	$\bar{x}_3=43$

(一) 组间的差异:

假定将某一居民点的居民分成三个组,由于种种原因,每一组居民在“态度得分”上是

随机波动的。因此从每一组中各随机地抽取 5 位居民,测量了他们的“态度得分”如表 6-25 所示,表中还给出了每个样本的平均态度得分值。

首先要问的一个问题是“这三个组的态度真的存在差异吗”,也就是说,表 6-25 中样本均值 $\bar{x}$ 的不同是由于潜在总体均值 μ 的不同 μ 产生的吗(μ 表示其中一个组的全体居民的平均态度分)?如果不是,那么样本均值 $\bar{x}$ 中的这些差异是否可以认为仅仅是由于随机波动造成的?

为了说明这一点,假定我们只从某一组中抽取三个样本,如表 6-26 所示。正如我们所预料的,尽管在这种情况下三个样本取自同一总体,因而其均值 μ 是相同的,但抽样的波动也引起了各个 $\bar{x}$ 之间的小小差别。因此可以将问题重新叙述如下:“表 6-25 中 $\bar{x}$ 间的差别和表 6-26 中 $\bar{x}$ 间的差别大体上阶数相同呢(因此说明表 6-25 中 $\bar{x}$ 间的差别也是由于随机波动造成的),还是表 6-25 中 $\bar{x}$ 间的差别大得多,从而足以说明潜在总体的均值 μ 之间存在差异呢?”由直观上看,似乎后一种解释更符合实际。那么,怎样给出一个正确的检验呢?和通常那样,在总体均值中“无差异”的假设称为原假设,即

$$H_0: \mu_1 = \mu_2 = \mu_3$$

检验 H_0 首先要求测量一下样本均值之间相差多少。为此要找到一个合适的能描述各组之间变动的量。我们先求出这三个样本的总平均值 $\bar{\bar{x}}$,

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{g} = \frac{42 + 40 + 44}{3} = 42$$

其中 g 表示组数。然后计算样本均值 $\bar{x}$ 相对于其总均值 $\bar{\bar{x}}$ 的总方差:

$$S_{\bar{x}}^2 = \frac{1}{g-1} \sum (\bar{x} - \bar{\bar{x}})^2 = \frac{1}{3-1} [(42-42)^2 + (40-42)^2 + (44-42)^2] = 4$$

这个方差公式和一般的方差公式是类似的,只是将 X 换成 $\bar{x}$,将 n 换成 g 而已。

由 $S_{\bar{x}}^2$ 的定义可知,它是一个描述组间变动的量。对于表 6-26 中的数据,由于三个样本取自同一总体,我们猜测样本间方差应该比较小。利用表 6-25 和表 6-26 中数据可求得:

$$\begin{aligned} \bar{\bar{x}} &= \frac{41 + 42 + 43}{3} = 42 \\ S_{\bar{x}}^2 &= \frac{1}{3-1} [(41-42)^2 + (42-42)^2 + (43-42)^2] = 1 \end{aligned}$$

与从表 6-25 求得的 $S_{\bar{x}}^2 = 4$ 相比,表 6-26 数据中各组间的差异要小多了。

(二) 组内的差异

我们前面给出的各组均值之间的方差 $S_{\bar{x}}^2$ 还不能完全说明问题。例如,考虑表 6-27 的数据,显然,它的总方差 $S_{\bar{x}}^2$ 和表 6-25 的相同。但是每一组的样本其态度得分都是十分不稳定的,每组都有很大的随机波动。为了进一步直观地比较表 6-25 和表 6-27 的数据,我们在图 6-65 的(A)和(B)中分别给出了潜在总体的可能形状。从(B)中可以看到,表 6-27 对应的三个组的态度得分是十分不稳定的,因此三个样本都有可能是取自同一总体的。也就是说,样本均值之间的差异可以解释为是随机波动产生的。但对于表 6-25,从图 6-65 的(A)可以看到样本均值间的差异却很难用随机因素来解释,因为在这种情况下三个组内的态度得分并非那么不稳定。

表 6-27 不稳定的三组居民的样本态度得分

样本 1	样本 2	样本 3
------	------	------

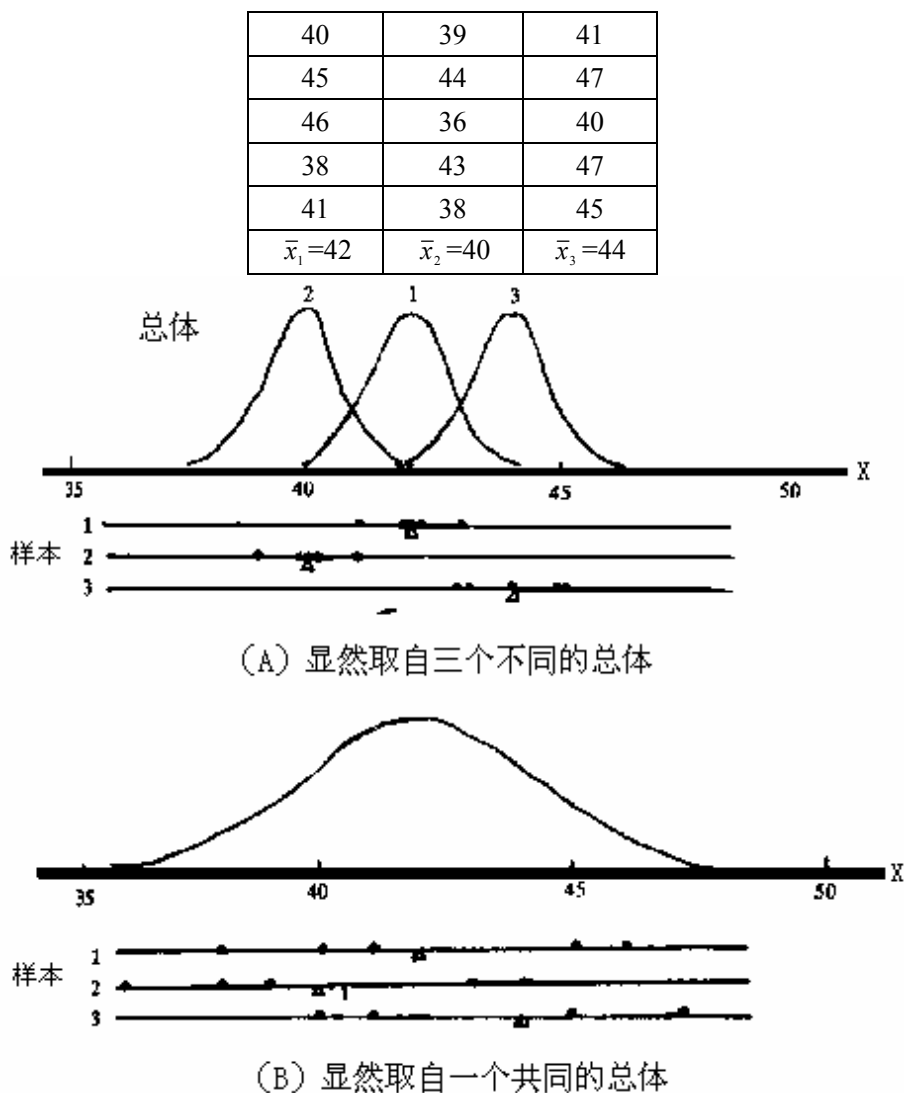


图 6-65 表 6-25与表 6-27的数据比较

现在我们就有了比较的标准。在图 6-65 的 (A) 中, 我们的结论是: 三个 μ 之间是不相同的, 因为样本均值的方差 $S_{\bar{x}}^2$ 相对于随机波动来说是比较大的, 因此我们拒绝零假设。那么我们怎样才能够度量这些随机波动即组内的变差呢? 从直观上看, 应当是每个样本内观测值的变化程度或偏离其均值的程度。为此我们先计算表 6-25 中第一个样本内的离差平方和:

$$\sum (x_i - \bar{x}_1)^2 = (42 - 42)^2 + (41 - 42)^2 + (42 - 42)^2 + (42 - 42)^2 + (43 - 42)^2 = 2$$

类似地计算第 2 个样本和第 3 个样本内的离差平方和, 将它们相加。然后用所有 3 个样本的总自由度 (每个样本的自由度都为 $n-1=4$) 去除, 这样就得到了两样本的联合方差 S_p^2 :

$$S_p^2 = \frac{2+2+4}{4+4+4} = \frac{8}{12} = \frac{2}{3}$$

联合方差的计算很容易推广到有 g 组的数据, 而每组内有 n 个观测值的情形:

$$S_p^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \cdots + \sum (x_g - \bar{x}_g)^2}{g(n-1)}$$

由 S_p^2 的定义可知，它是一个描述组内变动的量。对于表 6-27 的数据，我们可以猜测到其联合方差 S_p^2 一定比表 6-25 的大得多，计算得

$$S_p^2 = \frac{46+46+44}{4+4+4} = \frac{34}{3} \gg \frac{2}{3}$$

(三) F 值

那么现在就可以给出关键的式子（即检验统计量）了。是否拒绝原假设 H_0 ，要看组间变异相对于组内变异来说是否足够大。也就是说要考察比值 S_x^2/S_p^2 的大小。习惯上是用一个稍微修改一下的比值：

$$F = \frac{nS_x^2}{S_p^2}$$

其中分子多乘一个 n 是为了使当 H_0 为真时分子的值平均上来说等于分母，用会大于 S ；这时在 (8-5) 式中的 F 比值将倾向十比 1 大得多。因此， F 值越大，原假设 H_0 就越不可信。可以证明， F 检验统计量在成立时服从第一自由度为 $g-1$ 、第二自由度为 $g(n-1)$ 的 F 分布。

[例 6-9] 对表 6-25 的数据，我们已经求出了三个样本之间的总方差

$$S_x^2 = 4$$

以及三个样本内的联合方差

$$S_p^2 = \frac{2}{3}$$

问：总体均值之间是否存在显著差异？

解：1) 提出假设

H_0 ：总体均值之间没有差异

H_1 ：总体均值之间存在差异

2) 计算检验统计量

$$F = \frac{nS_x^2}{S_p^2} = \frac{5 \times 4}{\frac{2}{3}} = 30$$

在 H_0 为真条件下服从第一自由度为 $g-1=3-1=2$ ，第二自由度为 $g(n-1)=3(5-1)=12$ 的 F 分布。

3) 查附表得，

$$p \text{ 值} < 0.001$$

这意味着如果 H_0 为真，那么抽取到如表 6-25 那样有这么大差异的三个样本的机会小于 1%。因此可以认为表 6-25 中的三个组其态度得分均值真是不同的，即收看电视时间不同的三个组其对亚运会的态度是属于三个不同的总体。

下面再看看对于表 6-26 和表 6-27 的数据 F 检验说明了什么。

[例 6-10] 利用表 6-26 中的数据计算 H_0 的概值；

2) 利用表 6-27 中的数据计算 H_0 的概值；

解：1) 我们已在前面求得 $S_x^2 = 1$ ，利用表 6-26 数据可计算

$$S_p^2 = \frac{16+14+14}{4+4+4} = \frac{11}{3},$$

因此

$$F = \frac{5 \times 1}{11/3} = 1.36$$

查第一、二自由度分别为 2、12 的 F 分布表，得到

$$p \text{ 值} > 0.25$$

这说明不能否定 H_0 。这是正确的结论，因为我们正是从同一组（总体）中抽出表 6-26 的三个样本的。

2) 前面已求得表 6-27 的 $S_{\bar{x}}^2 = 4, S_p^2 = 34/3$ ，因此

$$F = \frac{5 \times 4}{34/3} = 1.76$$

查第一、二自由度分别为 2 和 12 的 F 分布临界值表，得

$$0.10 < p \text{ 值} < 0.25$$

就是说， H_0 也是不能否定的，我们真的没有什么根据去断有这个组的平均态度是不同的。样本均值中那么大的差异所以可能发生是因为每一组的态度得分都很不稳定（波动很大），并不一定因为各组的平均态度真的有什么显著的差异。

（四）方差分析表

下面介绍一种称之为方差分析表的标准形式的表格，它将面所讲述的计算以简洁的形式进行了总结，其形式如表 6-28 所示。表中的第一行说明 F 值中分子的计算，第二行是分母的计算。

表 6-28 方差分析表的一般形式

差异的来源	离差平方和	自由度	均方差(平均平方和)	F 值
组间差异(由于 $\bar{x}$ 的差异造成的)	$SS_b = n[(\bar{x}_1 - \bar{\bar{x}})^2 + (\bar{x}_2 - \bar{\bar{x}})^2 + \dots + (\bar{x}_g - \bar{\bar{x}})^2]$	(g-1)	$MSS_b = SS_b / (g-1) = nS_{\bar{x}}^2$	$F = \frac{MSS_b}{MSS_w} = \frac{nS_{\bar{x}}^2}{S_p^2}$
组内差异(由于随机波动造成的残差)	$SS_w = \sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \dots + \sum (x_g - \bar{x}_g)^2]$	g(n-1)	$MSS_w = SS_w / [g(n-1)] = S_p^2$	
总和	$SS_t = \sum \sum (x - \bar{\bar{x}})^2$	(gn-1)		

用方差分析表还可以检查你的计算是否正确。在第 2 列中， SS_b 表示组间的离差平方和， SS_w 表示组内的离差平方和，最后的 SS_t 表示每一个数据对总均值的离差平方和，叫做总离差平方和。可以证明在一般情况下总离差平方和等于组间离差平方和和组内离差平方和之和，即

$$SS_t = SS_b + SS_w$$

（总离差平方和=组间离差平方和+组内离差平方和）

此外，表 6-28 中的自由度也可以用同样的方法来检查，因为有

总自由度 = 第一自由度（分子自由度）+ 第二自由度（分母自由度）

在计算过程中可以利用这些关系确定你的离差平方和与自由度是否都加对了。表 6-29 给出了对应于表 6-25 的方差分析表。用自由度去除对应离差平方和，就得到了表 6-29 的均方差。根据各组可能属于不同的总体（态度有差异的总体）这一事实可以“解释”组间的均方差。组内的均方差是“不能解释的”，因为它们是无法系统地（用总体的差异）来解释的随机或偶然的均方差。因此 F 值有时也叫均方差比，即

$$F = \frac{\text{可以解释的均方差}}{\text{不能解释的均方差}}$$

表 6-29 表 6-25 数据的方差分析表

差异来源	离差平方和	自由度	均方差	F 值	p 值
组间	40	2	20	$\frac{20}{2/3} = 30$	$p < 0.001$
组内	8	12	2/3		
总和	48	14			

（五）样本容量不相等的情形

在表 6-25 中，对每一组所取的观测数（ $n=5$ ）是相同的，一般来说，这是收集数据的比较有效的办法，即让所有的样本有相同的容量 n 。不过，当样本容量 $n_1, n_2, n_3, \dots$ 不相同，也很容易适当地将方差分析的计算修改一下。

现在总观测数是 $n_1 + n_2 + \dots + n_g = N$ 而不再是 nc 。表中所有数值的总平均为：

$$\bar{\bar{x}} = \frac{\sum \sum x}{n_1 + n_2 + \dots + n_g} = \frac{\sum \sum x}{N}$$

或者，将总均值清楚地表示成各组均值的一种加权平均的形式，即

$$\bar{\bar{x}} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_g \bar{x}_g}{n_1 + n_2 + \dots + n_g} = \frac{\sum n_i \bar{x}_i}{N}$$

各组之间的离差平方和也相应变成

$$SS_b = n_1 (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 (\bar{x}_2 - \bar{\bar{x}})^2 + \dots$$

总自由度 $df = n_1 + n_2 + \dots + n_g - 1 = N - 1$

组内自由度 $df = (n_1 - 1) + (n_2 - 1) + \dots + (n_g - 1) = N - g$

表 6-30 方差分析表（样本容量不相等时）

差异的来源	离差平方和 SS	自由度 df	均方差 MSS	F 值
组间差异 (由于 $\bar{x}$ 的差异造成的)	$SS_b = \sum_{i=1}^g n_i (\bar{x}_i - \bar{\bar{x}})^2$	$(g-1)$	$MSS_b = SS_b / (g-1)$	$F = \frac{MSS_b}{MSS_w}$

组内差异 (由于随机波动造成的残差)	$SS_w = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	(N-g)	$MSS_w = SS_w / (N - g)$	
总和	$SS_t = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$	(N-1)		

(六) 方差分析的 SPSS 操作

以[例 6-1]的数据为例，在 SPSS 中进行方差分析的步骤如下：

(1) 定义“居民对亚运会的总态度得分”变量为 X (数值型)，定义组类变量为 G (数值型)，G=1、2、3 表示第一组、第二组、第三组。然后录入相应数据，如图 6-66 所示。

	g	x
1	1	42.00
2	1	41.00
3	1	42.00
4	1	42.00
5	1	43.00
6	2	39.00
7	2	40.00
8	2	40.00
9	2	41.00
10	2	40.00
11	3	43.00
12	3	44.00
13	3	43.00
14	3	45.00
15	3	45.00
16		

图 6-66 方差分析数据格式

(2) 选择[Analyze]>[Compare Means]>[One-Way ANOVA...]，打开[One-Way ANOVA]主对话框(如图 6-67 所示)。从主对话框左侧的变量列表选定 X，单击按钮使之进入[Dependent List]框，再选定变量 G，单击按钮使之进入[Factor]框。单击[OK]按钮完成。

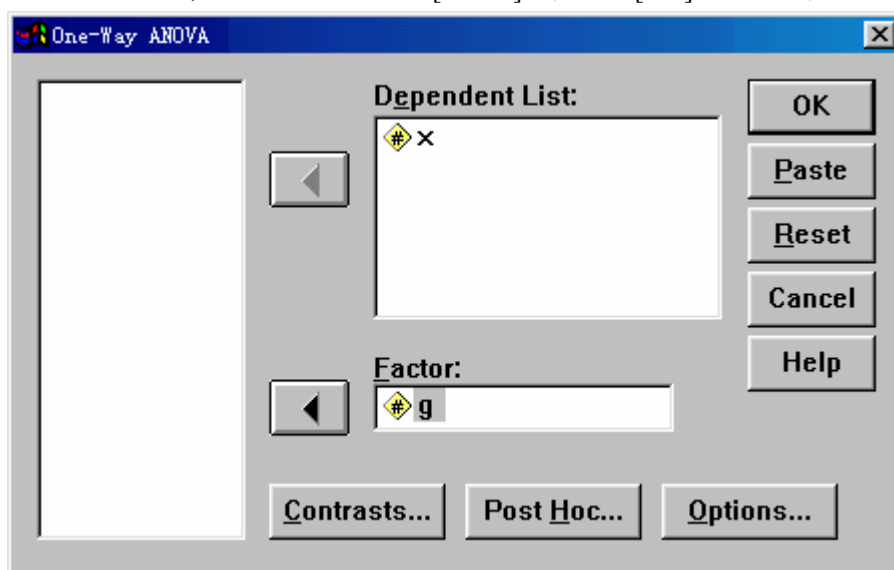


图 6-67 方差分析对话框

(3) 分析结果如下：

ANOVA X	Sum of Squares (离差平方和)	df (自由度)	Mean Square (均方)	F	Sig. (p值)
Between Groups (组间)	40.000	2	20.000	30	.000
Within Groups (组内)	8.000	12	.667		
Total (总和)	48.000	14			

该结果与表 6-29是一致的。

第二节 多因素方差分析

到现在为止，我们研究的是一种响应，例如态度得分，如何依赖于一个因子（因素），例如看电视的时间增减。又例如收入如何依赖于性别，产量如何依赖于机器的类型等等。通常称之为单因素的方差分析。但是在实践中，一种响应可能依赖于两个、三个或更多的因素。例如，态度不仅可能依赖于看电视时间的增减，还依赖于文化程度、年龄等等；收入不仅可能与性别有关，还可能与文化程度、工作业绩等有关；产量不仅取决于机器类型、还可能取决于操纵机器的经验、或原材料的质量等等。在这种情况下就要将表 6-28 的方差分析推广到多个因素的情况，所有这些因素都可能作为解释差异的来源。对应表就叫做多因素的方差分析表（如双因素或三因素方差分析表，等等）。这里仅通过一个双因素方差分析的例子给出对应的方差分析表。实际上这些计算用 SPSS 是十分容易解决的。因此读者完全没有必要去记表中繁琐的公式，只需掌握其主要思想并学会应用 SPSS 就可以了。

[例 6-11] 从由五名操作者操作的三台机器每小时产量中分别各抽取 1 个不同时段的产量，观测到的产量如表 6-31 所示。试进行产量是否依赖于机器类型和操作者的方差分析。

表 6-31 三台机器五名操作者的产量数据

$\begin{matrix} i \\ j \end{matrix}$	机器 1	机器 2	机器 3	操作者均值 $\bar{x}_{.j}$
操作者 1	53	61	51	55
操作者 2	47	55	51	51
操作者 3	46	52	49	49
操作者 4	50	58	54	54
操作者 5	49	54	50	51
机器均值 $\bar{x}_{i.}$	49	46	51	$\bar{\bar{x}} = 52$

为了给出一般形式的方差分析表，我们对常用的术语、各种记号及其意义作如下的简要说明：

(1) 因素(Factors)与处理(Treatments)

因素是影响因变量变化的客观条件；处理是影响因变量变化的人为条件；也可以统称为因素。

本例中，因素 1：机器类型（c 台机器，或 c 列）；

因素 2：操作者（ r 名操作者，或 r 行）；

（2）水平(Level)

因素的不同等级称水平。例如，性别因素在一般情况下只研究两个水平：男、女。

（3）单元(Cell)

单元指各种因素的水平之间的每个组合。例如，研究问题中的因素有性别 gender，取值为 0、1；还有年龄 age，分三个水平 1（10 岁）、2（11 岁）、3（12 岁）。两个变量的组合共可形成六个单元，即（1,1）、（1,2）、（1,3）、（2,1）、（2,2）、（2,3），代表两种性别与三种年龄的六种组合。

本例中，机器类型和操作者组合形成 15 个单元。

x_{ij} ：第 i 台机器由第 j 个操作者操作时的产量；

$\bar{x}_{i\cdot} = \sum_{j=1}^r x_{ij} / r$ ：第 i 台机器的平均产量；

$\bar{x}_{\cdot j} = \sum_{i=1}^c x_{ij} / c$ ：第 j 个操作者的平均产量；

一般地， x_{ij} 的预测值 $\hat{x}_{ij}$ 可表示为

$$\hat{x}_{ij} = \bar{\bar{x}} + (\bar{x}_{i\cdot} - \bar{\bar{x}}) + (\bar{x}_{\cdot j} - \bar{\bar{x}})$$

=总平均+由因素 1 所影响的部分+由因素 2 所影响的部分（例如， $\hat{x}_{21}=52+4+3=59$ ，

而 $x_{ij}=61$ ）

$$\text{残差} = x_{ij} - \hat{x}_{ij}$$

由此可以求出全部的残差 $x_{ij} - \hat{x}_{ij}$ 以及残差平方和 $\sum \sum (x_{ij} - \hat{x}_{ij})^2 = \sum \sum (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$ ，并

且可以证明下面的总离差平方和分解式：

$$SS_t = SS_c + SS_r + SS_{res}$$

（总离差平方和=列间离差平方和+行间离差平方和+残差平方和）

根据该式，即可给出如下的方差分析表。

表 6-32 双因素方差分析表的一般形式

差异的来源	离差平方和 SS	自由度 df	均方差 (平均平方和 MSS)	F 值
列间差异(由于列均值 $\bar{x}_{i\cdot}$ 的差异造成的)	$SS_c = r \sum_{i=1}^c (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$	$(c-1)$	$MSS_c = SS_c / (c-1)$	$\frac{MSS_c}{MSS_{res}}$

行 间 差 异(由于 行 均 值 $\bar{x}_{\cdot j}$ 的 差 异 造 成 的)	$SS_r = c \sum_{j=1}^r (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$	(r-1)	$MSS_r = SS_r / (r-1)$	$\frac{MSS_r}{MSS_{res}}$
残差(由 实 际 观 测 值 和 拟 合 值 间 的 差 异 造 成 的)	$SS_{res} = \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$	(c-1) · (r-1)	$MSS_{res} = \frac{SS_{res}}{[(c-1)(r-1)]}$	
总和	$SS_t = \sum_{i=1}^c \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2$	(rc-1)		

按表 6-32 的形式，可以给出对应于表 6-31 数据的方差分析表（如表 6-33 所示）。

表 6-33 方差分析表(表 6-31 的数据)

差异的来源	SS	df	MSS	F 值	p 值
机器间	130	2	65	23.6	p<0.001
操作者间	72	4	18	6.5	p<0.05
残差	22	8	2.75		
总和	224	14			

因此，可以认为机器类型和操作者的影响均是显著的。

SPSS 的操作步骤为：

（1）定义“操作者的产量”变量为 X（数值型），定义机器因素变量为 G1（数值型），操作者因素变量为 G2（数值型），G1=1、2、3 分别表示第一、二、三台机器，G2=1、2、3、4、5 分别表示第 1、2、3、4、5 位操作者。录入相应数据，如图 6-68 所示。

	g1	g2	x
1	1	1	53.00
2	1	2	47.00
3	1	3	46.00
4	1	4	50.00
5	1	5	49.00
6	2	1	61.00
7	2	2	55.00
8	2	3	52.00
9	2	4	58.00
10	2	5	54.00
11	3	1	51.00
12	3	2	51.00
13	3	3	49.00
14	3	4	54.00
15	3	5	50.00

图 6-68 双因素方差分析数据格式

(2) 选择[Analyze]=>[General Linear Model]=>[Univariate...], 打开[Univariate]主对话框(如图 6-69 所示)。从主对话框左侧的变量列表中选定 X, 单击按钮使之进入[Dependent List]框, 再选定变量 G1 和 G2, 单击按钮使之进入[Fixed Factor(s)]框。单击[OK]按钮即可得到与表 6-33 一致的结果。

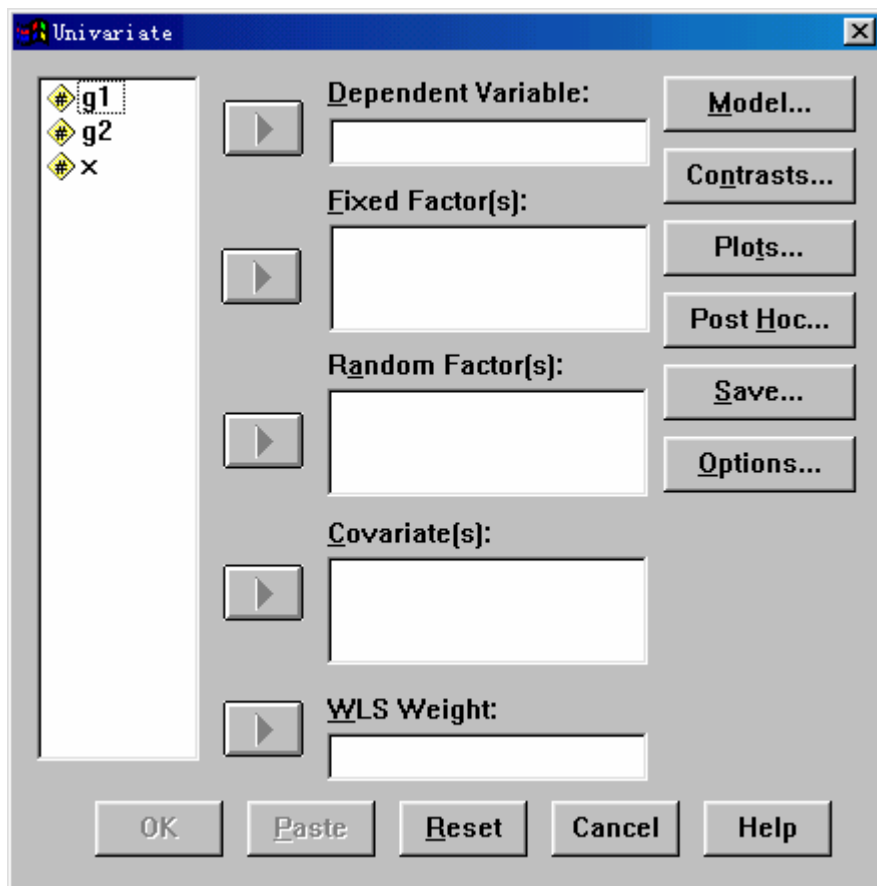


图 6-69 单变量多因素方差分析主对话框

第三节 案例:证券信息的定量分析

随着我国证券市场的迅速发展,对证券信息的分析显得越来越重要了。目前在上海和深圳两个证券交易所挂牌上市的股票已达上千家,投资者进行投资决策时不可能到各个公司去了解情况,而只能依据公司在公开传媒上披露的信息资料进行分析判断,所以多角度、多层次地获取上市公司的全面信息,并正确地加以分析和利用,对于投资者来说是至关重要的。按规定,上市公司的中期报告和年度报告必须在中国证监会指定的报刊上公开刊登。上市公司的中期报告和年度报告主要包括主营业务收入、净利润、总资产、股本收益率、每股净资产、净资产收益率等指标。

以上海证交所为例,上市公司分工业类、商业类、房地产类、公用事业类、综合类五大类,这里从《中国证券报》和《金融时报·每日证券》上收集了上海证交所 159 家上市公司的数据,其中:工业类 88 家、商业类 27 家、房地产类 7 家、公用事业类 13 家、综合类 24 家。净资产收益率反映了公司的盈利能力,我们对它很感兴趣,如何比较某一类上市公司 1994 年度与 1993 年度净资产收益率是否有显著性差异呢?如何比较 1994 年度五大类上市

参见翁小清等:《统计方法在证券信息分析中的应用》,《数理统计与管理》,1997.7。

公司净资产收益率之间是否有显著性差异呢？下面就对这些问题做一些初步探讨。

如何比较某一类上市公司 1994 年度与 1993 年度净资产收益率之间是否有显著性差异呢？可以采用配对 t 检验来解决。配对 t 检验是有前提条件的，都要求样本来自正态总体，而上市公司的净资产收益率的分布情况近似于正态分布，所以可采用配对 t 检验。以上海证交所商业类上市公司为例来说明这个问题，根据商业类上市公司 1993 年与 1994 年的净资产收益率（数据略）使用 SPSS 计算检验统计量 t 值为 2.89191，p 值为 0.00764，小于检验水准 0.05，可得结论：商业类上市公司 1994 年度与 1993 年度的净资产收益率之间差别显著，从数据表可看出 1994 年度净资产收益率高于 1993 年度净资产收益率。

如何比较这五大类上市公司 1994 年度净资产收益率之间是否有显著性差异呢？可以采用单因素方差分析来解决，这里因素就是上市公司的类别，上市公司分为五大类，也就是说类别这个因素有 5 个水平，用 SPSS 软件计算出单因素方差分析结果如表 7-34 所示。

表 7-34 单因素方差分析结果

方差来源	离差平方和	自由度	F 值	p 值
组间	480.8201	4	3.177	0.0153
组内	5826.5656	154		
总和	6307.3857	158		

从表 7-34 中可以看到，实际显著性水平（即 p 值）是 0.0153，小于检验水准 0.05，说明各大类上市公司 1994 年度净资产收益率之间差别显著。再从表 7-35 可以看到：房地产类上市公司净资产收益率最高，它的平均数是 20.0428%，其次是综合类上市公司，其净资产收益率的平均值是 15.2516%。

表 7-35 各大类上市公司净资产收益率的平均数及其 95%置信区间

分类	平均数%	95%置信区间	
		下限	上限
工业类	12.5475	11.2518	13.8431
商业类	12.4129	10.0739	14.7519
房地产类	20.0428	15.4491	24.6366
公用事业类	13.4453	10.0744	16.8163
综合类	15.2516	12.7707	17.7325

第七章 相关分析

第一节 简单相关分析

一、简单相关系数的定义

简单相关分析是对两个变量之间的相关程度进行分析。单相关分析所用的指标称为单相关系数，又称为单相关系数、Pearson（皮尔森）相关系数或相关系数。通常以 ρ 表示总体的相关系数，以 r 表示样本的相关系数。

正如第四章所给出的，总体相关系数的定义式是：

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \quad (7-41)$$

其中， $\text{Cov}(X, Y)$ 是随机变量 X 和 Y 的协方差； $\text{Var}(X)$ 和 $\text{Var}(Y)$ 分别为变量 X 和 Y 的方差。总体相关系数是反映两变量之间线性相关程度的一种特征值，表现为一个常数。

样本相关系数的定义公式是：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (7-42)$$

样本相关系数是根据样本观测值计算的，抽取的样本不同，其具体的数值也会有所差异。可以证明，样本相关系数是总体相关系数的一致估计量。

二、简单相关系数的检验

在实际的客观现象分析研究中，相关系数一般都是利用样本数据计算的，因而带有一定的随机性，样本容量越小其可信程度就越差。因此也需要进行检验，即对总体相关系数 ρ 是否等于 0 进行检验。

数学上可以证明，在 X 与 Y 都服从于正态分布，并且又有 $\rho = 0$ 的条件下，可以采用 t 检验来确定 r 的显著性。其步骤如下：

首先，计算相关系数 r 的 t 值：

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad (7-43)$$

其次，根据给定的显著性水平和自由度（ $n - 2$ ），查找 t 分布表中相应的临界值 $t_{\alpha/2}$ （或 p 值）。若 $|t| > t_{\alpha/2}$ （或 $p < \alpha$ ）表明 r 在统计上是显著的。若 $|t| \leq t_{\alpha/2}$ （或 $p \geq \alpha$ ），表明 r 在统计上是不显著的。

[例 7-12] 某地区统计了机电行业的销售额 Y 和汽车产量 X_1 以及建筑产值 X_2 （如表 7-36 所示），请使用 SPSS 计算 Y 与 X_1 的相关系数并进行显著性检验。

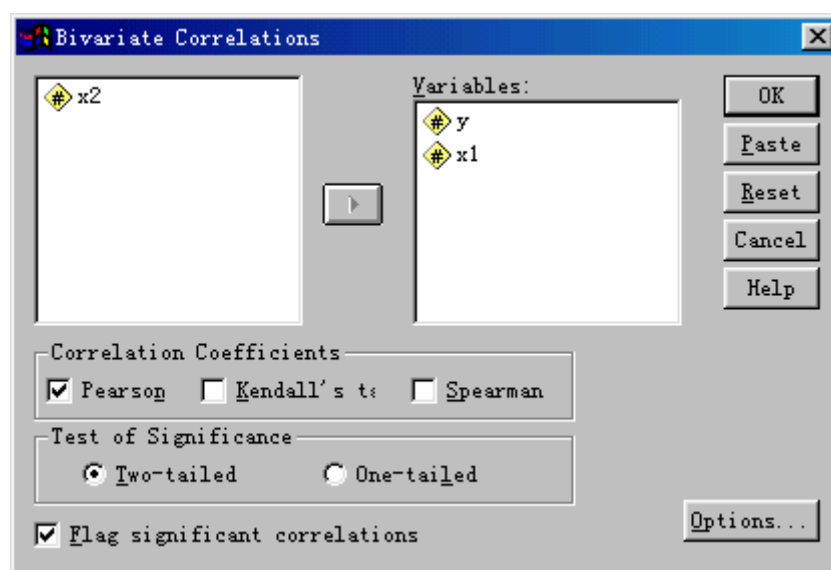
表 7-36 某地区机电行业销售额等数据

年份	销售额 Y (万元)	汽车 X_1 (万辆)	建筑 X_2 (千万元)
1983	280.00	9.43	3.91
1984	281.50	10.36	5.12

1985	337.40	14.50	6.67
1986	404.20	15.75	5.34
1987	402.10	16.78	4.32
1988	452.00	17.44	6.12
1989	431.70	19.77	5.56
1990	582.30	23.76	7.92
1991	596.60	31.61	5.82
1992	620.80	32.17	6.11
1993	513.60	35.09	4.26
1994	606.90	36.42	5.59
1995	629.00	36.58	6.68
1996	602.70	37.14	5.54
1997	656.70	41.30	6.93
1998	778.50	45.62	7.64
1999	877.60	47.38	7.75

解：(1) 根据表 7-36 的数据创建 SPSS 数据文件；

(2) 选择[Analyze]>[Correlate]>[Bivariate]，在显示的如下对话框中，选择变量 Y 和 X₁ 进入[Variables]框。采用默认设置，直接单击[OK]进行分析。



(3) 计算结果如下：

Correlations

		Y	X1
Y	Pearson Correlation	1.000	.948**
	Sig. (2-tailed)	.	.000
	N	17	17
X1	Pearson Correlation	.948**	1.000
	Sig. (2-tailed)	.000	.
	N	17	17

** Correlation is significant at the 0.01 level (2-tailed).

从结果可以看出，Y 与 X₁ 的相关系数 $r=0.948$ ， p 值=0.000，在 $\alpha=0.01$ 水平下线性关

系显著。

三、相关系数的直观意义

相关系数 r 度量了变量 X 和 Y 之间相互联系的程度，我们通过分析公式 (7-42) 来理解这一点。

我们知道，离差 $x - \bar{x}$ 告诉我们离开均值 $\bar{x}$ 有多远，类似地，离差 $y - \bar{y}$ 告诉我们离开均值 $\bar{y}$ 有多远。因此，当我们在二维空间中画出点 (x, y) 时，就可以看出它们离开数据的中心 $(\bar{x}, \bar{y})$ 有多远了。图 7-70 显示了数据点离开其中心（平均销售额、汽车产量）的分散程度。

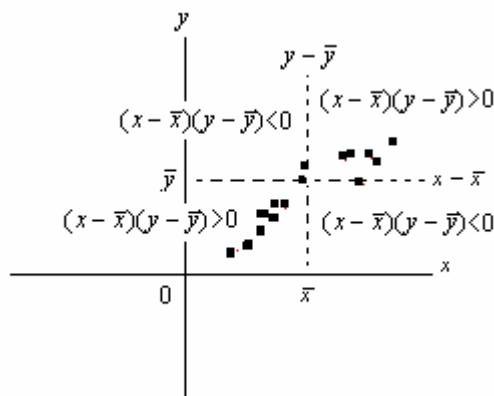


图 7-70 销售额和汽车产量的散点图

假定我们将对应于每年的 $x - \bar{x}$ 和 $y - \bar{y}$ 相乘，再将所有这些乘积相加得到 $\sum (x - \bar{x})(y - \bar{y})$ ，那么 $\sum (x - \bar{x})(y - \bar{y})$ 就给出了销售额和汽车产量如何倾向于一起变化（沿某条直线移动）的一个度量。从图 7-70 可以看到，对于在新坐标系 $(x - \bar{x}, y - \bar{y})$ 中的第 I 或 III 象限的点， $x - \bar{x}$ 和 $y - \bar{y}$ 的符号是相同的，因此乘积 $(x - \bar{x})(y - \bar{y})$ 是正的，相反，对于在第 II 或第 IV 象限的点， $x - \bar{x}$ 和 $y - \bar{y}$ 的符号不同，因此乘积 $(x - \bar{x})(y - \bar{y})$ 是负的，如果 $x - \bar{x}$ 和 $y - \bar{y}$ 是一起沿某条直线移动的，或者说如果 $x - \bar{x}$ 和 $y - \bar{y}$ 是倾向于一起增大或一起减小的，那么大多数的观测点将会落在第 I、III 象限，因此大多数的乘积 $(x - \bar{x})(y - \bar{y})$ 将是正的，它们的和 $\sum (x - \bar{x})(y - \bar{y})$ 也将是正的，这将反映了 X 和 Y 之间的某种正的联系。但是如果 X 和 Y 的联系是负的，即当一个增大时另一个则减小，那么大多数观测点将会落在第 II 和 IV 象限，这样 $\sum (x - \bar{x})(y - \bar{y})$ 就是负的。我们由此可以得出结论：作为度量 X 和 Y 相关的一个数值， $\sum (x - \bar{x})(y - \bar{y})$ 至少在符号上是对的（即 $\sum (x - \bar{x})(y - \bar{y})$ 的正与负表现了 X 与 y 相关的正与负）。而且，当 X 与 Y 之间没有什么线性联系时，观测点将均匀地散布在四个象限上，正项和负项抵消后 $\sum (x - \bar{x})(y - \bar{y})$ 将会是 0。

但是 $\sum (x - \bar{x})(y - \bar{y})$ 有一个缺陷，这就是它依赖于度量 X 与 Y 时的单位。我们希望用一个不随度量单位的变化而变化的量来表示 X 与 Y 之间的联系。怎样调整 $\sum (x - \bar{x})(y - \bar{y})$ 才能得到这样的一个量呢？如果把 $\sum (x - \bar{x})(y - \bar{y})$ 调整为式 (7-42)，就解决了问题，得到的是与 X 、 Y 的单位无关的一个量——相关系数 r 。为了进一步了解 r 的意义，我们在图 7-71 中给出了各种散点图以及相应的相关系数。例如图 7-71 中 (A) 的相关系数为 +1，(B) 的相关系数为 -1，(C) 的相关系数为 +0.8，(D) 相关系数为 -0.8，(E) 和 (F) 的相关系数均为 0。在 (E) 中， X 和 Y 之间根本没有什么联系；可是在 (F) 中， X 和 Y 之间有着很强的联系（实际上是一种曲线联系）。因此， $r=0$ 并不意味着“没有联系”。实际上，它只意味着“没有线性联系”（没有直线关系）。因此， r 仅是线性关系的一种度量。

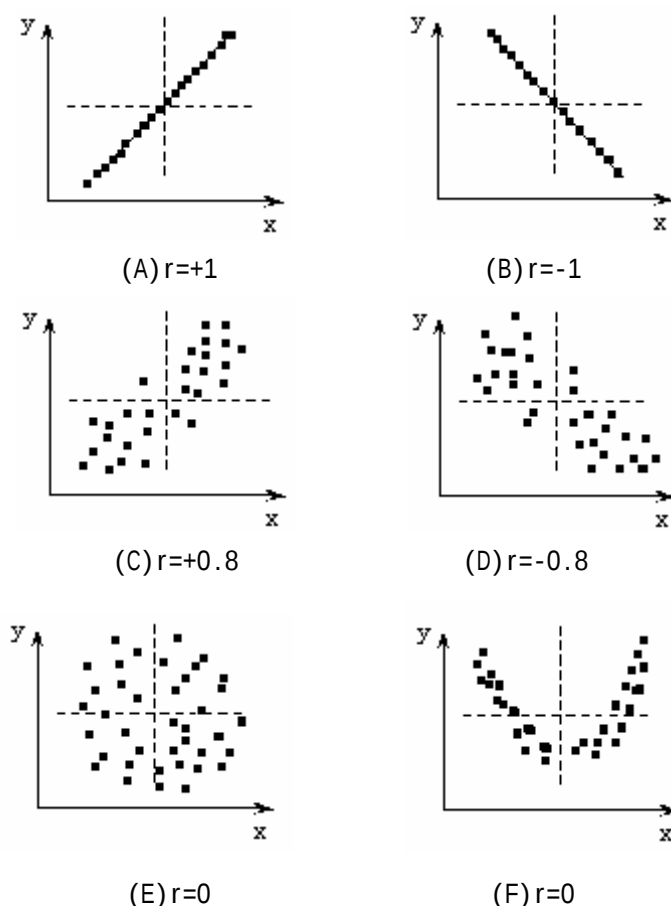


图 7-71 各种相关

第二节 偏相关分析

在多变量的情况下,变量之间的相关关系是很复杂的。因此,多元相关分析除了要利用上一节的简单相关系数外,还要计算偏相关系数和复相关系数。复相关系数留待下一章讨论,这里仅讨论偏相关系数。

在对其他变量的影响进行控制的条件下,衡量多个变量中某两个变量之间的线性相关程度的指标称为偏相关系数。偏相关系数不同于前面所介绍的简单相关系数。在计算简单相关系数时,只需要掌握两个变量的观测数据,并不考虑其他变量对这两个变量可能产生的影响。而在计算偏相关系数时,需要掌握多个变量的数据,一方面考虑多个变量相互之间可能产生的影响,一方面又采用一定的方法控制其他变量,专门考察两个特定变量的净相关关系。在多变量相关的场合,由于变量之间存在错综复杂的关系,因此偏相关系数与简单相关系数在数值上可能相差很大,有时甚至符号都可能相反。简单相关系数受其他因素的影响,反映的往往是表面的非本质的联系,而偏相关系数则较能说明现象之间真实的联系。例如,一种商品的需求既受收入水平的影响又受其价格的影响。按照经济学理论,在一定的收入水平下,该商品的价格越高,商品的需求量就越小。也就是说,需求与价格之间应当是负相关。可是,在现实经济生活中,由于收入和价格常常都有不断提高的趋势,如果不考虑收入对需求的影响,仅仅利用需求和价格的时间序列数据去计算简单相关系数,就有可能得出价格越高需求越大的错误结论。

在明确偏相关系数与简单相关系数区别的基础上,我们再来讨论偏相关系数的定义公式。在偏相关中,根据固定变量数目的多少,可分为零阶偏相关、一阶偏相关、...、 $(p-1)$

阶偏相关。零阶偏相关就是简单相关。如果用下标 0 代表 Y，下标 1 代表 X1，下标 2 代表 X2，则变量 Y 与变量 X1 之间的一阶偏相关系数为：

$$r_{01 \cdot 2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1-r_{02}^2}\sqrt{1-r_{12}^2}} \quad (7-44)$$

$r_{01 \cdot 2}$ 是剔除 X2 的影响之后，Y 与 X1 之间的偏相关程度的度量； r_{01}, r_{02}, r_{12} 分别是 Y, X1, X2 两两之间的简单相关系数。设增加变量 X3，则变量 Y 与 X1 的二阶偏相关系数为：

$$r_{01 \cdot 23} = \frac{r_{01} - r_{02}r_{12} - r_{03}r_{13}}{\sqrt{1-r_{02}^2-r_{03}^2}\sqrt{1-r_{12}^2-r_{13}^2}} \quad (7-45)$$

一般地，考察多个变量时，Y 与 $X_i (i=1, 2, \dots, p)$ 的 $p-1$ 阶偏相关系数，可由 (7-44) (7-45) 和以下 (7-46) 式组成一组递推公式进行计算。

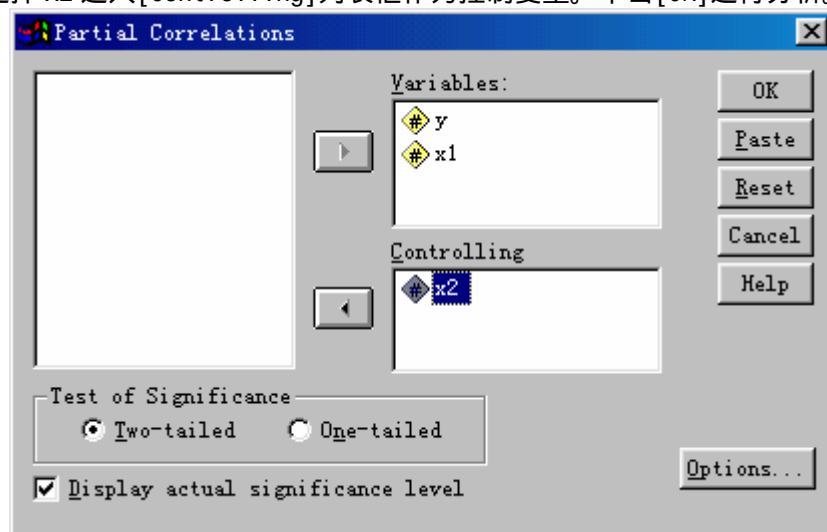
$$r_{0i \cdot 12 \dots (i-1)(i+1) \dots p} = \frac{r_{0i \cdot 12 \dots (i-1)(i+1) \dots (p-1)} - r_{0p \cdot 12 \dots (p-1)} r_{ip \cdot 12 \dots (i-1)(i+1) \dots (p-1)}}{\sqrt{1-r_{0p \cdot 12 \dots (p-1)}^2} \sqrt{1-r_{ip \cdot 12 \dots (i-1)(i+1) \dots (p-1)}^2}} \quad (7-46)$$

对偏相关系数的显著性检验与简单相关系数的显著性检验类似。读者无须记住上面的繁琐公式，只要理解其基本思想，就可以用 SPSS 可直接计算出偏相关系数大小及 p 值并进行判断。

[例 7-13] 根据 [例 7-12] 的数据，计算销售额 Y 在固定建筑产值 X2 的影响后与汽车产量 X1 的偏相关系数。

解：(1) 根据 [例 7-12] 的数据文件，选择 [Analyze] => [Correlate] => [Partial]。

(2) 在 [Partial Correlations] 主对话框中，选择 y 和 x1 进入 [Variables] 列表框作为分析变量，选择 x2 进入 [Controlling] 列表框作为控制变量。单击 [OK] 进行分析。



(3) 输出结果如下：

- - - PARTIAL CORRELATION COEFFICIENTS - - -		
Controlling for.. X2 (固定变量X2)		
	Y	X1
Y	1.0000 (0) P= .	.9482 (14) P= .000
X1	.9482 (14) P= .000	1.0000 (0) P= .

(Coefficient / (D.F.) / 2-tailed Significance)
 (偏相关系数 / (自由度) / 双侧显著性水平即 p 值)
 " . " is printed if a coefficient cannot be computed

从结果可以看出,在固定变量X2下Y与X1的偏相关系数为0.9482,在 $\alpha=0.01$ 下线性关系显著。

第三节 其它相关系数分析

前面介绍了简单相关系数和偏相关系数的计算及检验,它们是最常用的相关系数,但仅适用于度量定距变量(或定比变量)与定距变量(或定比变量)之间的线性相关程度。在实际问题中,经常要计算定类变量或定序变量的“相关系数”,这时必须选用其它合适的度量方法。本节仅简单介绍 Spearman(斯皮尔曼)等级相关系数和 Kendall(肯德尔)的 tau 相关系数。

一、Spearman 等级相关系数及其检验

(一) Spearman 等级相关系数的定义和计算

Spearman 等级相关系数适用于度量定序变量与定序变量之间的相关系数,是由统计学家斯皮尔曼(C. Spearman)首先提出的,其计算公式为:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (7-47)$$

其中, $d_i = (x_i - y_i)$, x_i 和 y_i 分别是两个变量按大小(或优劣等)排位的等级(称为秩), n 是样本的容量。

与简单相关系数类似, Spearman 等级相关系数的取值区间为: $-1 \leq r_s \leq 1$ 。 r_s 为正值时,存在正的等级相关, r_s 取负值时,存在负的等级相关。 $r_s=1$, 表明两个变量的等级完全相同,存在完全正相关。 $r_s=-1$, 表明两个变量的等级完全相反,存在完全的负相关。

(二) Spearman 等级相关系数检验

Spearman 等级相关系数是根据一定的样本计算的。两个变量的总体是否存在显著的等级相关也需要进行检验。当样本容量 n 大于 20 时,可利用以下 t 统计量,进行等级相关系数的显著性检验。

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \quad (7-48)$$

当总体等级相关系数 $\rho_s = 0$ 时,可证明 t 服从自由度为 $(n-2)$ 的 t 分布。在给定的显著水平 α 下,如按上式计算的 t 值(或者 p 值)大于临界值 $t_{\alpha/2}(n-2)$ (或 $p < \alpha$),则可以认为 ρ_s 与 0 显著差别,即两种现象(两个变量)的总体是否存在显著的等级相关。

[例 7-14] 某校对学生某专业课程的复习时间和考试成绩进行调查。抽查的 10 位同学的有关原始数据如表 7-37 第 1 栏和第 3 栏所示。要求计算复习时间与考试成绩的简单相关系数和等级相关系数。试问根据以上结果能否得出,复习时间越长考试成绩越高的结论?

解:这里分别使用手工和 SPSS 进行计算:

(1) 手工计算

首先,对复习时间和考试成绩按从少(低)到多(高)的顺序确定等级。如遇到复习时间(或考试成绩)相同的,取其应得的等级的平均数。得到的结果列在表 7-37 的第 2 栏和第 4 栏。

其次,计算简单相关系数。根据前面介绍的计算公式,利用表中的资料计算可得:简单相关系数为 0.587,其 t 检验值为 2.05(计算过程略)。查表可知:显著水平为 5%,自由度为 8 的临界值 $t_{\alpha/2} = 2.306$,上式中的 t 值小于 2.306,因此, r 不能通过显著性检验。也就是说,从原始数据看,难以判断复习时间与考试成绩之间存在显著的线性相关。

表 7-37 复习时间与考试成绩的等级相关分析

复习时间(小时)		考试成绩(分)	
原始数据T	排队等级 x_i	原始数据S	排队等级 y_i
3	3	86	3
4	4	87	4
1	1	4	1
2	2	85	2
5	5	93	6
8	6	91	5
10	8	95	8.5
9	7	94	7
11	9	95	8.5
13	10	96	10

最后，计算 Spearman 等级相关系数。将表中的有关数据代入 (7-47) 式，可得：

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 0.985$$

将以上结果代入 (7-48) 式可得：

$$t = \frac{0.985 \sqrt{10-2}}{\sqrt{1-0.985^2}} = 16.04$$

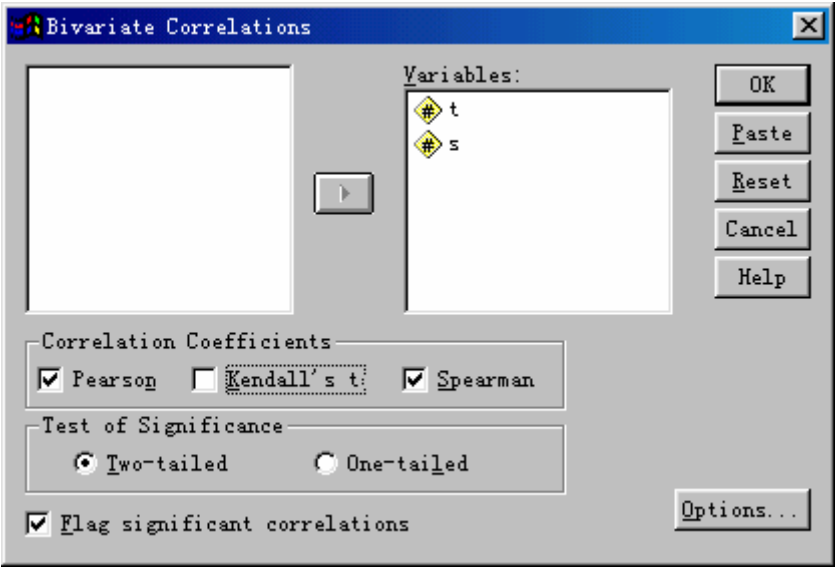
以上得到的 r_s 接近 1， t 值远大于 2.306，因此可以认为复习时间与考试成绩之间存在相当显著的等级相关关系。也就是说，仅根据简单相关分析，判断复习时间与考试成绩之间不存在显著相关关系并不妥当。从等级相关系数看，确实存在复习时间越长，考试成绩越高的现象。

(2) SPSS 计算

首先，把表 7-37 中原始数据（设复习时间为变量 T、考试成绩为变量 S）录入存为 SPSS 数据文件。

然后，选择 [Analyze] => [Correlate] => [Bivariate]，在显示的 [Bivariate Correlations] 主对话框中，把 T、S 选入 [Variables] 列表框中作为分析变量，并选择 [Correlation Coefficients] 选项中的 [Spearman]。

严格地说，本例中样本容量小于 20，不宜直接用 t 检验。但可以此作为一种参考。无须录入排队等级。



最后，单击[OK]就可得到如下结果。Correlations

		T	S
T	Pearson Correlation	1.000	.587
	Sig. (2-tailed)	.	.075
	N	10	10
S	Pearson Correlation	.587	1.000
	Sig. (2-tailed)	.075	.
	N	10	10

Correlations				
			T	S
Spearman's rho (Spearman 相关系数)	T	Correlation Coefficient	1.000	.985
		Sig. (2-tailed)	.	.000
		N	10	10
	S	Correlation Coefficient	.985	1.000
		Sig. (2-tailed)	.000	.
		N	10	10
** Correlation is significant at the .01 level (2-tailed).				

该结果除了给出的是实际显著性水平（p值）外，均与手工计算结果相同。

二、Kendall（肯德尔）的 τ （ τ ）相关系数及其检验

Kendall（肯德尔）的 τ 相关系数由统计学家Kendall提出，适用于度量两个定序变量X与Y之间的相关。共有三种形式：tau-a、tau-b和tau-c，公式分别为：

$$\tau - a = \frac{N_s - N_d}{n(n-1)/2} \tag{7-49}$$

$$\tau - b = \frac{N_s - N_d}{\sqrt{n(n-1)/2 - T_x} \sqrt{n(n-1)/2 - T_y}} \tag{7-50}$$

$$\tau - a = \frac{2m(N_s - N_d)}{n^2(m-1)} \tag{7-51}$$

其中, N_s 为X和Y的同序对的数目; N_d 为X和Y的异序对的数目; T_x 为X中同分对的数目; T_y 为Y中同分对的数目; n 为样本容量; m 为X与Y等级数较小者。所谓同序对是指变量大小顺序相同的两个样本观测值, 即其X的等级高低顺序与Y的等级顺序相同, 否则称为异序对; 所谓同分对是指等级相同的一对样本观测值, 如果样本容量为 n , 则样本观测值两两组对的话一共可以有 $n(n-1)/2$ 对。

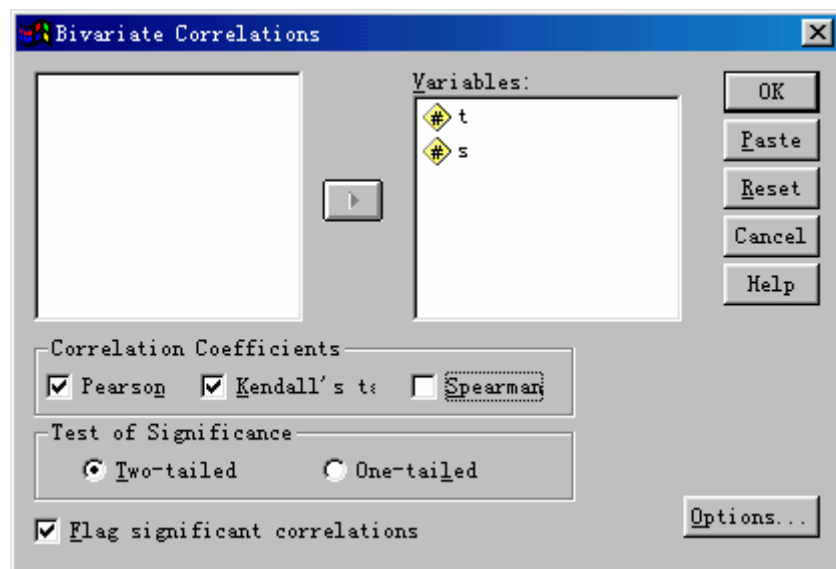
一般情况下, tau-a是在没有同分对时采用, 它表示同序对的数目与异序对的数目的差在全部可能对数中所占的比例。如果有同分对时常用tau-b和tau-c; 如果X和Y的等级数相同, 则可用tau-b, 否则用tau-c。在SPSS中采用tau-b。

[例7-15]用SPSS计算表7-37中数据的tau相关系数。

用SPSS计算表7-37中数据的tau相关系数的步骤如下:

(1) 录入数据;

(2) 选择[Analyze]>[Correlate]>[Bivariate], 在[Bivariate Correlations]主对话框中, 把变量T、S选入[Variables]列表框中, 并选择[Kendall's tau-b]。单击[OK]。



(3) 输出结果如下:

(简单相关系数表同[例7-14])

Correlations				
			T	S
Kendall's tau_b	T	Correlation Coefficient	1.000	.944**
		Sig. (2-tailed)	.	.000
		N	10	10
	S	Correlation Coefficient	.944**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10

** Correlation is significant at the .01 level (2-tailed).

可以看出, 结论与[例7-14]相同。

第八章 回归分析

第一节 一元线性回归分析

一、基本概念

在数量分析中,我们经常会看到变量与变量之间存在着一定的联系,而不只是前面所讨论的单个变量的某些孤立的特性,如均值、方差的特性等。我们要了解的是变量之间是如何发生相互影响的,这就是所谓的相关分析和回归分析。

在实际问题中,我们常常要研究两个变量之间的联系,例如:汽车生产数量 Y 与所需车轮数量 X 之间的关系,某产品的价格 X 与社会对该产品的需求 Y 之间的关系,人的身高 X 与体重 Y 之间的关系,家庭收入 X 与消费支出 Y 之间的关系等等。这些变量之间的关系可以分为两类:函数关系(确定性关系)和相关关系(随机性关系)。

如果给定解释变量 X 的值,被解释变量 Y 的值就唯一地确定了,那么 Y 与 X 的关系就是函数关系,即 $Y=F(X)$ 。例如,生产一辆汽车要配四个车轮,只要知道了汽车的生产数量 X ,所需的车轮数量 Y 也就唯一地确定了,其函数关系式为: $Y=4X$ 。

如果给定了解释变量 X 的值,被解释变量 Y 的值不是唯一的, Y 与 X 的关系就是相关关系,例如,身高与体重的关系是很密切的,但已知某人的身高 X ,我们无法确切地推断出他的体重。这是因为,身高不是决定体重的唯一因素,从而身高相同的人未必体重一样。因此身高与体重的关系就是相关关系。研究变量之间相关关系密切程度的分析叫相关分析。如果在研究变量之间的相关关系时,把其中的一些因素作为所控制的变量(自变量),而另一些随机变量作为它们的因变量,这种关系分析就称为回归分析。

应该指出的是,变量之间的函数关系和相关关系,在一定条件下是可以互相转化的。本来具有函数关系的经济变量,当存在观测误差时,其函数关系往往以相关的形式表现出来。而具有相关关系的变量之间的联系,如果我们对它们有了深刻的规律性认识,并且能够把影响因变量变动的因素全部纳入方程,这时的相关关系也可能转化为函数关系。另外,相关关系也具有某种变动规律性,所以,相关关系经常可以用一定的函数形式去近似地描述。经济现象的函数关系可以用数学分析的方法去研究,而研究社会经济现象的相关关系必须借助于统计学中的相关与回归分析方法。

为了具体说明,考虑家庭月可支配收入如何影响消费支出。如果把不同的可支配收入 X (千元)对应的消费支出 Y (千元)画在平面图上,那么可以得到如图 8-72 的散点图。

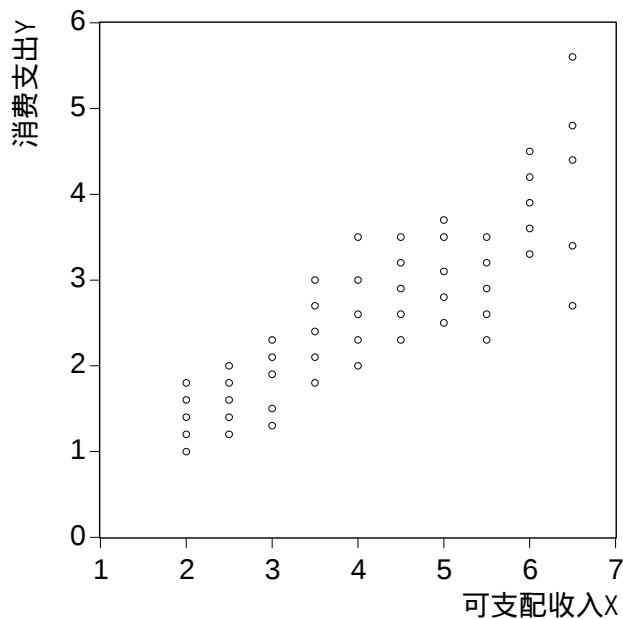


图 8-72 家庭月消费支出与可支配收入的关系

从该散点图似乎可以看到可支配收入确实对消费支出有影响。也应该可能通过拟合一条穿过这一散点图的直线或曲线来描述可支配收入 X 是如何影响消费支出 Y 的。这里的消费支出 Y 取决于可支配收入，作为因变量（或被解释变量、响应变量），可支配收入 X 不依赖于消费，作为自变量（或解释变量、独立变量、预测因子、回归子等）。

[例-16]假设由于条件限制，我们只能进行 10 次观测，观测值如所示。

表 8-38 消费支出与可支配收入的观测值

消费支出 Y (千元)	可支配收入 X (千元)
1.6	2.0
2.0	2.5
2.3	3.0
2.4	3.5
3.0	4.0
3.2	4.5
3.1	5.0
3.5	5.5
3.6	6.0
4.4	6.5

根据观测值，在平面上描出 10 个点，并找出一条拟合这些点的直线，如图 8-73 所示。为了拟合这样一条直线，需要某种准则。准则不同，拟合的方法也就不同，拟合出来的直线就不一样。最常用的准则是最小二乘准则。

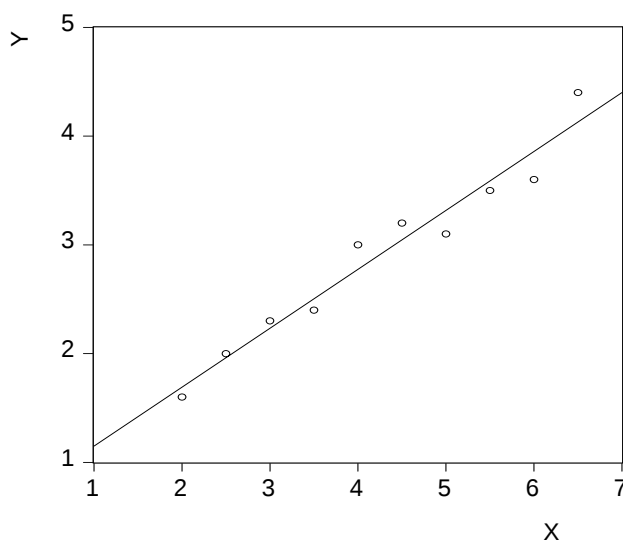


图 8-73 观测值的散点图及其拟合直线

二、用最小二乘法拟合回归直线

我们的目标是从代数上对数据拟合一条直线，直线方程的形式为：

$$\hat{Y} = b_0 + b_1 X \quad (8-52)$$

其中， $\hat{Y}$ 表示当 X 取某个值时 Y 的预测值，即拟合直线上对应的高度。为此，我们要找到计算 b_0 （截距）和 b_1 （斜率）的公式。在拟合这条直线时，一个合理的准则就是使图 8-73 中的所有观测点与直线的垂直距离 $e = Y - \hat{Y}$ （称为残差 Residual）都尽可能地小，即让所有的观测点与直线的垂直距离之和 $\sum e$ 为最小。不过由于有些观测点在该直线之上，有些观测点在直线之下，因此有些 e 是正的，有些是负的。相加后正负抵销，有可能总和 $\sum e$ 很小但是个别的 e 还是很大。为了克服这个问题，我们先将 e 平方使它们都变成正的，然后再求和并使之变成最小，这就是所谓的“普通最小二乘法（OLS——Ordinary Least Squares）准则”：选择 b_0 和 b_1 使得 $\sum e^2 = \sum (Y - \hat{Y})^2$ 为最小。根据这条准则选择出来的一条最佳拟合直线，叫做最小二乘回归直线。

由于残差平方和 $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1 X)]^2$ 是 b_0 和 b_1 的二次函数，

并且是非负和连续可微的。根据微积分中求极小值的原理，可知残差平方和存在极小值，同时欲使它达到最小，残差平方和 $\sum e^2$ 对 b_0 和 b_1 的偏导数必须等于零。从而可推导出 b_1 和 b_0 的公式（推导过程略）：

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \quad (8-53)$$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (8-54)$$

对于[例-16]中的数据，根据上面的公式求得 $b_0 = 0.607$ 和 $b_1 = 0.542$ 。将它们代入式(8-52)，即得到了最小二乘回归直线的方程为

$$\hat{Y} = 0.607 + 0.542X$$

这一直线在图 8-73 中给出。根据定义，直线的斜率等于沿着 X 方向移动一个单位时高度 Y 的变化量，即斜率 b_1 表示当 X 变化一个单位时，Y 平均变化 b_1 个单位。这里 $b_1 = 0.542$ ，表示家庭可支配收入增加（或减少）1000 元时，消费支出将平均增加（或减少）542 元。

在 SPSS 中进行一元线性回归方程估计的操作步骤为：

- (1) 建立数据文件，定义“消费支出”变量为 Y，定义“可支配收入”变量为 X，并录入相应数据；
- (2) 选择主菜单[Analyze]⇒[Regression]⇒[Linear]（如图 8-74 所示），打开[Linear Regression]主对话框（如图 8-75 所示）。在左边列表框中选定变量 Y，单击  按钮，使之进入[Dependent]框，选定变量 X，单击按钮使之进入[Independent(s)]框。

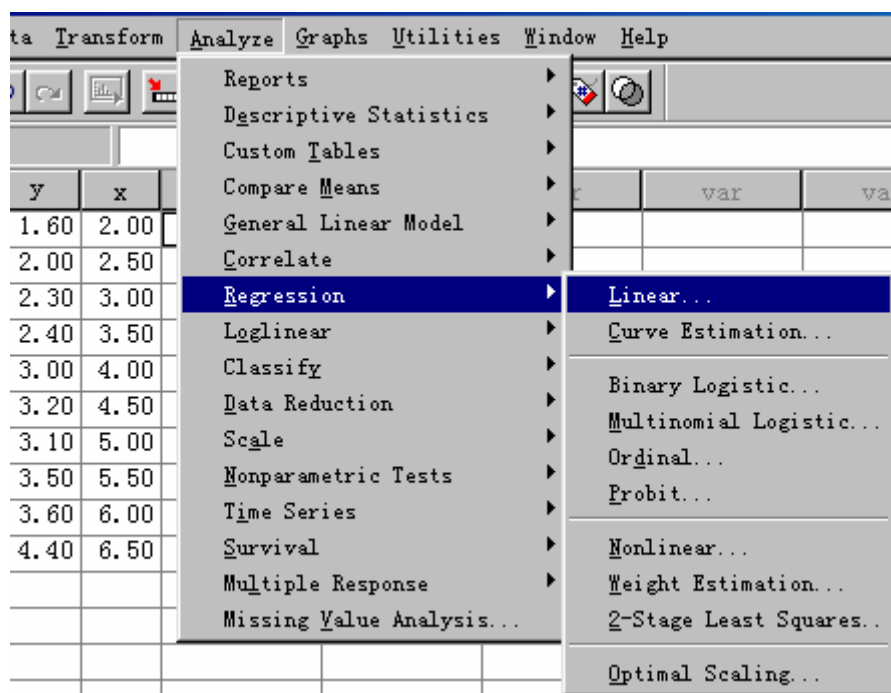


图 8-74 回归分析菜单

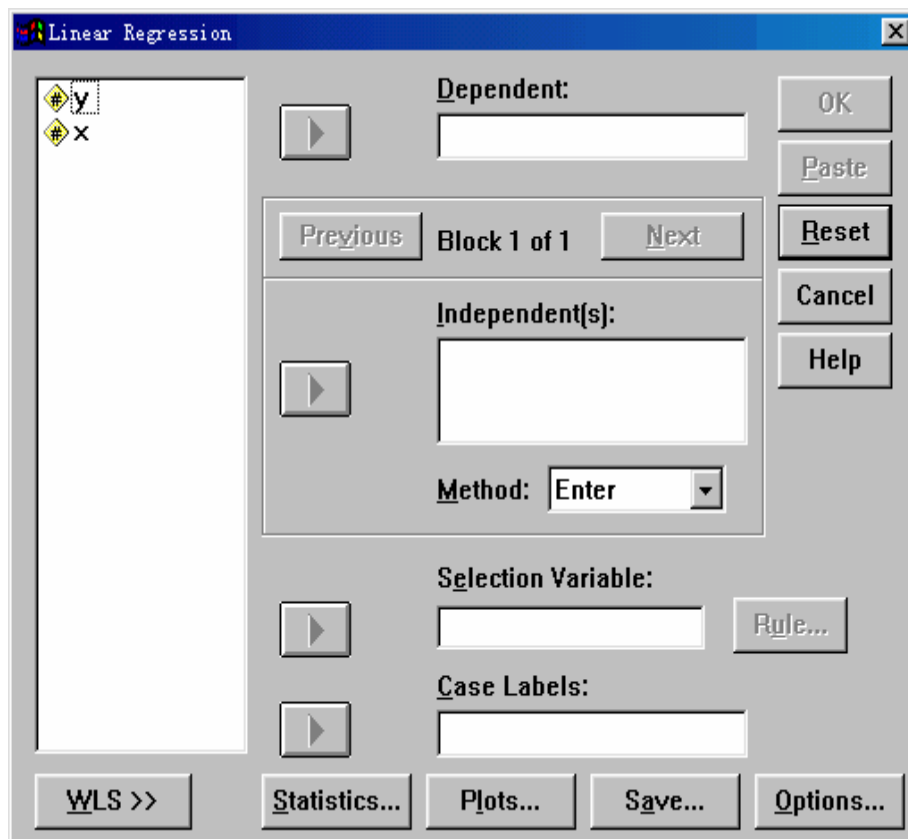


图 8-75 回归分析主对话框

(3) 单击[OK]按钮，得到如下结果：

表 8-39 一元线性回归分析结果输出

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977 ^a	.954	.948	.1918

a Predictors: (Constant), X

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.055	1	6.055	164.655	.000
	Residual	.294	8	3.677E-02		
	Total	6.349	9			

a Predictors: (Constant), X

b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.607	.189		3.206	.013
	X	.542	.042	.977	12.832	.000

a Dependent Variable: Y

输出结果中的[Unstandardized Coefficients]指未标准化的系数估计值 (B) 及其标准误 (Std. Error)。可以看出, 系数估计值分别为 $b_0 = 0.607$ 和 $b_1 = 0.542$, 与手工计算相同。输出结果的其它指标留待后面解释。

三、标准线性回归模型

上面我们对样本观测点的处理仅仅是用一条直线去拟合, 得到的方程 $\hat{Y} = b_0 + b_1 X$ 称为样本回归方程。如果希望对抽取这个样本的总体进行推断, 那么必须建立数学模型, 以便构造置信区间和进行假设检验。

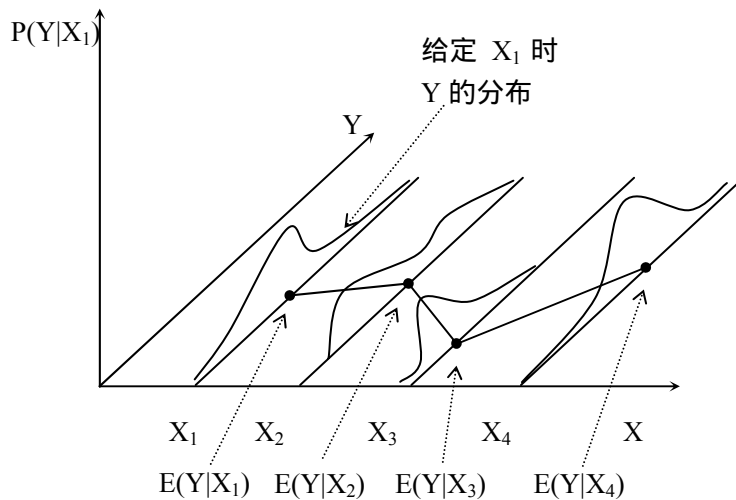


图 8-76 一般情况下的总体回归模型

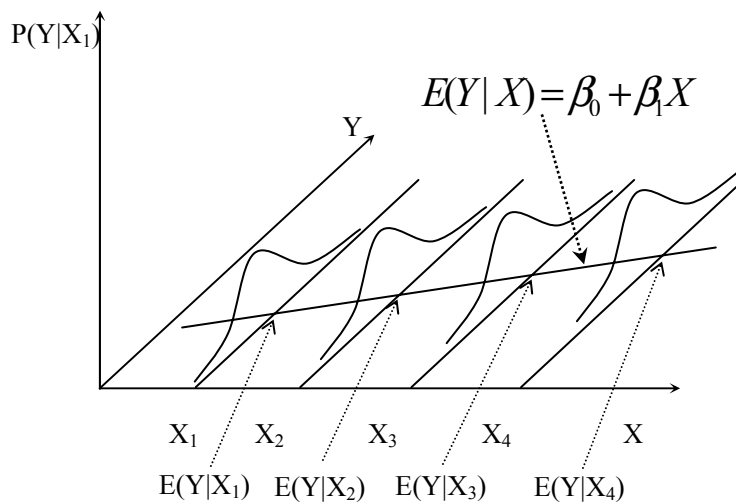


图 8-77 假定条件下的总体回归模型

在图 8-76 中, 收入水平为 X_1 的某一家庭的消费支出 Y 不会相同, 因为不同家庭的消费结构不一样。在同等收入情况下, 有的家庭会多消费一些, 有的少一些。这样就可以得到消费支出 Y 的一个分布 (或总体), 称为给定 (固定) X_1 时 Y_1 的分布, 记作 $P(Y_1|X_1)$ 。类似的, 在上 X_2 上也会有一个 Y_2 上的分布, 等等。因此, 我们可以直观地看到如图 8-76 中所示的全部 Y 总体的一个集合。很明显, 分析这些各个特殊的总体会有很大的困难。因此, 为

了便于处理这个问题，我们对这些 Y 分布的规律性作一些假定，所图 8-77 所示。我们的基本假定为：

1) 所有的 Y 的分布的均值都正好在一条直线上，称之为总体的（真实的）回归直线：

$$E(Y|X) = \beta_0 + \beta_1 X \quad (8-55)$$

总体参数 β_0 和 β_1 确定了该直线，它们是要通过样本信息来估计的。

2) 所有的 Y 分布都有同样的形状。这意味着对所有的 $X_i (i=1,2,\dots,n)$ ，概率分布 $P(Y_i|X_i)$ 都有着相同的方差 σ^2 ，即 $Var(Y_i|X_i) = \sigma^2, i=1,2,\dots,n$ 。

3) 随机变量 Y 是相互独立的。也就是说，Y2 和 Y1 没有统计关系，Y3 和 Y4 也没有统计关系，等等。

4) 给定 X 时 Y 分布的形状是正态的，即 Y 服从正态分布。

我们把符合以上假定的回归模型称为标准（古典）的回归模型。

通常把 Y_i 对其均值的离差记为扰动项 ε_i ，因而标准回归模型可写为：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (8-56)$$

其中， ε_i 为相互独立的正态变量，它们的均值为 0，方差为 σ^2 。

从这里可以看出，随机变量 Y 由两部分组成，一部分是其均值部分： $E(Y) = \beta_0 + \beta_1 X$ ，

另一部分是随机部分：扰动项 ε 。扰动项产生的原因主要有以下两个方面：

- 1、客观现象的随机性质。人的行为的随机性，社会环境与自然环境影响的随机性决定了回归模型中必须引入扰动项。
- 2、测量误差。在收集、整理数据时，总要产生某些主观或客观上的测量误差、登记误差，致使有些变量的观测值并不精确等实际值。

四、样本回归模型

以上我们给出了一元线性回归模型的总体回归方程。在图 8-78 中，直线 $E(Y|X) = \beta_0 + \beta_1 X$ 表示真实的总体回归直线。根据样本观测值，采用最小二乘法，得到了一条估计的样本回归直线 $\hat{Y} = b_0 + b_1 X$ 。在实际问题中，由于所要研究的现象的总体单位数一般是很多的，在许多场合甚至是无限的，因此无法掌握因变量 Y 总体的全部取值。也就是说，总体回归方程事实上是未知的，需要利用样本的信息对其进行估计。

根据样本数据拟合的直线，称为样本回归直线。显然，样本回归直线的函数形式应与总体回归线的函数形式一致。一元线性回归模型的样本回归直线可表示为：

$$\hat{Y} = b_0 + b_1 X \quad (8-57)$$

式中的 $\hat{Y}$ 是样本回归直线上与 X 相对应的 Y 值，可视为 $E(Y)$ 的估计； b_0 是样本回归方程的截距， b_1 是样本回归方程的斜率，它们分别是对总体回归参数 β_0 和 β_1 的估计。

实际观测到的因变量 Y 值，并不完全等于 $\hat{Y}$ ，其二者之差为 $e = Y - \hat{Y}$ 则有：

$$Y = b_0 + b_1 X + e$$

上式称为样本回归模型。式中 e 称为残差，在概念上， e 与总体误差项 ε 相互对应。

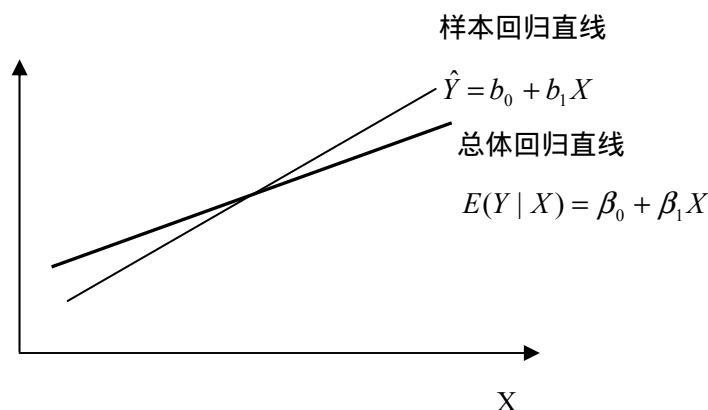


图 8-78 真实的总体回归直线与估计的样本回归直线

样本回归直线与总体回归直线之间的联系显而易见。这里需要特别指出的是它们之间的区别。第一、总体回归直线是未知的，它只有一条。而样本回归直线则是根据样本数据拟合的，每抽取一组样本，便可以拟合一条样本回归直线。第二、总体回归直线中的 β_0 和 β_1 是未知的参数，表现为常数。而样本回归函数中的 b_0 和 b_1 是随机变量，其具体数值随所抽取的样本观测值不同而变动。第三、总体回归直线中的 ε 是 Y 与未知的总体回归直线之间的纵向距离，它是不可直接观测的。而样本回归函数中的 e 是 Y 与样本回归直线之间的纵向距离，当根据样本观测值拟合出样本回归直线之后，可以计算出 e 的具体数值。

综上所述，样本回归直线是对总体回归直线的近似反映。回归分析的主要任务就是要采用适当的方法，充分利用样本所提供的信息，使得样本回归函数尽可能地接近于真实的总体回归函数。无论如何，所估计的样本回归直线都不可能与真实的总体回归直线完全一致。

五、“回归”名称的产生背景

回归分析的基本思想和方法以及“回归(Regression)”名称的由来归功于英国统计学家 F·Galton (1822—1911 年)。F·Galton 和他的学生、现代统计学的奠基者之一 K·Pearson(1856—1936 年)在研究父母身高与其子女身高的遗传问题时，观察了 1078 对夫妇，以每对夫妇的平均身高作为解释变量 X ，而取他们的一个成年儿子的身高作为被解释变量 Y ，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为 $\hat{Y} = 33.73 + 0.516X$ 。这种趋势及回归方程表明父母身高 X 每增加一个单位时，其成年儿子的身高 Y 也平均增加 0.516 个单位。这个结果表明，虽然高个子父辈有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。平均来说，一群高

参见何晓群：《现代统计分析方法与应用》，中国人民大学出版社；彼得·伯恩斯坦著，毛二万、张顺明译：《与天为敌——风险探索传奇》，清华大学出版社，1999 年。

个子父辈的儿子们的平均身高要低于他们父辈的平均身高，他们儿子的身高没有比他们更高，高个子父辈偏离其父辈平均身高的一部分被其子代拉回来了，即子代的平均高度向中心回归了。但是，低个子父辈的儿子们虽然仍为低个子，平均身高却比他们的父辈增加了，即父辈偏离中心的部分在子代被拉回来一些。就是说，子代的平均高度没有比他们的父辈更低。正是因为子代的身高有回到父辈平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。

F·Galton 用他最有说服力且最风趣的一段话来概括了这个结论：

孩子的遗传一部分来自父母，一部分来自祖先。家谱向前推得越远，其祖先越多样越不同，直到他们成为从一个大种族随机抽取的多样性的样本为止。这个规律解决了为何天才无法全部遗传给其后代的问题……这个规律是公正的；无论好的方面还是坏的方面的遗传都会打相同的折扣。如果它使一些有天赋的父母期待其子女也很有天赋的愿望化为泡影，那么它同样也会使另一些父母减少担心，因为他们的子女同样也不会全部继承他们的缺陷和疾病。

这生动地说明了生物学中“种”的概念的稳定性。正是为了描述这种有趣的现象，F·Galton 引进了“回归(regression)”这个词来描述父辈身高 X 与子代身高 Y 的关系。尽管“回归”这个名称的由来具有其特定的含义，人们在研究大量的问题中变量 X 与 Y 之间的关系并不具有这种“回归”的含义，但借用这个词把研究变量 X 与 Y 之间的统计关系的数学方法称为“回归分析”，也算是对 F·Galton 这个伟大的统计学家的一种纪念。

第二节 一元线性回归模型估计量的性质与分布

一、最小二乘估计量的性质

最小二乘法是多种估计方法中的一种。按照最小二乘法求得的估计总体回归参数的公式 (8-53) 和 (8-54) 是样本观测值的函数，通常称之为最小二乘估计量。最小二乘估计量的形式是不变的，但根据所选取的样本不同，其具体数值即回归参数的估计值却会随之变化，因此，它是一种随机变量。可以证明，在标准假定能够得到满足的条件下，回归参数的最小二乘估计量是因变量观测值 Y 的线性函数，而且均值等于其真值，即有：

$$E(b_0) = \beta_0, E(b_1) = \beta_1 \quad (8-58)$$

其方差为：

$$Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X - \bar{X})^2} \right) \quad (8-59)$$

$$Var(b_1) = \frac{\sigma^2}{\sum (X - \bar{X})^2} \quad (8-60)$$

其中 n 为样本容量， σ^2 为 Y 的方差、随动项的方差。

最小二乘估计量是因变量观测值 Y 的线性函数，其均值等于总体回归参数的真值。因此，最小二乘估计量是总体回归参数的线性无偏估计量。还可以证明，在所有的线性无偏估计量中，回归参数的最小二乘估计量的方差最小；同时随着样本容量 n 的增大，其方差会不断缩

小。也就是说,回归参数的最小二乘估计量是最佳线性无偏估计量,即 BLUE 估计量(the Best Linear Unbiased Estimator)和一致估计量。

标准线性回归模型中,回归参数的最小二乘估计量所具有的上述性质。通俗地讲,这一性质表明,在标准的假定条件下,最小二乘估计量是一种最佳的估计方式。但是应当明确,这并不意味着根据这一方式计算的每一个具体的估计值都比根据其他方式计算的具体估计值更接近真值,而只是表明如果反复多次进行估计值计算或是扩大样本的容量进行估计值计算,按最佳估计方式计算的估计值接近真值的可能性(概率)最大。

二、最小二乘估计量的抽样分布

b_0 和 b_1 是 Y 的线性组合函数。因此, b_0 和 b_1 的分布取决于 Y 的分布。由于正态随机变量的线性组合仍服从正态分布,因此 b_0 和 b_1 服从正态分布,其分布密度由其均值和方差唯一决定。由式(8-58)、式(8-59)和式(8-60),可知 b_0 和 b_1 的抽样分布分别为:

$$b_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}}{\sum (X - \bar{X})^2} \right)\right) \quad (8-61)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (X - \bar{X})^2}\right) \quad (8-62)$$

在式(8-59)、(8-60)、(8-61)和(8-62)中关于参数 b_0 和 b_1 的方差的表达式中,均含有随机扰动项方差 $\sigma^2 = \text{Var}(\varepsilon) = \text{Var}(Y)$ 。 σ^2 又称为总体方差,是未知的,因而 b_0 和 b_1 实际上无法计算。由于随机扰动项 ε 无法观测,只能从残差 e_i 出发,对总体方差 σ^2 进行估计。

可以证明总体方差 σ^2 的估计量为:

$$S^2 = \frac{\sum e_i^2}{n-2} \quad (8-63)$$

因此, b_0 和 b_1 的方差的估计量分别是:

$$S^2(b_0) = S^2 \left(\frac{1}{n} + \frac{\bar{X}}{\sum (X - \bar{X})^2} \right) \quad (8-64)$$

$$S^2(b_1) = \frac{S^2}{\sum (X - \bar{X})^2} \quad (8-65)$$

在表 8-39 中, [Unstandardized Coefficients Std. Error]指系数的标准误差:

$$S(b_0) = \sqrt{S^2(b_0)} = 0.189, \quad S(b_1) = \sqrt{S^2(b_1)} = 0.042。$$

以 b_1 的抽样分布为例, b_1 是 β_1 的无偏估计,其分布中心(均值)为 β_1 ,其标准误差

$\frac{S}{\sqrt{\sum (X - \bar{X})^2}}$ 用来衡量估计量 b_1 接近真值 β_1 的程度, 判定估计量 b_1 的可靠性。

因此可以看出, 要想使 b_0 和 b_1 更稳定, 在收集数据时, 就应该考虑 X 的取值尽可能分散一些; 样本容量也应尽可能大一些, 样本量太小时, 估计量的稳定性肯定不会很好。

第三节 一元线性回归模型的检验

一、一元线性回归模型检验的种类

根据变量 X 和 Y 的样本观测值, 应用最小二乘法求得了样本回归直线, 作为总体回归直线的近似, 这种近似是否合理, 必须对其进行检验。如果通过检验发现模型有缺陷, 则必须回到重新设定模型或估计参数。一元线性回归模型的检验包括经济意义检验、统计检验和计量检验。

(一) 经济意义检验

经济意义检验主要涉及参数估计值的符号和取值范围, 如果它们与经济理论以及人们的实践经验不相符, 就说明模型不能很好地解释现实的经济现象。例如, 前面的家庭消费支出与可支配收入例子中, β_2 的取值范围应在 0 和 1 之间, 如果估计出来的 b_2 小于 0 或大于 1, 则不能通过经济意义检验。在对实际的经济现象进行回归分析时, 常常会遇到经济意义检验不能通过的情况。造成这一结果的主要原因是: 经济现象的统计数据无法象自然科学中的统计数据那样通过有控制的实验去取得, 因而所观测的样本容量有可能偏小, 不具有足够的代表性, 或者不能满足标准线性回归分析所要求的假定条件。

(二) 统计检验

统计检验是利用统计学中的抽样理论来检验样本回归方程的可靠性, 具体又可分为拟合程度检验、相关系数检验、参数显著性检验 (t 检验) 和回归方程显著性检验 (F 检验), 是对所有现象进行回归分析时都必须通过的检验。

(三) 计量检验

计量检验是对标准线性回归模型的假定条件是否满足进行检验, 具体包括序列相关检验、异方差性检验等。计量检验对于经济现象的定量分析具有特别重要的意义。

二、拟合优度检验

所谓拟合程度, 是指样本观测值聚集在样本回归直线周围的紧密程度。判断回归模型拟合程度优劣最常用的数量指标是判定系数 (Coefficient of Determination)。该指标是建立在对总离差平方和进行分解的基础之上的。

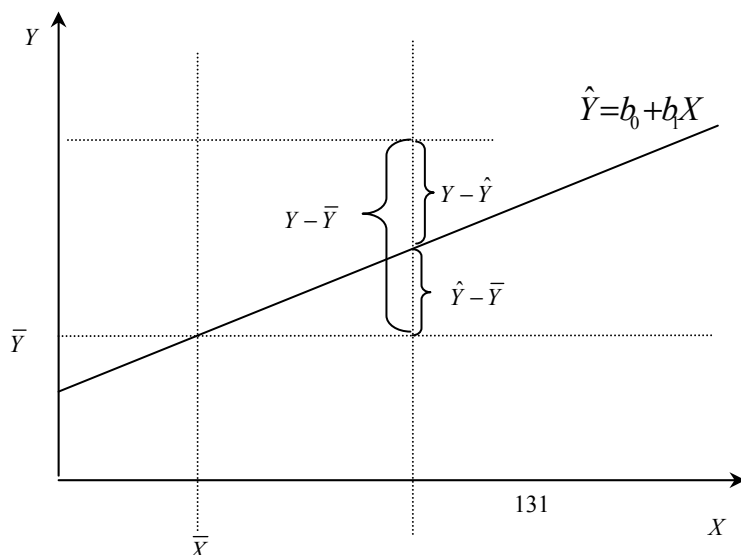


图 8-79 总离差的分解

如图 8-79 所示, 因变量的实际观测值与其样本均值的离差即总离差 ($Y - \bar{Y}$) 可以分解为两部分: 一部分是因变量的理论回归值与其样本均值的离差 ($\hat{Y} - \bar{Y}$), 它可以看成是能够由回归直线解释的部分, 称为可解释离差; 另一部分是实际观测值与理论回归值的离差 ($Y - \hat{Y}$), 它是不能由回归直线加以解释的残差 e 。对任一实际观测值 Y 总有:

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y}) \quad (8-66)$$

对 (8-66) 式两边平方并求和并计算, 可得到:

$$\sum(Y - \bar{Y})^2 = \sum(Y - \hat{Y})^2 + \sum(\hat{Y} - \bar{Y})^2 \quad (8-67)$$

$$TSS = RSS + ESS$$

上式中, $TSS = \sum(Y - \bar{Y})^2$ 是总离差平方和(Total Sum of Squares); $RSS = \sum(Y - \hat{Y})^2$ 是用回归直线无法解释的离差平方和, 称为残差平方和(Residual Sum of Squares); $ESS = \sum(\hat{Y} - \bar{Y})^2$ 是由回归直线可以解释的那一部分离差平方和, 称为回归平方和(Explained Sum of Squares)。式 (8-67) 的两边同除以 TSS, 得:

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS} \quad (8-68)$$

显而易见, 各个样本观测点与样本回归直线靠得越紧, ESS 在 TSS 中所占的比例就越大。因此, 可定义这一比例为判定系数, 即有:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (8-69)$$

判定系数是对回归模型拟合程度的综合度量, 判定系数越大, 模型拟合程度越高。判定系数越小, 则模型对样本的拟合程度越差。

判定系数 R^2 具有如下性质:

1、判定系数 R^2 具有非负性。

由判定系数的定义式可知, R^2 的分子分母均是不可能为负值的平方和, 因此其比值必大于零。

2、判定系数的取值范围为 $0 \leq R^2 \leq 1$ 。

由 R^2 的计算公式可以看出: 当所有的观测值都位于回归直线上时, $RSS = 0$, 这时 $R^2 = 1$, 说明总离差可以完全由所估计的样本回归直线来解释; 当观测值并不是全部位于回归直线上时, $RSS > 0$, 则 $RSS / TSS > 0$, 这时 $R^2 < 1$; 当回归直线没有解释任何离差, 即模型中解释变量 X 与因变量 Y 完全无关时, Y 的总离差全部归于残差平方和, 即 $RSS = TSS$, 这时 $R^2 = 0$ 。

3、判定系数是样本观测值的函数, 它也是一个统计量。

例如 根据[例-16]的数据, 可计算出 $R^2 = 0.954$ (见表 8-39 Model Summary 中的 R Square 栏目), 说明消费支出的变动中有 95.4% 可以由可支配收入来解释。

三、相关系数的显著性检验

经济变量之间通常是相关的。问题是相关程度如何, 如果在相关程度过低的变量之间建立回归模型, 就没有很大的意义。这里讨论两个变量之间的线性相关, 称为简单相关。正如上一章所讨论的, 两个变量 X 和 Y 之间真实的线性相关程度用总体相关系数 ρ 来表示, 即

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

由于总体未知， ρ 无法计算，我们利用相本相关系数

$$r = \frac{S_{XY}}{S_X S_Y}$$

作为 ρ 的估计。

其中， $S_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}$ 称为 X 与 Y 的样本协方差，是 $\text{Cov}(X, Y)$ 的无偏估计；

$S_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$ 称为 X 的样本标准差，它的平方即 X 的样本方差 S_X^2 是 $\text{Var}(X)$ 的无偏估计；

$S_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n-1}}$ 称为 Y 的样本标准差，它的平方即 Y 的样本方差 S_Y^2 是 $\text{Var}(Y)$ 的无偏估计。因此，样本相关系数 r 是总体相关系数 ρ 的无偏估计。

相关系数 r 具有以下性质：

(1) 相关系数 r 的取值范围在 -1 至 +1 之间，当 $r > 0$ 时，称 X 与 Y 正相关；当 $r < 0$ 时，称 X 与 Y 负相关；当 $r = 0$ 时，称 X 与 Y 不（线性）相关。当 $|r|$ 接近于 1 时，相关程度越高。

(2) $r = \pm \sqrt{R^2}$ ，但两者的概念不同。判定系数 R^2 是对变量 Y 和 X 作回归分析得出的，用于衡量拟合优度，而相关系数 r 是对变量 X 和 Y 作相关分析得出的，用以判定 X 与 Y 的线性相关程度。

(3) 相关系数 r 接近于 1 的程度与样本容量 n 有关：当 n 较小时，相关系数的绝对值容易接近于 1；当 n 较大时，相关系数的绝对值容易偏小。特别是当 $n=2$ 时，相关系数的绝对值总为 1。因此仅凭相关系数较大就说变量 X 与 Y 之间有密切的线性关系，则显得匆忙。

样本相关系数在统计上是否显著，即总体 X 与 Y 是否显著相关，必须进行相关系数的显著性检验。相关系数的显著性检验可采用上一章的检验方法，也可以按如下步骤进行检验：

(1) 计算样本相关系数 r ；

(2) 根据给定的显著性水平 α 和样本容量 n ，查相关系数表（附表）得到临界值 r_{α} 。

(3) 若 $|r| > r_{\alpha}$ ，则 X 与 Y 有显著的线性关系，否则 X 与 Y 的线性相关关系不显著。

例如，根据[例-16]的数据计算得出 $r=0.977$ ，给定 $\alpha=0.05$ ， $n=10$ ，查表得 $r_{0.05}=0.632$ 。由于 $r > r_{0.05}$ ，所以家庭可支配收入与消费支出存在显著的正线性关系。

应该注意的是，这里的 $r=0.977$ 刚好等于表 8-39 Model Summary 中的 R （复相关系数），因为在一元线性回归模型中指标 R 的计算公式与样本相关系数 r 的计算公式相同，但在多元线性回归模型中这些指标有着不同的含义。

四、回归参数的显著性检验（t 检验）

所谓回归参数的显著性检验，就是根据样本估计的结果对总体回归参数的有关假设进行检验。为了进行回归参数的显著性检验，首先有必要了解 b_0 和 b_1 的抽样分布。在总体方差已知的情况下，根据式（8-61）（8-62）可以按照第五章中所介绍的 Z 检验方法去对总体回归参数进行假设检验。可是，一般来说，总体方差是未知的，要用其无偏估计量 S^2 去代替。我们用式（8-64）和（8-65）分别作为 b_0 和 b_1 的方差估计值，可以证明，当样本为小样本时，回归参数估计值的标准化变换变量并不遵循正态分布规律，而是服从自由度为 $n-2$ 的 t 分布，即

$$t = \frac{b_1 - \beta_1}{S(b_1)} \sim t(n-2)$$

式中的 n 为样本容量， $n-2$ 为自由度。

利用以上结论可以对回归参数进行显著性检验。 β_0 与 β_1 的检验方法是相同的，但 β_1 的检验更为重要，因为它表明自变量对因变量线性影响的程度。下面我们以前 β_1 的检验为例，

介绍回归参数显著性检验的基本步骤：

1、提出假设。

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

式中， H_0 表示零假设； H_1 表示备择假设。如果零假设成立，则说明X对Y没有显著的影响，反之X对Y具有显著的影响。

2、计算回归参数的t统计量值。

$$t = \frac{b_1 - \beta_1}{S(b_1)} = \frac{b_1 - 0}{S(b_1)} = \frac{b_1}{S(b_1)}$$

3、根据给定的显著水平 α 确定临界值，或者计算t值所对应的p值。

t检验的临界值是由显著水平 α 和自由度决定的。应该注意的是，这里进行的检验是双侧检验，所以临界值为 $t_{\frac{\alpha}{2}}(n-2)$ 。

4、做出判断。

如果t的绝对值大于临界值（或者 $p < \alpha$ ），就拒绝原假设，接受备择假设，说明X对Y具有显著的影响作用；反之，如果t的绝对值小于临界值的绝对值（或者 $p > \alpha$ ），则接受原假设，说明X对Y没有显著的影响。

例如 根据[例-16]的数据 对家庭消费支出与可支配收入的回归参数 β_1 进行显著性检验。

根据表 8-39 [Coefficients]中的T和Sig栏可以看出， $t=12.832$ ， $p=0.000$ （Sig表示实际显著性水平，即p值），所以拒绝回归参数 $\beta_1=0$ 的假设，可支配收入X对消费支出Y具有显著的影响。

五、回归方程的显著性检验（F检验）

F检验是对回归总体线性关系是否显著的一种假设检验。根据式（8-67）有 $TSS=ESS+RSS$ 。如果比例值 ESS/RSS 较大，说明X对Y的解释程度高，可以认为总体存在线性关系，反之总体可能不存在线性关系。做利用这个值 ESS/RSS 进行推断。由于对不同的样本，这个比值可能不同，因此对给定的样本，利用这个比值进行推断，必须在统计假设检验的基础上进行。可以证明，在一元线性回归条件下， ESS 和 RSS 分别服从自由度为1和 $n-2$ 的 χ^2 分布，即 $ESS \sim \chi^2(1)$ ， $RSS \sim \chi^2(n-2)$ 。因此构造统计量

$$F = \frac{ESS/1}{RSS/(n-2)}$$

分子称为回归均方(差)，分母称为残差均方(差)。由抽样分布理论，统计量F服从第一自由度为1、第二自由度为 $n-2$ 的F分布。即

$$F = \frac{ESS/1}{RSS/(n-2)} \sim F(1, n-2)$$

利用F统计量进行回归方程显著性检验的步骤如下：

1、提出假设。

$$H_0: \beta_0 = \beta_1 = 0$$

$$H_1: \beta_0, \beta_1 \text{不全为} 0$$

式中，如果零假设 H_0 成立，则说明回归总体是显著线性的，反之表明回归总体不存在线性关系，即所有解释变量对Y没有显著的线性作用。

2、计算回归方程的F统计量值。

$$F = \frac{ESS/1}{RSS/(n-2)}$$

3、根据给定的显著水平 α 确定临界值 $F_{\alpha}(1, n-2)$ ，或者计算F值所对应的p值。

4、做出判断。

如果F值大于临界值 $F_{\alpha}(1, n-2)$ （或者 $p < \alpha$ ），就拒绝原假设，接受备择假设；反之，

如果 F 值小于临界值 $F_{\alpha}(1, n-2)$ (或者 $p > \alpha$) , 则接受原假设。

例如 根据[例-16]的数据 对家庭消费支出与可支配收入的回归参数 β_1 进行显著性检验。

根据表 8-39SPSS 输出结果的 ANOVA(方差分析表)可得到 $F=164.655$, $p=0.000 < 0.05$, 所以拒绝零假设, 回归方程的线性关系是显著的。

在一元线性回归模型中, 由于只有一个解释变量 X , 对 $\beta_1 = 0$ 的 t 检验与对整个回归方程的 F 检验是等价的, 可以证明 F 统计量与 t 统计量具有如下关系:

$$F=t^2$$

但在一般的多元回归条件下, 两种检验要说明的问题不同、作用不同, 不能相互替代。

第四节 多元线性回归基本概念

一、多元线性回归模型

前面介绍的一元线性回归分析所反映的是一个因变量与一个自变量之间的关系。但是, 在实际的经济活动中, 某一现象的变动常受多种现象变动的影 响。例如, 家庭消费支出除了受可支配收入水平的影响外, 还会受以往消费和收入水平的影响; 汽车的需求量除了受到人们的收入水平的影响外, 还会受到汽车价格水平的影响。这就是说, 影响因变量的自变量通常不是一个, 而是多个。在许多场合, 仅仅考虑单个变量是不够的, 还需要就一个因变量与多个自变量的联系来进行考察, 才能获得比较满意的结果。这就产生了测定多因素之间相关关系的问题。

研究在线性相关条件下, 两个和两个以上自变量对一个因变量的数量变化关系, 称为多元线性回归分析, 表现这一数量关系的数学公式, 称为多元线性回归模型。多元线性回归模型是一元线性回归模型的扩展, 其基本原理与一元线性回归模型相类似, 只是在计算上比较麻烦一些而已。

假定因变量 Y 与 p 个自变量 $X_1, X_2, \dots, X_p$ 之间的回归关系可以用线性函数来近似反映。多元线性总体回归模型的一般形式如下:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (1)$$

其中, ε 是随机扰动项; $\beta_1, \beta_2, \dots, \beta_p$ 是总体回归参数。 β_j 表示在其他自变量保持不变的情况下, 自变量 X_j 变动一个单位所引起的因变量 Y 平均变动的单位数, 因而又叫做偏回归参数。

固定解释变量的每一组观察值时, 因变量 Y 的值是随机的, 其可能值的集合形成一个总体, 记为 $E(Y|X_1, X_2, \dots, X_p)$, 从而因变量 Y 的条件期望函数为

$$E(Y|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

该式称为多元线性总体回归方程。同一元总体回归方程, 总体回归参数是未知的, 必须利用有关的样本观测值来进行估计。

特别地, 对于只有两个自变量的二元线性回归方程为:

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

从几何上看, 这个方程表示的是如图 8-80所示的三维空间中的一个平面(多元线性回归方程则表示多维空间中的一个超平面)。对于给定的 (X_1, X_2) , Y 的均值就是该平面上正对 (X_1, X_2) 的那个点的 Y 坐标的值, 用实心的小圆圈来表示这个点。对应于实际观测值 Y 的点用空心的小圆圈来表示, 它与实心小圆圈的差别就对应着随机扰动项。

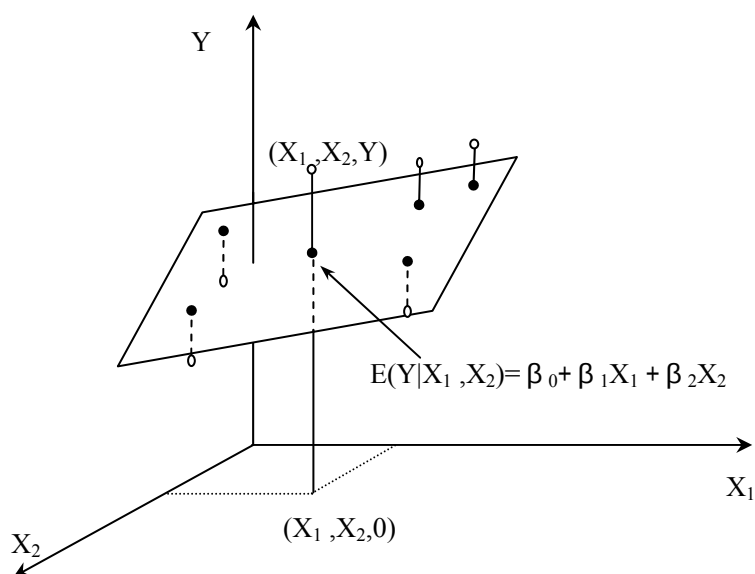


图 8-80 观测点（空心）关于真实回归平面的散点图

假设已给出了 n 次观测值，同时 $b_0, b_1, b_2, \dots, b_p$ 为总体回归参数的估计，则多元线性回归模型的样本回归模型如下：

$$Y = b + b_1 X_1 + b_2 X_2 + \dots + b_p X_p + e$$

式中， e 是 Y 与其估计 $\hat{Y}$ 之间的离差，称为残差。样本回归方程为：

$$\hat{Y} = b + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

对于二元线性回归方程，在几何上是一个平面（如图 8-81 所示），对于不同的观测值，就得到不同的样本回归面。

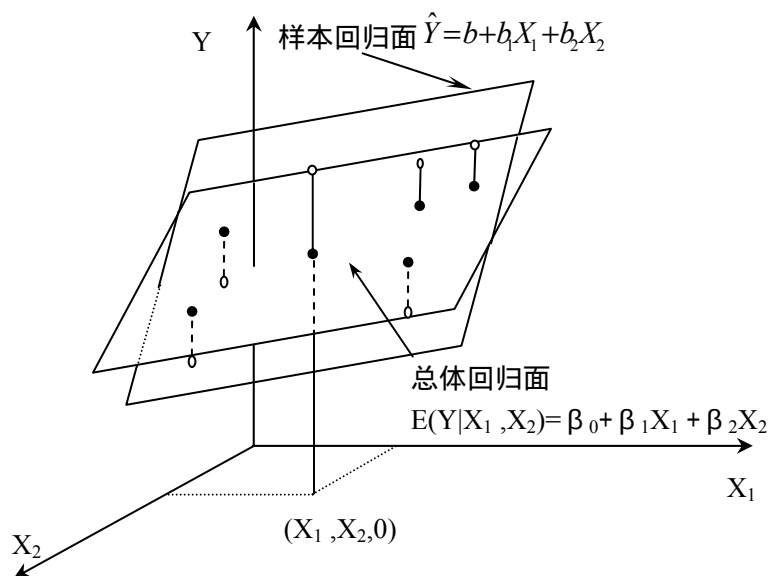


图 8-81 总体回归平面与样本回归平面

二、多元线性回归模型的基本假定

与一元线性回归分析类似，为了估计总体回归参数，也需要提出一些必要的假定。多元线性回归分析的标准假定除了包括上一节中已经提出的关于随机误差项的假定外，还要追加一条假定。这就是回归模型所包含的自变量之间不能具有较强的线性关系。

多元线性回归模型基本假定如下表所示：

	假定名称	假定条件	说明
对扰动项 ε 的假定	1、正态性	$\varepsilon \sim N(0, \sigma^2)$ 且 $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 (i \neq j)$	$Y \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p, \sigma^2)$ 且 $\text{Cov}(Y_i, Y_j) = 0 (i \neq j)$
	2、零均值		
	3、同方差		
	4、互独立		
对自变量 X 的假定	5、非随机	解释是确定性变量	
	6、不相关	解释变量间不存在线性相关关系	
对 X 与 ε 的假定	7、不相关	$\text{Cov}(X, \varepsilon) = 0$	

符合基本假定的多元线性回归模型称为标准的多元线性回归模型。这些假定对于回归模型的估计和检验是很重要的，如果无法满足这些假定，模型参数的普通最小二乘估计将存在一系列问题。我们将在第六、七、八节讨论违背基本假定的主要情况。

第五节 多元线性回归模型的估计和检验

一、回归参数的最小二乘估计及其性质

多元线性回归模型中回归参数的估计同样可采用最小二乘法。由残差平方和

$$\begin{aligned} RSS &= \sum e^2 = \sum (Y - \hat{Y})^2 \\ &= \sum (Y - b_0 - b_1 X_1 - \dots - b_p X_p)^2 \end{aligned}$$

根据微积分中求极小值的原理，可知残差平方和 RSS 存在极小值，欲使 RSS 达到最小， RSS 对 b_0 、 b_1 、 $\dots$ 、 b_p 的偏导数必须等于零。将 RSS 对 b_0 、 b_1 、 $\dots$ 、 b_p 求偏导数，并令其等于零，加以整理后可得到 $p+1$ 个方程式（称为正规方程组或标准方程组），通过求解这一方程组便可以得到 b_0 、 b_1 、 $\dots$ 、 b_p 。实际求解回归参数的估计值，当自变量个数较多时，计算十分复杂简直无法用手算，必须依靠电子计算机完成。在电子计算机技术发达的今天，多元回归分析的计算已经变得相当简单。利用 SPSS，只要将有关数据输入电子计算机，并指定因变量和相应的自变量，立刻就能得到计算结果。因此，对于从事应用研究的人们来说，更为重要的是要能够理解输入和输出之间相互对应的关系，以及对电子计算机输出的结果做出正确的解释。

与一元线性回归模型类似，多元线性回归模型中回归参数的最小二乘估计量也是随机变

量。数学上可以证明，在标准假定条件可以得到满足的情况下，多元回归模型中回归参数最小二乘估计量是最优线性无偏估计量(BLUE)和一致估计量。也就是说，在标准的多元线性回归模型中，高斯——马尔可夫定理同样成立。

除了回归参数以外，多元线性回归模型中还包含了另一个未知参数，那就是随机扰动项的方差 σ^2 。与一元回归分析相类似，多元线性回归模型中的 σ^2 也是利用残差平方和除以其自由度来估计的。即有：

$$S^2 = \frac{\sum e_i^2}{n-p-1} \quad (8-70)$$

上式中， n 是样本容量； p 是方程中解释变量的个数；在 p 元回归模型中，标准方程组有 $p+1$ 个方程式，残差必须满足 $p+1$ 个约束条件，因此其自由度为 $(n-p-1)$ 。可以证明， S^2 是 σ^2 的无偏估计。 S^2 的正平方根 S 又叫做回归估计的标准误差。 S 越小表明样本回归方程的代表性越强。

同一元线性回归相类似，可以计算参数估计量的方差 $S^2(b_j)$ ($j=0,1,2,L,p$) 和标准差（即标准误差） $S(b_j)$ ($j=0,1,2,L,p$)。

[例 8-17] 某种商品的需求量 Y 、价格 X_1 和消费者收入 X_2 的统计资料如所示，试估计 Y 对 X_1 和 X_2 的线性回归方程。

表 8-40 某商品的统计资料

年份	需求量 Y (吨)	价格 X_1 (元)	收入 X_2 (元)
1	59190	23.56	76200
2	65450	24.44	91200
3	62360	32.07	106700
4	64700	32.46	111600
5	67400	31.15	119000
6	64440	34.14	129200
7	68000	35.3	143400
8	72400	38.7	159600
9	75710	39.63	180000
10	70680	46.68	193000

用 SPSS 估计参数步骤如下：

- 1、在 SPSS 中输入变量数据，设变量名分别为 Y 、 X_1 、 X_2 。
- 2、选择主菜单[Analyze]=>[Regression]=>[Linear...]，显示如下图所示的对话框。

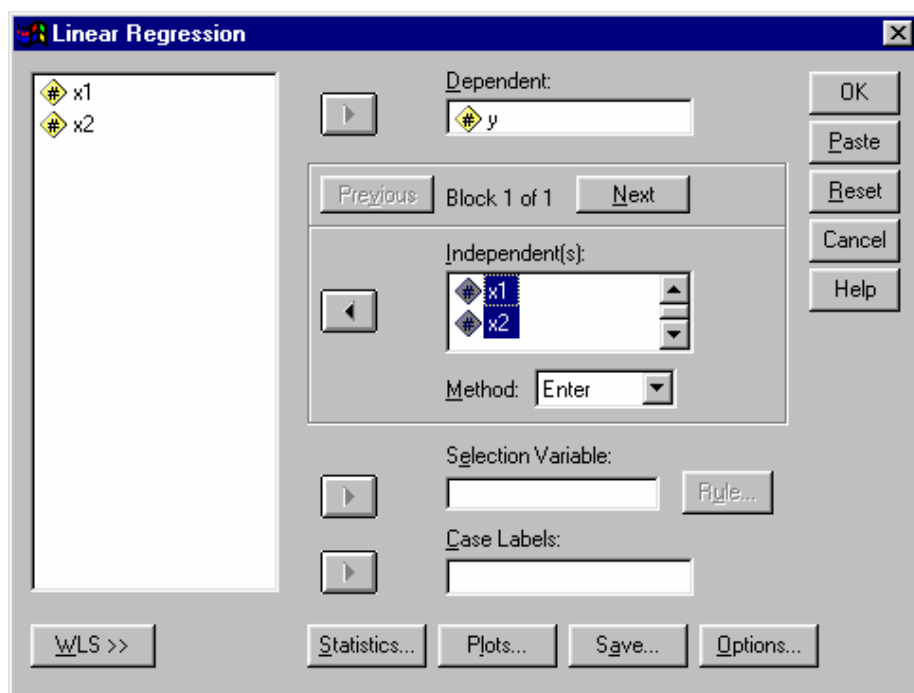


图 8-82 回归分析主对话框

3、选择 Y 进入[Dependent]因变量框，选择 X1、X2 进入[Independent(s)]自变量列表框，单击[OK]。

4、回归结果输出：

表 8-41 [例 8-17]的计算结果输出

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.950	.902	.874	1738.9846

a Predictors: (Constant), X2, X1

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	195318937.424	2	97659468.712	32.294	.000
	Residual	21168472.576	7	3024067.511		
	Total	216487410.000	9			

a Predictors: (Constant), X2, X1

b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	62650.928	4013.010		15.612	.000
	X1	-979.057	319.784	-1.381	-3.062	.018
	X2	.286	.058	2.211	4.902	.002

a Dependent Variable: Y

输出结果说明：

(1) Unstandardized Coefficients B：参数估计值，有样本回归方程：

$$\hat{Y} = 62650.928 - 979.057X_1 + 0.286X_2$$

(2) Unstandardized Coefficients Std. Error : 参数估计值对应的标准误差, 分别为 $S(b_0)=4013.010$ 、 $S(b_1)=319.784$ 、 $S(b_2)=0.058$ 。

(3) Std. Error of the Estimate : 估计的标准误差 S, 本例中 $S=1738.9846$ 。

(4) 其它输出结果在以后加以解释。

二、拟合优度检验

1、样本判定系数 R^2

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (8-71)$$

同一元线性回归分析相类似, $0 \leq R^2 \leq 1$, R^2 越接近于 1, 回归超平面拟合程度越高。

在[例 8-17]的计算结果输出中, R Square : R^2 , 即样本判定系数。 $R^2=0.902$ 说明 Y 的变动中有 90.2%可以由自变量 X_1 和 X_2 解释。

2、调整的判定系数 $\bar{R}^2$

判定系数 R^2 的大小受到自变量 X 的个数 p 的影响。可以证明, 增加自变量 X 的个数, 回归平方和增大, 从而使得 R^2 增大。由于增加自变量个数引起的 R^2 增大与拟合好坏无关, 在含自变量个数 p 不同的模型之间比较拟合程度时, R^2 就不是一个合适的指标, 必须加以调整。调整方法为: 把残差平方和与总离差平方和之比的分子分母分别除以各自的自由度, 变成均方差之比, 以剔除自变量个数对拟合优度的影响。调整的判定系数为:

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \\ &= 1 - \frac{RSS}{TSS} \cdot \frac{n-1}{n-p-1} \\ &= 1 - \frac{TSS-ESS}{TSS} \cdot \frac{n-1}{n-p-1} \\ &= 1 - (1-R^2) \frac{n-1}{n-p-1} \end{aligned} \quad (8-72)$$

由上式可以看出, 只要 $p>0$, 就有 $\bar{R}^2 < R^2$ 。模型中包含的解释变量个数越多, $\bar{R}^2$ 与 R^2 就相差越大。另外, 尽管 R^2 始终是非负的, 但 $\bar{R}^2$ 却可能小于零, 如果出现这种情况, 就作为零值处理。

在[例 8-17]的计算结果输出中, Adjusted R Square : $\bar{R}^2$, 即调整的判定系数, [例 8-17]中 $\bar{R}^2=0.874$ 。

3、复相关系数 R

复相关系数表示所有解释变量与被解释变量 Y 的线性相关程度, 实际上反映的是样本观测值 Y 与拟合值 $\hat{Y}$ 之间的线性相关程度。复相关系数 R 可以由判定系数 R^2 的平方根求得, 也可以通过计算 Y 与 $\hat{Y}$ 的简单相关系数得到。

在[例 8-17]的计算结果输出中, R 为复相关系数。 $R=0.950$, 说明 Y 与自变量 X_1 、 X_2 之间的相关程度为 95.0%。

三、偏相关系数检验

在现实经济活动中, 两个变量之间的相关性大小总要受到其它变量的影响。例如, 某旅游地的冷饮料销售量 Y 与该地游客数量 X_1 之间的关系要受到天气条件 X_2 的影响。这时, Y 与 X_1 的简单相关系数不能反映 Y 与 X_1 之间的真实相关程度。如果要研究 Y 与 X_1 的真实相关性大小, 就必须剔除 X_2 对它的影响。一般地, 在考察多个变量 Y, X_1 , X_2 , ..., X_p 之间的关系时, 如果其它变量不变, 仅考虑 Y 与 X_i ($i=1, 2, \dots, p$) 之间的关系, 这种相关叫偏相关。衡量偏相关程度的指标是偏相关系数。

在偏相关中,根据回定变量数目的多少,可分为零阶偏相关、一阶偏相关、...、(p-1)阶偏相关。零阶偏相关就是简单相关。如果用下标 0 代表 Y,下标 1 代表 X1,下标 2 代表 X2,则变量 Y 与变量 X1 之间的一阶偏相关系数为:

$$r_{01\cdot 2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1-r_{02}^2}\sqrt{1-r_{12}^2}} \quad (8-73)$$

$r_{01\cdot 2}$ 是剔除 X2 的影响之后, Y 与 X1 之间的偏相关程度的度量; r_{01}, r_{02}, r_{12} 分别是 Y, X1, X2 两两之间的简单相关系数。设增加变量 X3, 则变量 Y 与 X1 的二阶偏相关系数为:

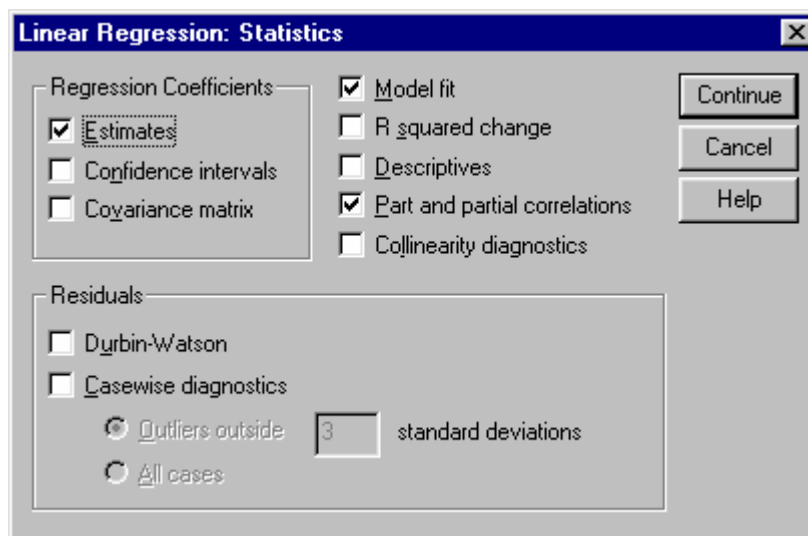
$$r_{01\cdot 23} = \frac{r_{02} - r_{03\cdot 2}r_{13\cdot 2}}{\sqrt{1-r_{03\cdot 2}^2}\sqrt{1-r_{13\cdot 2}^2}} \quad (8-74)$$

一般地,考察多个变量时, Y 与 $X_i(i=1,2,\dots,p)$ 的 p-1 阶偏相关系数,可由(7-44)(7-45)和以下(7-46)式组成一组递推公式进行计算。

$$r_{0i\cdot 12\cdots(i-1)(i+1)\cdots p} = \frac{r_{0i\cdot 12\cdots(i-1)(i+1)\cdots(p-1)} - r_{0p\cdot 12\cdots(p-1)}r_{ip\cdot 12\cdots(i-1)(i+1)\cdots(p-1)}}{\sqrt{1-r_{0p\cdot 12\cdots(p-1)}^2}\sqrt{1-r_{ip\cdot 12\cdots(i-1)(i+1)\cdots(p-1)}^2}} \quad (8-75)$$

对偏相关系数的显著性检验与简单相关系数的显著性检验类似,可以利用相关性系数表进行。用 SPSS 可直接计算出偏相关系数大小。

用 SPSS 对[例 8-17]的资料进行回归分析时,由于使用默认设置进行回归,表 8-41 的输出结果中并没有直接给出偏相关系数。要获得偏相关系数,要改变默认设置,在[Linear Regression]主对话框中单击[Statistics]按钮,出现次级对话框[Linear Regression: Statistics],选择[Part and partial correlations]选项,单击[Continue]返回主对话框。



输出结果中的[Coefficients]表将增加如下几列:

Coefficients

Model		Correlations		
		Zero-order	Partial	Part
1	(Constant)			
	X1	.753	-.757	-.362
	X2	.878	.880	.579

a Dependent Variable: Y

该表中的 Partial Correlations 就是偏相关系数。本例中, $r_{01\cdot 2} = -0.757$, $r_{02\cdot 1} = 0.880$ 。

这里没有给出对应的 p 值。如果要得到偏相关系数及 p 值，必须使用主菜单 [Analyze] => [Correlate] => [Partial] 进行偏相关分析（见上一章）。

四、回归参数的显著性检验（t 检验）

同一元线性回归一样，要检验解释变量 X_j 对因变量 Y 的线性作用是否显著，要使用 t 检验。步骤如下：

1、提出假设。

$$H_0: \beta_j = 0 \quad (j=0, 1, 2, \dots, p)$$

$$H_1: \beta_j \neq 0 \quad (j=0, 1, 2, \dots, p)$$

2、在 H_0 成立条件下，计算 t 统计量

$$t = \frac{b_j - \beta_j}{S(b_j)} = \frac{b_j}{S(b_j)}$$

3、给定显著性水平 α ，查表得临界值 $t_{\frac{\alpha}{2}}(n-p-1)$ （或者计算 t 统计量的 p 值）。

若 $|t| > t_{\frac{\alpha}{2}}(n-p-1)$ （或者 $p < \alpha$ ），就拒绝 H_0 ， X_j 对 Y 有显著线性作用；

若 $|t| < t_{\frac{\alpha}{2}}(n-p-1)$ （或者 $p > \alpha$ ），就接受 H_0 ， X_j 对 Y 线性作用不显著。

例如，在显著性水平 $\alpha=0.05$ 条件下，对[例 8-17]中回归方程的系数 β_1 作 t 检验。

表 8-41 的计算结果中， t 就是 t 统计量值， Sig 是实际显著性水平即 p 值。所以 $t_1 = -3.062$ ， $p_1 = 0.018$ 。因为 $p < \alpha$ ，所以拒绝 H_0 ， X_1 对 Y 的线性作用显著。

五、回归方程的显著性检验（F 检验）

多元线性回归方程的 F 检验就是检验总体回归方程是否显著。即检验所有的参数是否都等于 0。具体步骤如下：

1、提出假设。

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1: \beta_j \text{ 不全为 } 0 \quad (j=0, 1, 2, \dots, p)$$

2、在 H_0 成立条件下， F 统计量

$$F = \frac{ESS/p}{RSS/(n-p-1)} \sim F(p, n-p-1)$$

由样本观测值，计算 F 值。

3、给定显著性水平 α ，查表得临界值 $F_{\alpha}(p, n-p-1)$ （或者计算 F 统计量的 p 值）。

若 $F > F_{\alpha}(p, n-p-1)$ （或者 $p < \alpha$ ），就拒绝 H_0 ，回归方程显著成立，所有自变量对 Y 的影响是显著的；

若 $F < F_{\alpha}(p, n-p-1)$ （或者 $p > \alpha$ ），就接受 H_0 ，回归方程不显著，所有自变量对 Y 的线性作用不显著。

例如，在显著性水平 $\alpha=0.05$ 条件下，对[例 8-17]的回归方程作 F 检验。

计算 F 统计量基于方差分析表。表 8-41 的方差分析表中， F 就是 F 统计量值， Sig 是 F 值的实际显著性水平即 p 值。所以 $F=32.294$ ， $p=0.000$ 。因为 $p < \alpha$ ，所以拒绝 H_0 ，回归方程

线性关系显著。

第六节 非线性回归与曲线回归

一、非线性回归模型的类型

前面讨论过的线性回归模型有这样的特点，即因变量 Y 的均值 $E(Y)$ 不仅是自变量 X 的线性函数，而且同时也是参数 β_1 的线性函数。但是，在现实问题中，变量之间的关系往往不是这样的线性关系，而是非线性的。变量之间的非线性回归模型可以分为三类：

第一类是变量为非线性参数为线性的模型，如抛物线方程和双曲线方程；

第二类是参数为非线性变量为线性的模型，如指数曲线方程；

第三类是变量和参数都是非线性的模型。

对于这三类非线性模型的回归分析是不同的。这里仅考虑可线性化的非线性回归模型。

在对实际的经济现象进行定量分析时，选择恰当的模型形式是很重要的。选择模型具体形式时，必须以经济理论为指导，使模型具体形式与经济学的基本理论相一致，而且模型必须具有较高的拟合优度和尽可能简单的数学形式。

下面介绍常用的几种非线性模型。

(一) 抛物线模型(二次曲线模型)

抛物线模型的具体形式为：

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \quad (8-76)$$

式中 β_0 、 β_1 和 β_2 为待估计参数。

判断某种现象是否适合应用抛物线，可以利用“差分法”。其步骤如下：

首先将样本观察值按 X 的大小顺序排列，然后按以下两式计算 X 和 Y 的一阶差分 X_t 、 Y_t 以及 Y 的二阶差分 Y_{2t} 。

$$X_t = X_t - X_{t-1}; \quad Y_t = Y_t - Y_{t-1} \quad (8-77)$$

$$Y_{2t} = Y_t - Y_{t-1} \quad (8-78)$$

当 X_t 接近于一常数，而 Y_{2t} 的绝对值接近于常数时， Y 与 X 之间的关系可以用抛物线模型近似加以反映。

(二) 双曲线模型

假如 Y 随着 X 的增加而增加（或减少），最初增加（或减少）很快，以后逐渐放慢并趋于稳定，则可选用双曲线来拟合。双曲线模型形式是：

$$Y = \beta_0 + \beta_1 (1/X) + \varepsilon \quad (8-79)$$

(三) 幂函数模型

幂函数模型的一般形式是：

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \cdots X_p^{\beta_p} e^{\varepsilon} \quad (8-80)$$

这类函数的优点在于：方程中的参数可以直接反映因变量 Y 对于某一个自变量的弹性。所谓 Y 对于 X_j 的弹性，是指在其他情况不变的条件下， X_j 变动 1% 时所引起 Y 变动的百分比。弹性是一个无量纲的数值，它是经济定量分析中常用的一个尺度。它在生产函数分析和需求函数分析中，得到了广泛的应用。

(四) 指数函数模型

指数函数模型为：

$$Y = \beta_0 e^{\beta_1 X + \varepsilon} \quad (8-81)$$

这种曲线被广泛应用于描述社会经济现象的变动趋势。例如产值、产量按一定比率增长，成本、原材料消耗按一定比例降低。

(五) 对数函数模型

对数函数是指数函数的反函数，其方程形式为：

$$Y = \beta_0 + \beta_1 \ln X + \varepsilon \quad (8-82)$$

式中, $\ln$ 表示取自然对数。对数函数的特点是随着 X 的增大, X 的单一变动对因变量 Y 的影响效果不断递减。

(六) 逻辑曲线模型

逻辑曲线的方程式如下:

$$Y = \frac{L}{1 + \beta_0 e^{-\beta_1 X + \varepsilon}} \quad (L > 0) \quad (8-83)$$

逻辑曲线具有以下性质。 Y 是 X 的非减函数, 开始时随着 X 的增加, Y 的增长速度也逐渐加快, 但是 Y 达到一定水平之后, 其增长速度又逐渐放慢。最后无论 X 如何增加, Y 只会趋近于 L , 而永远不会超过 L 。由于逻辑曲线的这一特点, 它常被用来表现耐用消费品普及率的变化。

(七) 多项式模型

多项式模型在非线性回归分析中占有重要的地位。因为根据数学上级数展开的原理, 任何曲线、曲面、超曲面的问题, 在一定的范围内都能够用多项式任意逼近。所以, 当因变量与自变量之间的确实关系未知时, 可以用适当幂次的多项式来近似反映。

当所涉及的自变量只有一个时, 所采用的多项式方程称为一元多项式, 其一般形式如下:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon \quad (8-84)$$

前面介绍过的一元线性模型、抛物线函数模型和双曲线函数模型都是一元多项式模型的特例。

当所涉及的自变量在两个以上时, 所采用的多项式称为多元多项式。例如, 二元二次多项式模型的形式如下:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \varepsilon \quad (8-85)$$

一般来说, 涉及的变量越多, 变量的幂次越高, 计算量就越大。因此, 在实际的经济定量分析中, 一般尽量避免采用多元高次多项式。

下面将讨论如何将非线性模型线性化和曲线回归。

二、非线性模型的线性化估计

对于非线性回归模型, 除了要考虑如何根据所要研究的问题的性质并结合实际的样本资料确定其具体形式外, 还要考虑如何估计模型中的参数。非线性回归模型最常用的方法仍然是最小二乘估计法, 但需要根据模型的不同类型, 作适当的变换。

许多具有实用价值的非线性回归函数, 可以通过适当的变换, 转化为线性回归函数, 然后再利用线性回归分析的方法进行估计和检验。常用的非线性函数的线性变换方法有以下几种:

(一) 倒数变换

倒数变换是用新的变量来替换原模型中变量的倒数, 从而使原模型变成线性模型的一种方法。例如, 对于双曲线函数, 令 $X^* = 1/X$ 代入原方程 (8-79) 式, 得 $Y = \beta_0 + \beta_1 X^* + \varepsilon$, 从而转化为一元线性回归模型。

(二) 半对数变换

这种方法主要应用于对数函数模型的线性变换。对于对数函数模型 (8-82), 令 $X^* = \ln X$, 代入原方程, 同样可得: $Y = \beta_0 + \beta_1 X^* + \varepsilon$ 。

(三) 双对数变换

这种方法通过用新变量替换原模型中变量的对数, 从而使原模型变换为线性模型。例如, 对幂函数模型 (8-80) 的两边求对数, 可得:

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \cdots + \beta_p \ln X_p + \varepsilon \quad (8-86)$$

令 $Y^* = \ln Y$; $\beta_0^* = \ln \beta_0$; $X_1^* = \ln X_1, \dots, X_k^* = \ln X_k$, 代入上式可得:

$$Y^* = \beta_0^* + \beta_1 X_1^* + \beta_2 X_2^* + \dots + \beta_p X_p^* \quad (8-87)$$

(四) 多项式变换

这种方法适用于多项式方程的变换。例如,对于(8-85)式给出的二元二次多项式,可令 $X_2^* = X_1$, $X_3^* = X_2$, $X_4^* = X_1 X_2$, $X_5^* = X_1^2$, $X_6^* = X_2^2$, 代入原方程, 可得:

$$Y = \beta_0 + \beta_1 X_2^* + \beta_2 X_3^* + \beta_3 X_4^* + \beta_4 X_5^* + \beta_5 X_6^* \quad (8-88)$$

以上所述的线性变换的方法具有简便易行的优点。但是,在实际应用时要注意以下几个问题:

第一、对于一些比较复杂的非线性函数,常常需要综合利用上述的几种方法。例如,对于(8-83)式给出的逻辑曲线模型,若假定 $L=20$,则可以采用以下方式进行线性变换。首先,(8-83)式两边同时取倒数,可得:

$$\frac{1}{Y} = \frac{1 + \beta_0 e^{-\beta_1 X + \varepsilon}}{20} \quad (8-89)$$

进而又有:

$$\frac{20}{Y} - 1 = \beta_0 e^{-\beta_1 X + \varepsilon}$$

上式两边取对数,可得:

$$\ln\left(\frac{20}{Y} - 1\right) = \ln \beta_0 - \beta_1 X + \varepsilon \quad (8-90)$$

令 $Y^* = \ln(20/Y - 1)$, $\beta_0^* = \ln \beta_0$, $\beta_1^* = -\beta_1$, 代入上式, 可得:

$$Y^* = \beta_0^* + \beta_1^* X + \varepsilon \quad (8-91)$$

在以上变换过程中,综合利用了倒数变换和对数变换。

第二、为了能够根据样本观测值,对通过变换得到的线性回归方程式进行估计,该方程中的所有变量都不允许包含未知的参数。例如,如果 L 未知,上面所述的逻辑曲线模型的线性变换就是不正确的。这是因为这种情况下 Y^* 包含了未知的 L , 因而是不可观测的。

第三、与线性回归分析的场合一样,非线性回归分析也要考虑随机扰动项的问题。只有当变换后的新模型中包含的扰动项能够满足各种基本假定时,新模型中回归参数最小二乘估计量的各种理想性质才能成立。

第四、严格地说,上述的各种线性变换方法只是适用于变量为非线性的函数。对于参数为非线性或参数与变量均为非线性的模型来说,即使有可能进行线性变换和回归估计,也无法得到原模型中非线性参数的无偏估计量。

最后,并不是所有的非线性模型都可以通过变换得到与原方程完全等价的线性模型。在遇到这种情况时,还需要利用其他一些方法如泰勒级数展开法等去进行估计。这里不作进一步的介绍。

[例-18] 1996 年我国城镇居民收入情况如所表 8-42 示。表中资料共有 16 组, X 是各组的人均生活费收入, Y 是各组的人均生活费支出。试建立 Y 对 X 的回归模型。

表 8-42 人均生活费收入和人均生活费支出 (单位: 元)

Y	X	Y	X
1493.47	1017.52	4996.12	6160.77
1762.82	1643.86	5692.75	6785.27
2298.40	2300.04	6102.06	7503.47
2784.98	2917.52	5712.40	8106.87
3345.43	3567.42	6886.65	8814.28
3769.01	4205.01	8877.78	9427.21
3981.00	4881.93	6561.70	10001.90
4805.03	5521.33	8311.68	12582.52

用 SPSS 估计步骤如下：

1、绘制 X 与 Y 的散点图。选择[Graphs]>[Scatter]，选择[Simple]后单击[Define]，在出现的[Simple Scatterplot]主对话框(如图 8-83所示)中把 Y 和 X 分别选入[Y Axis]和[X Axis]，单击[OK]开始绘散点图。



图 8-83

2、从如下的散点图可以看出，人均生活费支出先是随着人均生活费收入的提高而快速提高，但当收入达到一定水平后，生活费支出的增幅明显趋缓。因此，用线性回归模型表示 Y 和 X 的关系是不恰当的。

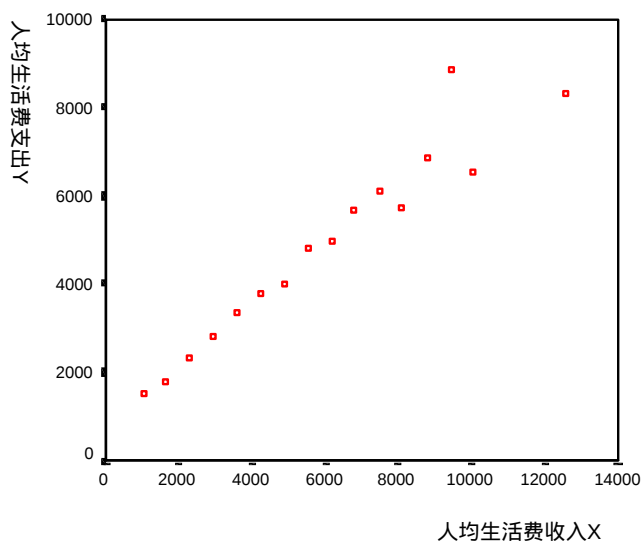


图 8-84

3、对 Y 和 X 分别取自然对数 $\ln Y$ 和 $\ln X$ ，画出 $\ln Y$ 和 $\ln X$ 的散点图。

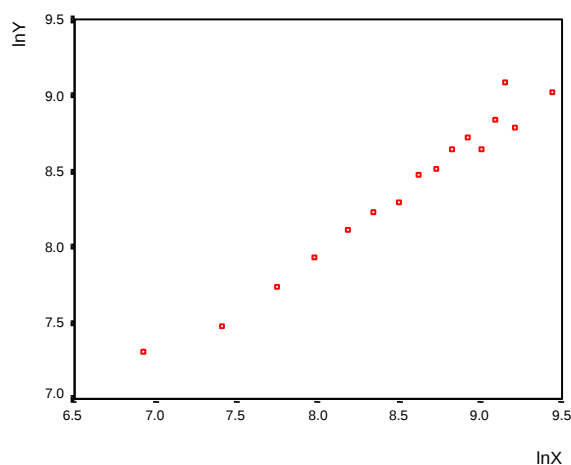


图 8-85

可以看出, $\ln Y$ 和 $\ln X$ 在散点图上近似为线性关系。于是把回归模型高定为幂函数模型:

$$Y = \beta_0 X^{\beta_1} e^{\varepsilon} \quad (8-92)$$

并进行双对数变换, 得

$$\ln Y = \ln \beta_0 + \beta_1 \ln X + \varepsilon \quad (8-93)$$

分别令 $Y^* = \ln Y, \beta_0^* = \ln \beta_0, X^* = \ln X$, 得到线性回归模型:

$$Y^* = \beta_0^* + \beta_1 X^* + \varepsilon \quad (8-94)$$

4、估计 (8-94)。选择主菜单[Analyze]=>[Regression]=>[Linear], 在出现的主对话框中, 选择因变量 $\ln y$ 和自变量 $\ln x$, 单击[OK]。

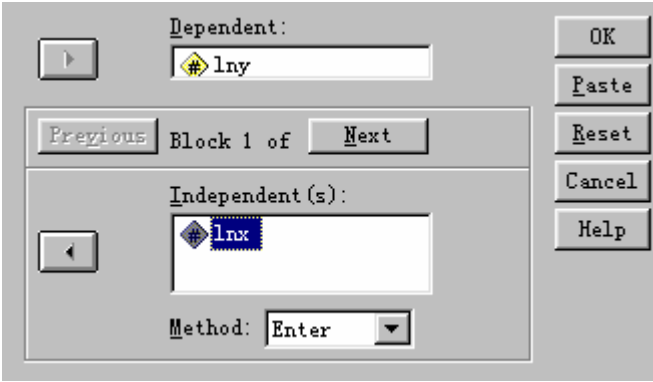


图 8-86

5、回归结果如下:

$$\begin{aligned} \hat{Y}^* &= 2.006 + 0.748X^* \\ t\text{值: } &7.098 \quad 22.580 \\ p\text{值: } &0.000 \quad 0.000 \\ R^2 &= 0.973 \quad F = 509.847 \quad p_F = 0.000 \end{aligned}$$

模型通过参数显著性检验。注意到这里的 b_1 是弹性值, 即人均生活费收入每提高 1%, 人均生活费支出平均增加 0.748%。

三、曲线估计

在 SPSS 中, 可以对一些特殊的非线性模型直接进行估计。这类模型的方程式如所示。

表 8-43 非线性模型方程式

名 称		方程式
Linear (一元线性)	LIN	$Y = \beta_0 + \beta_1 t$
Quadratic(二次函数)	QUA	$Y = \beta_0 + \beta_1 t + \beta_2 t^2$
Compound(复合函数)	COM	$Y = \beta_0 (\beta_1)^t$
Growth(生长函数)	GRO	$Y = e^{(\beta_0 + \beta_1 t)}$
Logarithmic(对数函数)	LOG	$Y = \beta_0 + \beta_1 \ln t$
Cubic(三次函数)	CUB	$Y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$
S(S 形曲线)	S	$Y = e^{(\beta_0 + \beta_1 / t)}$

Exponential(指数函数)	EXP	$Y = \beta_0 e^{\beta_1 t}$
Inverse(逆函数)	INV	$Y = \beta_0 + \beta_1 / t$
Power(幂函数)	POW	$Y = \beta_0 t^{\beta_1}$
Logistic(逻辑函数)	LGS	$Y = \frac{1}{\frac{1}{u} + \beta_0(\beta_1 t)}$

表中， t 为时间或自变量， β_0 为常数项， β_1 为回归参数， e 表示自然对数的底， $\ln$ 表示以 e 为底的自然对数。

在 SPSS 中进行曲线回归分析的操作步骤如下：

1、选择主菜单[Analyze]=>[Regression]=>[Curve Estimation(曲线估计)](如图 8-87 所示)。

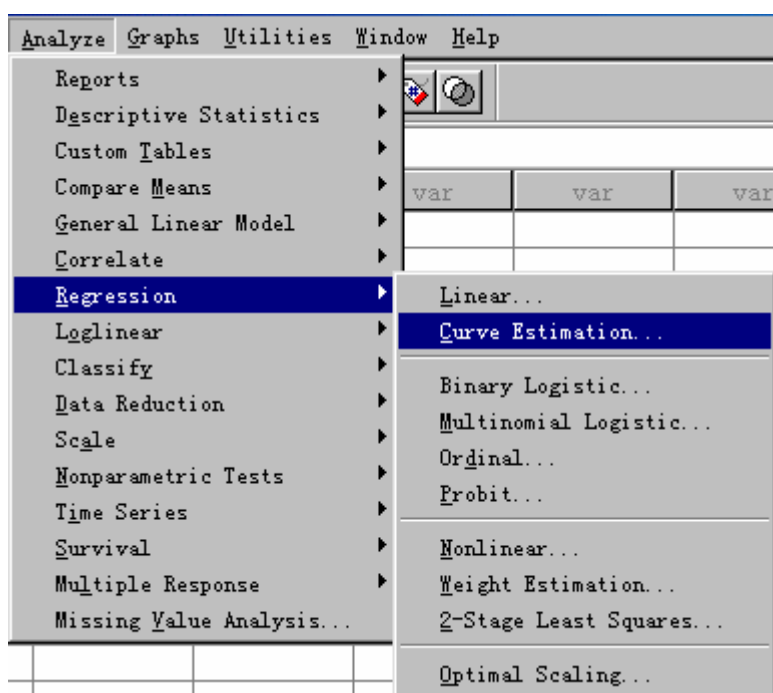


图 8-87

2、显示[Curve Estimation]主对话框，如图 8-88所示。选择因变量 Y ，并选择某一变量作为自变量或以时间作为自变量。根据实际问题选择合适的模型（各模型方程式见表 8-43）。主对话框中其它选项的含义为：

Include constant in equation：选择此项表示在方程式中计算常数项；

Plot models：选择此项根据所确定的模型产生模型图；

Display ANOVA table：选择此项将列出方差分析结果。

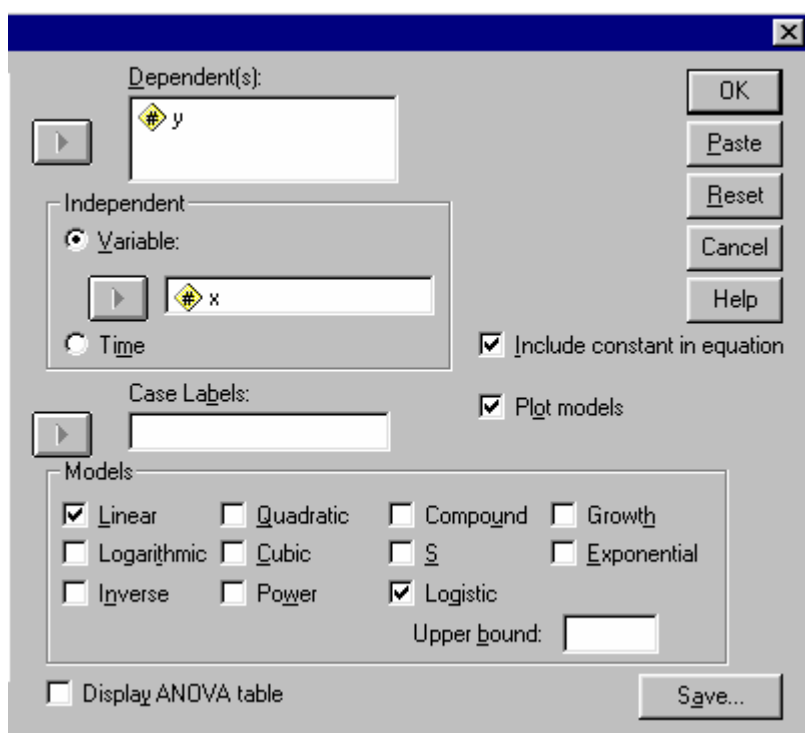


图 8-88

3、如果要保存预测值、预测区间、显著性水平等，在主对话框中单击[Save]，显示如图 8-89 所示的二级对话框。

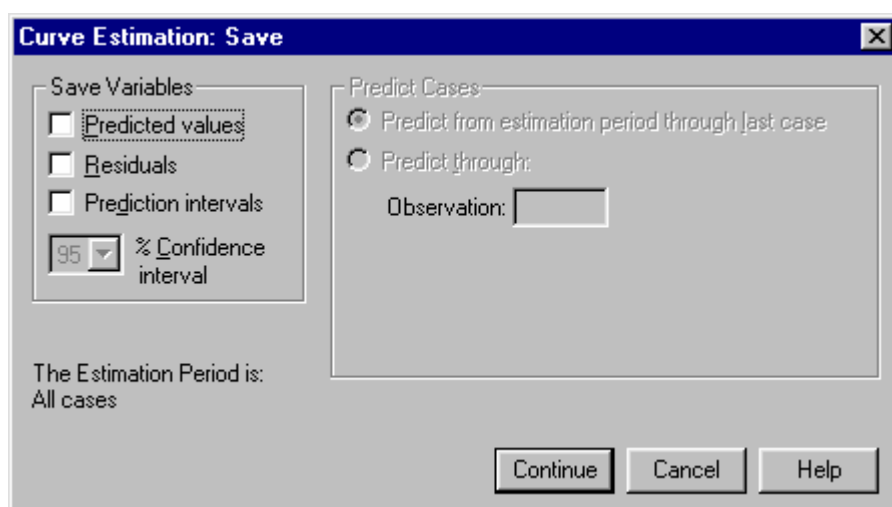


图 8-89

Predicted values：因变量的预测值；

Residuals：残差值；

Prediction intervals：预测区间；

Confidence interval：预测区间的置信区间。如果选择了 Prediction intervals 选项，可以选择置信水平，默认为 95%。

如果选择时间作为自变量，则可以通过在单击[Save]按钮后，在[Predict cases]选项中确定一种超过数据时间序列的预测周期。

Predict from estimation period through last cases：依据估计周期为所有的观测值提供预测值。

Predict through: 如果预测的范围超过了当前数据文件中的最后一个观测值, 选择此项在随后的[Observation]中键入一个周期数值。

4、返回对主对话框, 单击[OK]进行曲线回归分析。

[例-19]根据[例-18]的资料, 用 SPSS 进行曲线回归分析。

步骤如下:

1、选择主菜单[Analyze]>[Regression]>[Curve Estimation]打开曲线回归主对话框, 把 Y 和 X 分别选入因变量框和自变量框中, 选择 Linear、Compound、Growth、Logarithmic、S、Exponential、Inverse、Power 模型。单击[OK]进行估计。

2、输出结果如下:

Independent: X

Dependent	Mth	Rsq	df	F	Sigf	b0	b1
Y	LIN	.927	14	178.32	.000	975.301	.6473
Y	LOG	.873	14	95.91	.000	-20282	2954.27
Y	INV	.624	14	23.27	.000	6788.94	-7.E+06
Y	COM	.888	14	111.08	.000	1736.20	1.0002
Y	POW	.973	14	509.85	.000	7.4312	.7480
Y	S	.820	14	63.63	.000	8.9017	-2024.8
Y	GRO	.888	14	111.08	.000	7.4595	.0002
Y	EXP	.888	14	111.08	.000	1736.20	.0002

结果中, Mth表示模型形式(见表8-43), Rsq表示 R^2 , df表示自由度, F表示F检验值, Sigf表示F检验值的实际显著性水平即p值, b0表示常数项, b1表示回归参数。

可以看出, 模型为POW形式时的 R^2 最大, LIN次之。我们选用幂函数模型拟合人均生活费收入和人均生活费支出是合适的。

第七节 多重共线性

一、多重共线性的产生与后果

在经济实践中, 人们对建立的回归模型进行估计时所用的样本数据, 往往不是专门为此目的而在科学设计的受控试验中采集的, 它们基本上都是在事物的自然发展变化过程中观测记录取得的。这样的样本数据, 对于所要估计的回归模型而言, 往往不能提供充足的信息, 如某些变量观测值的变化范围很狭小, 又如对某些变量的观测只有很少几个值, 等等。其中, 样本数据中最常遇到且需要特别重视的一个问题就是多重共线性问题。

所谓多重共线性, 是指线性回归模型中的若干解释变量或全部解释变量的样本观测值之间具有某种线性关系。多重共线性可分为完全和不完全多重共线性两种情况。完全多重共线性, 是指解释变量之间存在完全的线性关系(例如, 解释变量 X_1 与 X_2 的相关系数为 1); 不完全多重共线性, 是指解释变量之间存在着近似的线性关系(例如, 解释变量 X_1 与 X_2 的相关系数近似等于 1)。如果解释变量之间没有任何线性关系, 则称无多重共线性(X_1 与 X_2 的相关系数为零), 这正是标准线性回归模型的基本假定之一, 但这时并不排除解释变量之间存在非线性关系。

虽然, 样本来自总体, 样本中包含有总体信息, 当总体密切相关时, 多重共线性往往严重。但是, 即使总体并非密切相关, 样本数据间也可能存在线性关系, 多重共线性本质上是个样本现象。

回归模型中存在多重共线性问题,将给模型的估计带来一系列后果。如果解释变量之间存在完全的多重共线性,那么无法估计模型参数,参数估计的方差将为无穷大,这将使得回归模型的普通最小二乘法估计完全失效。当然,在经济实践中,解释变量之间的完全多重共线性是比较少见的,但是解释变量之间具有高度相关性却是十分常见的。这是因为许多经济变量之间往往都存在一定的相互联系,例如影响家庭消费支出的可支配收入与家庭财富两个变量之间就存在明显的高度线性相关关系。尤其是在使用时间序列数据进行回归分析时,由于许多解释变量都有随着时间的推移而同方向变动的趋势,往往使用解释变量之间的线性相关性更为严重。在此情形之下,虽然可以计算出参数的估计值,但是参数的估计很不稳定,参数估计值对样本数据或样本容量变化极为敏感,甚至会改变参数的原有正确符号,同时估计量的方差可能会很大,从而导致对模型参数进行假设检验时,出现各参数估计量的 t 值均很低而 F 值却很高的取舍矛盾。一般来说,对于一个特定的样本,这种不完全多重共线性可能产生的后果主要有:

1、各个解释变量对被解释变量的影响很难精确鉴别。回归参数是在其它变量不变的条件下某个解释变量的变动对被解释变量的影响,但是由于解释变量之间多重共线性的存在,某个解释变量的变动将也引起其它解释变量的变动,从而将使我们难以区分各个解释变量对被解释变量的影响大小。

2、由于存在多重共线性,模型回归参数估计量的方差会很大,这将使得进行显著性检验时认为回归参数的值与零无显著差异。从而导致将相应的解释变量从模型中剔除,但这并不是因为该解释变量对被解释变量无影响作用,而只是由于样本数据不适于精确区分各解释变量的单独影响。

3、模型参数的估计量对删除或增加少量的观测值以及删除一个不显著的解释变量都可能非常敏感。

二、多重共线性的检验

对于一个给定的线性回归模型和一组样本数据,其解释变量的样本数据中是否存在多重共线性以及共线性的严重程度有多大?显然需要用一定的方法来进行检验。如果在对多元线性回归模型进行统计检验时,发现参数估计值的大小或(和)符号违背经济理论,或者判定系数 R^2 、 F 检验值很大(p 值小)而各个偏回归参数的 t 检验值均偏小(其 p 值大于 α),那么很有可能是因为解释变量之间存在多重共线性,这是实际问题中是经常出现的。实践中,常用的检验方法主要有简单相关系数检验法、容限度(Tolerance)法、方差扩大因子(VIF, Variance Inflation Factor)法、特征值和条件指数(Eigen-Value and Condition Indexes)法、Theil 多重共线性效应系数法等。这里只简要介绍容限度法和方差扩大因子法。

容限度和方差扩大因子是检验多重共线性的两个重要指标。容限度是由每个自变量 X_j 作为因变量对其他自变量回归时得到的余差比例,即:

$$Tolerance_j = 1 - R_j^2$$

其中, R_j^2 表示第 j 个自变量对其他自变量进行回归得到的判定系数 R^2 。容限度很大时, R_j^2

很小,说明所 X_j 包含的独立信息很多,可能成为重要解释变量;反之,容限度很小, R_j^2 很大,说明 X_j 与其它自变量的信息重复性越大,其对因变量 Y 的解释能力越小。容限度的大小是根据研究者的具体需要制定的,通常当容限度小于 0.1 (这里 $R_j^2 > 0.9$) 时,便认为变量 X_j 与其他变量之间的多重共线性超过了容许界限。

方差扩大因子(以下简称 VIF)是容限度的倒数。即:

$$VIF_j = \frac{1}{Tolerance_j} = \frac{1}{1 - R_j^2}$$

它表示所对应的偏回归系数的方差由于多重共线性而扩大的倍数。当容限度为 0.1 时，VIF 为 10（倍）。通常当 $VIF_j > 10$ 时，便认为变量 X_j 与其他变量之间存在多重共线性。

需要 SPSS 输出多重共线性检验指标时，必须改变默认设置。在如图 8-82 所示的回归分析主对话框中单击 [Statistics (统计量)] 按钮，打开 [Linear Regression: Statistics] 子对话框，选择 [Collinearity diagnostics (共线性诊断)]，单击 [Continue] 返回主对话框并单击 [OK] 按钮。这样 SPSS 便可输出所有检查多重共线性的指标。从表 8-44 某商品需要量对价格和消费者收入回归的部分输出结果可以看出， X_1 与 X_2 存在多重共线性问题。

表8-44 例8-2某商品需要量对价格和消费者收入回归的部分输出结果

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	62650.928	4013.010		15.612	.000		
X1	-979.057	319.784	-1.381	-3.062	.018	.069	14.559
X2	.286	.058	2.211	4.902	.002	.069	14.559

三、多重共线性的处理

对于给定的样本数据，如果其存在较严重的多重共线性，那么就必须采取一些措施进行处理，以减轻其不良影响。常用的处理方法有删除不重要的解释变量、追加样本信息、利用非样本先验信息、改变解释变量形式、逐步回归法等等。

（一）删除不重要的解释变量

解释变量之间存在共线性，说明解释变量所提供的信息是重叠的，可以删除不重要的解释变量减少重复信息。但从模型中删去解释变量时应该注意：从实际经济分析确定为相对不重要并从偏相关系数检验证实为共线性原因的那些变量中删除。如果删除不当，会产生模型设定误差，造成参数估计严重有偏的后果。

（二）追加样本信息

多重共线性问题的实质是样本信息的不充分而导致模型参数的不能精确估计，因此追加样本信息是解决该问题的一条有效途径。但是，由于资料收集及调查的困难，要追加样本信息在实践中有时并不容易。

（三）利用非样本先验信息

非样本先验信息主要来自经济理论分析和经验认识。充分利用这些先验的信息，往往有助于解决多重共线性问题。

（四）改变解释变量的形式

改变解释变量的形式是解决多重共线性的一种简易方法，例如对于横截面数据采用相对数变量，对于时间序列数据采用增量型变量。

（五）逐步回归法

逐步回归 (Stepwise Regression) 是一种常用的消除多重共线性、选取“最优”回归方程的方法。其做法是将逐个引入自变量，引入的条件是该自变量经 F 检验是显著的，每引入一个自变量后，对已选入的变量进行逐个检验，如果原来引入的变量由于后面变量的引入而变得不再显著，那么就将其剔除。引入一个变量或从回归方程中剔除一个变量，为逐步回

归的一步，每一步都要进行 F 检验，以确保每次引入新变量之前回归方程中只包含显著的变量。这个过程反复进行，直到既没有不显著的自变量选入回归方程，也没有显著自变量从回归方程中剔除为止。

这里对自变量的显著性检验不是采用通常的 t 检验，而是采用 F 检验（不同于回归方程的显著性检验）。F 检验采用的检验统计量是 t 检验统计量的平方，即 $F = t^2$ 。我们把引入变量的 F 检验的临界值和 p 值分别记为 F_{entry} 和 p_{entry} ，把剔除变量的 F 检验的临界值和 p 值分

别记为 $F_{removal}$ 和 $p_{removal}$ ，通常取 $F_{entry} > F_{removal}$ 或 $p_{entry} < p_{removal}$ ，在 SPSS 的默认设置条件下，

$F_{entry} = 3.84$ ， $p_{entry} = 0.05$ ， $F_{removal} = 2.71$ ， $p_{removal} = 0.1$ ，并采用 p_{entry} 和 $p_{removal}$ 进行检验。

对例 8-2 进行逐步回归的 SPSS 操作步骤如下：

- (1) 在 SPSS 中输入变量数据，设变量名分别为 Y、X1、X2。
- (2) 选择主菜单[Analyze]⇒[Regression]⇒[Linear...]，显示如图 8-82 所示的对话框。选择 Y 进入因变量框，选择 X1、X2 进入自变量列表框。选择[Method（方法）]为[Stepwise]。
- (3) 单击[Options]按钮，打开[Linear Regression: Options]子对话框（如图 8-90 所示），该对话框的[Stepping Method Criteria]栏给出了引入/删除自变量的默认 F 临界值/p 值。本例采用 SPSS 默认值，单击[Continue]按钮返回。

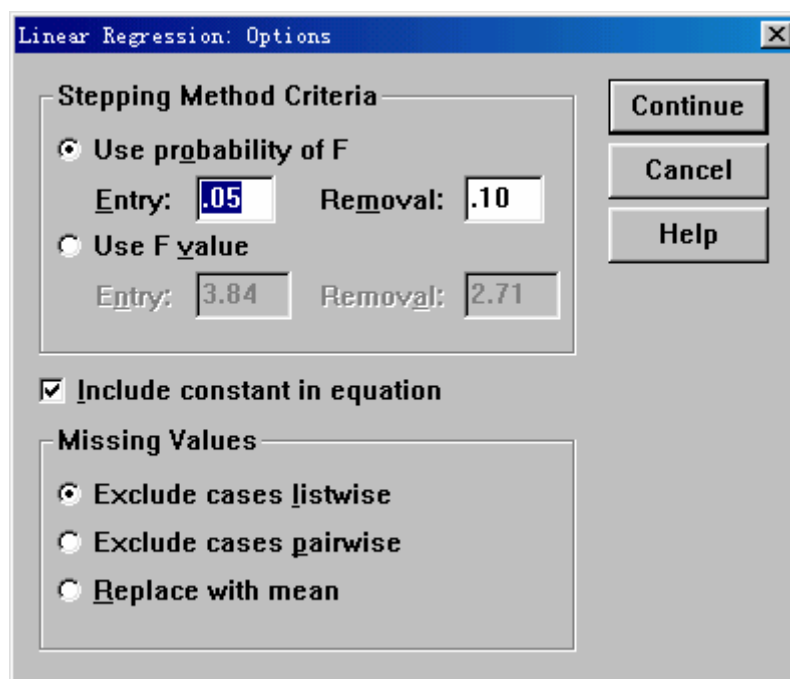


图 8-90 回归分析选项对话框

- (4) 单击主对话框中的[OK]按钮，输出结果为：

Coefficients（系数）

	Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
	B	Std. Error			Tolerance	VIF
1 (Constant)	52140.580	2973.212	17.537	.000		
X2	.114	.022	5.194	.001	1.000	1.000

2	(Constant)	62650.928	4013.010	15.612	.000		
	X2	.286	.058	4.902	.002	.069	14.559
	X1	-979.057	319.784	-3.062	.018	.069	14.559

a Dependent Variable: Y

Excluded Variables (剔除的变量)

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
					Tolerance	VIF	Minimum Tolerance
1 X1	-1.381	-3.062	.018	-.757	6.869E-02	14.559	6.869E-02

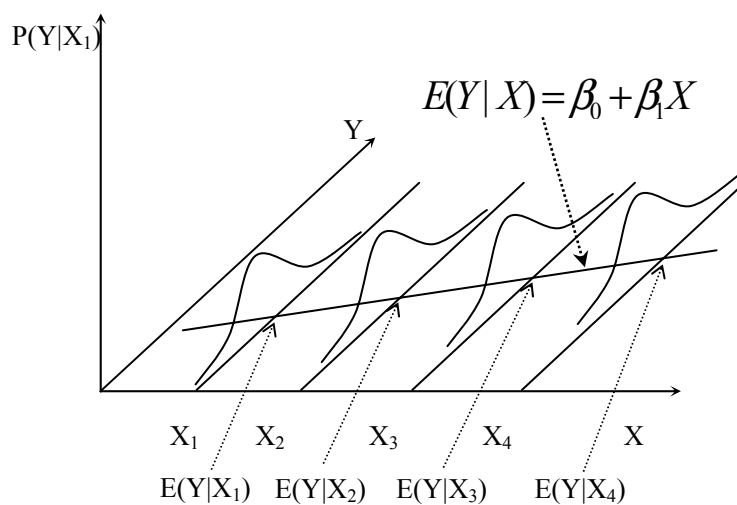
a Predictors in the Model: (Constant), X2

b Dependent Variable: Y

第八节 异方差

一、异方差的产生与后果

异方差是对同方差假定的违反,是指随机扰动项的方差(也是因变量 Y 的方差)随着自变量取值的变化,而不是一个常数。在讨论异方差时,我们仍假设其它所有假定仍然满足。同方差与异方差情况下,因变量 Y 的分布如图 8-91 所示。



(A) 同方差

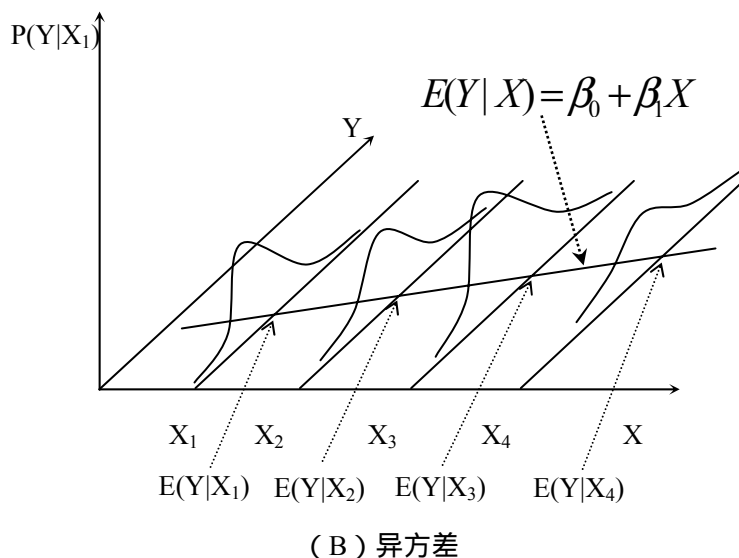


图 8-91 同方差与异方差情况下的分布

异方差在现实经济中是经常出现的。例如，在[例-16]中认为随机扰动项或因变量满足同方差假定是不符合实际情况的。对于那些低收入家庭来说，除去购买生活必需品后的余钱不多，其消费支出的差异性将不会很大，即方差较小；而对于那些高收入家庭来说，除去购买生活必需品以后的余钱还很多，这些余钱可以用于购买奢侈消费品，也可用于储蓄或投资，其消费支出的差异性将会很大，即方差较大；显然这里存在异方差现象。

当回归模型中的扰动项存在异方差时，参数的普通最小二乘法估计量仍然是线性无偏的，但不再是最小方差的估计量，即不再是有效的；而且参数估计量的方差是有偏的，导致参数的假设检验也是非有效的。

二、异方差的检验

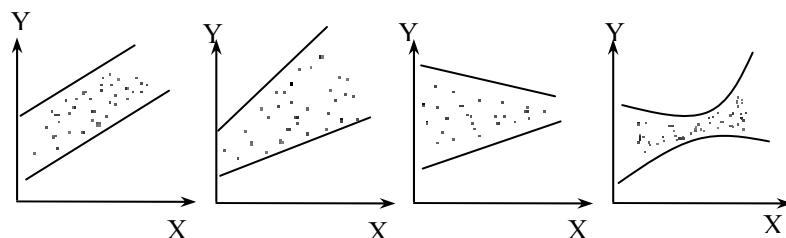
由于扰动项存在异方差时会造成严重的后果，所以必须对模型中的扰动项是否存在异方差进行检验。常用的检验方法有：图示法、样本分段比较法（如 Goldfeld-Quandt 检验法）、残差回归检验法（如 Glejser（格里奇）检验法和 Park（帕克）检验法、White（怀特）检验法）等。

（一）图示法

异方差是扰动项的方差随着解释变量的变化而变化，故可以利用因变量 Y 与解释变量 X 的散点图、残差平方 e^2 与 X 的散点图或者学生化残差与因变量标准化预测值的散点图，来判断是否存在异方差。

1、因变量 Y 与解释变量 X 的散点图

当 Y 与 X 的散点图分布的区域逐渐变宽、变窄或出现不规则的复杂变化时，都可能存在异方差。如图 8-92 所示，(A) 为同方差的情形，(B) 为递增异方差的情形，(C) 为递减异方差的情形，(D) 为复杂异方差的情形。



(A) 同方差 (B) 递增异方差 (C) 递减异方差 (D) 复杂异方差

图 8-92 异方差条件下 Y 与 X 的散点图

在 SPSS 中, 通过绘散点图进行异方差检验。步骤如下:

(1) 选择菜单[Graphs]⇒[Scatter], 在显示如下图所示的对话框中选择[Simple], 并单击[Define]按钮。

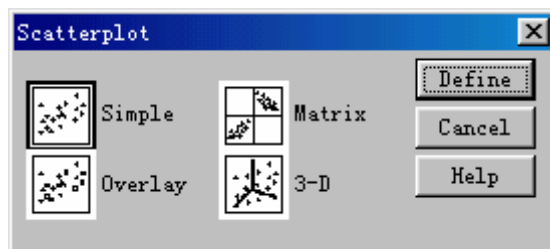


图 8-93

(2) 显示[Simple Scatterplot]对话框, 把 Y 和 X 分别选入[Y Axis]、[X Axis], 然后单击[OK]。

(3) 在输出窗口将显示 Y 与 X 的散点图。

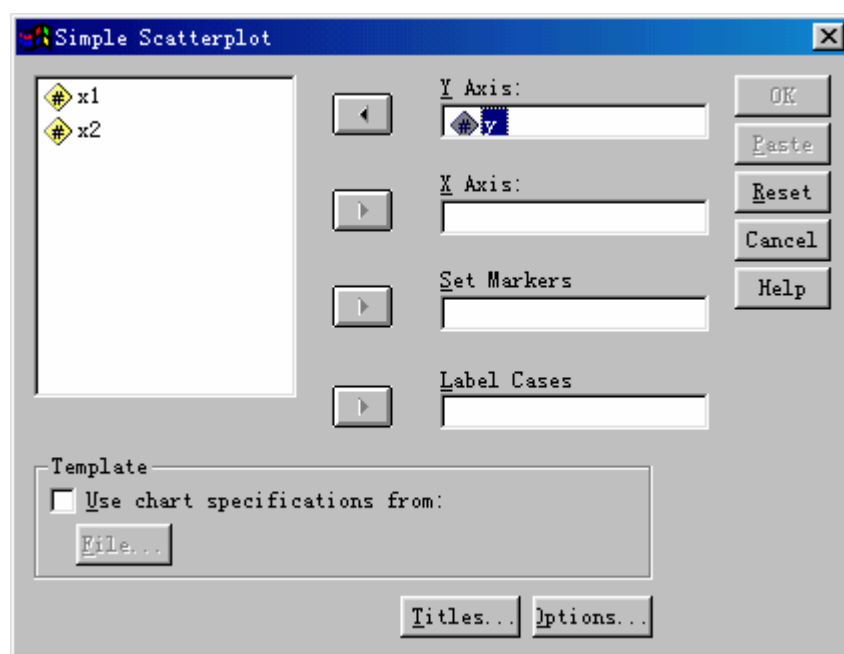
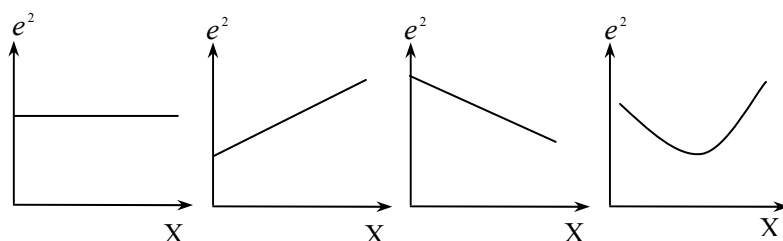


图 8-94

2、残差平方 e^2 与 X 的散点图

如果扰动项 ε 存在异方差, 在残差平方 e^2 与 X 的散点图上, e^2 不会近似于某一常数。因此可依此判断是否存在异方差。各种异方差情形所对应的 e^2 与 X 的散点图如图 8-95 所示。



(A) 同方差 (B) 递增异方差 (C) 递减异方差 (D) 复杂异方差

图 8-95 异方差条件下的残差平方 e^2 与 X 散点图

在 SPSS 中进行检验时，要绘制残差平方 e^2 与 X 的散点图，首先必须根据回归结果保存残差，并计算残差的平方，然后画残差平方与 X 的散点图。步骤如下：

(1) 回归计算时，在回归主对话框中单击[Save]按钮，显示如所示的对话框，选择[Residuals]中的[Unstandardized]选项，单击[Continue]返回主对话框。输出结果的同时，将把残差保存为一新变量 res_1 。

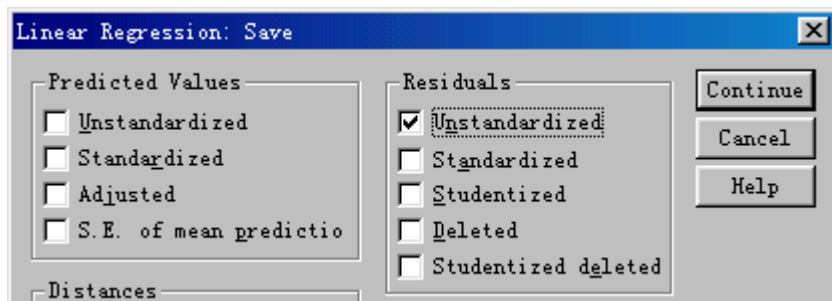


图 8-96 保存残差

(2) 计算残差的平方。选择 [Transform] => [Compute]，在显示的对话框中输入残差平方的变量名（如 $e2$ ）和计算残差平方的表达式（如 $res_1 ** 2$ ）。单击[OK]后将产生一个新变量。

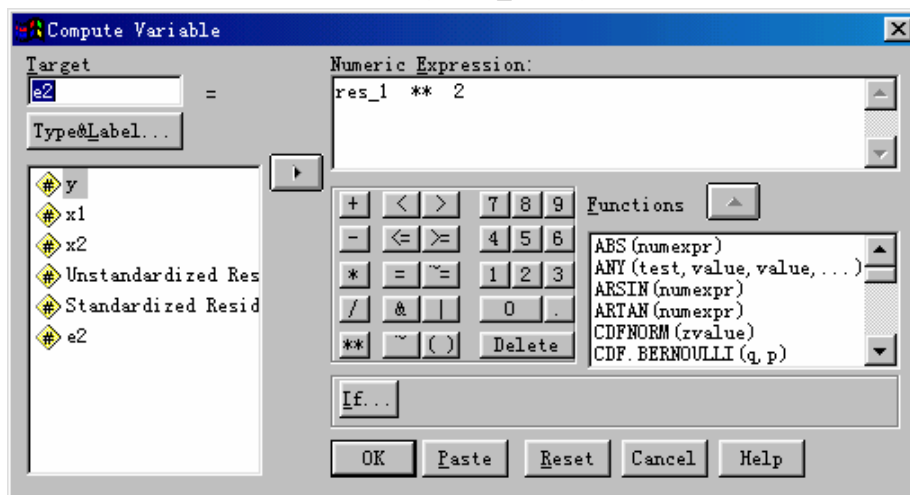


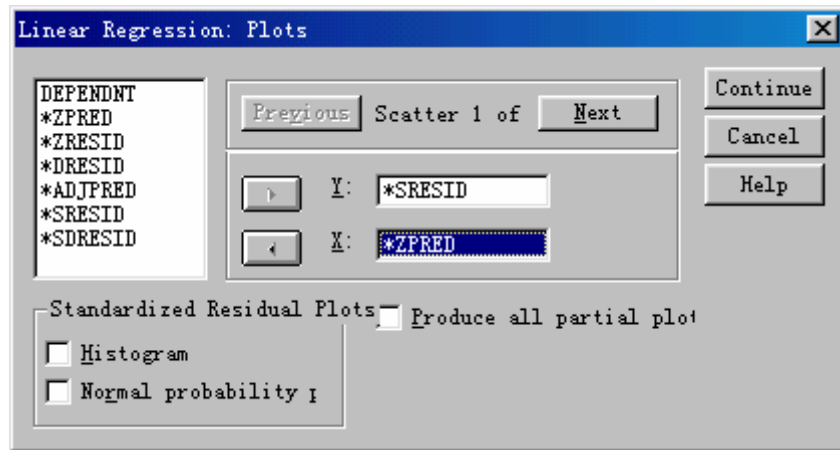
图 8-97 计算残差平方

(3) 画出新变量与 X 的散点图，并进行判断。

3、学生化残差与因变量标准化预测值（即拟合值 $\hat{Y}$ ）的散点图。

因变量标准化预测值指对预测值进行标准化变换的结果，其均值为 0，标准差为 1。学生化残差指残差除以残差的标准误差的点估计值的结果。在 SPSS 中绘制学生化残差与因变量标准化预测值的散点图，可在回归过程中改变默认设置实现。其步骤如下：

(1) 在回归分析主对话框中单击[Plots]按钮，出现如所示的二级对话框，把*SRESID（指学生化残差 Studentized Residual）选入[Y]作为纵轴，把*ZPRED（指标准化预测值 Standardized Predicted Value）选入[X]作为横轴。



(2) 输出结果中将给出学生化残差与标准化预测值的散点图。若学生化残差分布很不均匀, 则说明存在异方差。

(二) 样本分段比较法

这种方法是由 Goldfeld 和 Quandt 于 1972 年提出的, 又称为 Goldfeld-Quandt 检验法(戈德菲尔德—七匡检验法, 简称 G-Q 检验)。这种检验方法以 F 检验为基础, 适用于样本容量较大、异方差明显递增或递减的情况。其基本思想是把样本按某个解释变量的大小顺序排列, 并将样本分成两段; 然后各段分别用普通最小二乘法估计模型, 然后利用两段样本的残差均方之比值构造统计量进行异方差检验。

具体步骤如下:

1、将 n 个样本观测值 (X_i, Y_i) ($i=1, 2, \dots, n$) 按解释变量观察值 X_i 的大小顺序排列。

在 SPSS 中选择菜单[Data]⇒[Sort Cases], 显示如所示的对话框。把 X 选入[Sort by]列表作为排序变量, 并单击[OK]将进行排序。

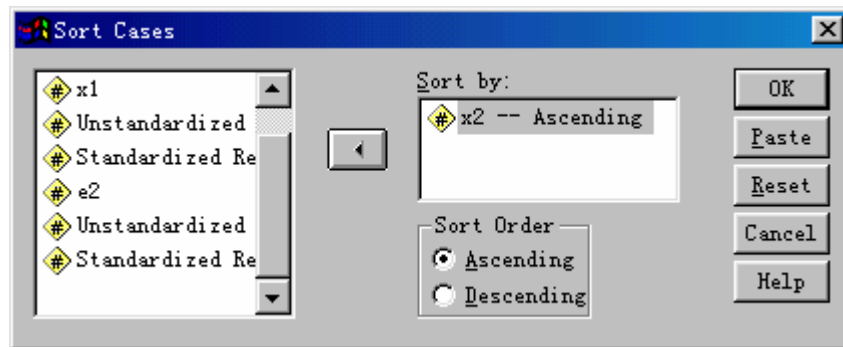


图 8-98 排序

2、将序列中间的 $C=n/4$ 个观测值剔除, 并将剩下的观测值划分为大小相同的两个子样。每个子样的样本容量均为 $(n-C)/2$ 。

3、对每个子样分别求出回归方程, 并计算各自的残差平方和。用 RSS_1 表示对应 X_i 较大值的子样的残差平方和, 用 RSS_2 表示对应 X_i 较小值的子样的残差平方和。它们的自由度均为 $\frac{n-C}{2} - p - 1$, p 为模型中解释变量的个数。

在 SPSS 中选择菜单[Data]⇒[Select Cases], 显示[Select Cases]主对话框。选择[Base on time or case range]作为筛选观测值条件, 并单击该选项下的[Range]按钮, 并在出现的对话框中输入观测值范围[1 至 $(n-C)/2$]。返回主对话框单击[OK]。可以看到, 指定范围之外的观测值所对应的行标签增加 “/” 表示暂时划去这些观测值。

对第一个子样进行回归分析, 在输出结果中的方差分析表中, 可找到残差平方和

[Residual Sum of Squares]，并记下该值。

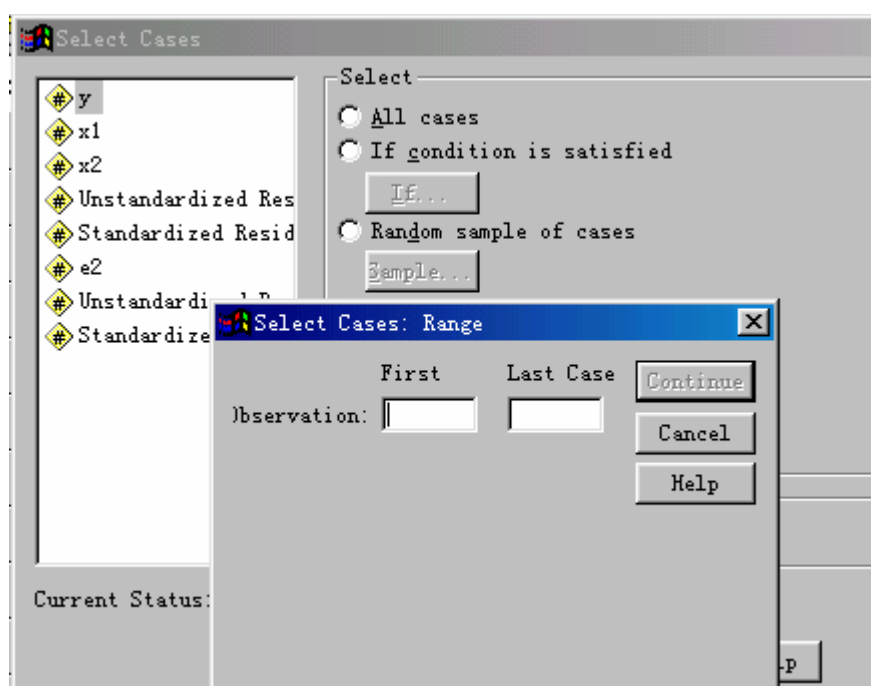


图 8-99

对第二个子样进行类似的操作，记下其残差平方和 RSS2。

4、提出假设。

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

σ_1^2 和 σ_2^2 分别为与两个子样对应的随机扰动项方差。

5、构造统计量

$$F = \frac{RSS2 / (\frac{n-C}{2} - p - 1)}{RSS1 / (\frac{n-C}{2} - p - 1)}$$

当零假设成立时，该统计量服从第一自由度和第二自由度均为 $(\frac{n-C}{2} - p - 1)$ 的 F 分布。

6、根据给定的显著性水平 α 查表得临界值 $F_\alpha(\frac{n-C}{2} - p - 1, \frac{n-C}{2} - p - 1)$ 。

若 $F > F_\alpha(\frac{n-C}{2} - p - 1, \frac{n-C}{2} - p - 1)$ ，则拒绝同方差的零假设，认为存在异方差，而且是递增异方差；否则，认为不存在异方差。

(二) 残差回归检验法

这种方法是用模型普通最小二乘法估计的残差或其绝对值与平方作为被解释变量，建立各种回归方程，然后通过检验回归参数是否显著为 0，来判断模型的随机扰动项是否有某种变动规律，以确定异方差是否存在。该方法的优点是可以近似地给出异方差 $\sigma_i^2 = f(X_i)$ 的具体形式。该方法具体有：

1、Glejser (格里奇) 检验法

这是 Glejser 于 1969 年提出的方法,用残差的绝对值 $|e_i|$ 对每个解释变量建立各种回归模型,如

$$|e| = \gamma_0 + \gamma_1 X + u, |e| = \gamma_0 + \gamma_1 \sqrt{X} + u, |e| = \gamma_0 + \gamma_1 1/X + u \text{ 等等,并检验回归参数 } \gamma_1 \text{ 是否为}$$

零。如果这里的回归参数 γ_1 显著不为 0,说明存在异方差;否则不存在异方差。

在 SPSS 中的操作步骤为:

(1) 选择主菜单[Analyze]=>[Regression]=>[Linear],选择因变量和自变量。并单击[Save]在二级对话框[Linear Regression: Save]中选择[Unstandardized Residuals],保存残差为一新变量(一般为 res_1)。

(2) 选择[Transform]=>[Compute],把残差的绝对值存为另一新变量(如 abse)。类似地,分别生成 X 的各种形式。

(3) 把残差的绝对值变量(如 abse)对 X 的各种形式进行回归,并根据输出结果判断是否存在异方差。

2、Park (帕克) 检验法

该方法与 Glejser 基本思想相同,不同的是 Park 认为方差的形式为

$$\sigma_i^2 = \sigma^2 X_i^r e^{u_i} \quad (8-95)$$

对上式两边取对数,得

$$\ln \sigma_i^2 = \ln \sigma^2 + r \ln X_i + u_i \quad (8-96)$$

由于 σ_i^2 未知,用 e_i^2 近似替代,上式可写成

$$\ln e_i^2 = \ln \sigma^2 + r \ln X_i + u_i \quad (8-97)$$

先求残差,然后利用残差平方 e_i^2 ,求出 $\ln e_i^2$ 对 $\ln X_i$ 的回归方程,最后对回归方程

作统计检验,如果通过显著性检验,则说明存在异方差,否则不存在异方差。

3、White (怀特) 检验法

这是 White 在 1980 年提出的一种方法。它用残差平方 e^2 对所有解释变量及其平方项和交叉乘积项 $X_1, X_2, \dots, X_1^2, X_2^2, \dots, X_1 X_2, \dots$ 进行线性回归,并检验各回归系数是否显著为 0。

三、异方差的处理

由于异方差存在时普通最小二乘估计量是非有效的,所以对于已经检验出存在异方差的回归模型,就不应再直接采用普通最小二乘法来估计模型的参数。常用于处理异方差问题的方法是采用加权最小二乘法(WLS, Weighted Least Squares)。

若已知模型中扰动项的方差与某个变量成比例,如 $\sigma_i^2 = \text{Var}(\varepsilon_i) = \sigma^2 L_i^2$,其中 L_i 已知,则可用 L_i 去除模型的两边。例如,将一元线性回归模型 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 两边同除以 L_i 得:

$$\frac{Y_i}{L_i} = \beta_0 \frac{1}{L_i} + \beta_1 \frac{X_i}{L_i} + \frac{\varepsilon_i}{L_i} \quad (8-98)$$

记 $u_i = \frac{\varepsilon_i}{L_i}$,则有 $\text{Var}(u_i) = \sigma^2$,可见 u_i 具有同方差性,从而可对变换后的模型使用普通

最小二乘法来估计其参数。显然,变换后的模型是 $\frac{Y_i}{L_i}$ 对 $\frac{1}{L_i}$ 和 $\frac{X_i}{L_i}$ 的一个没有常数项的二元线性回归模型。这种方法实际上就是用 $\frac{1}{L_i}$ 对第 i 个样本观测值或残差进行加权,然后再采用普通最小二乘法估计模型的参数,所以称为加权最小二乘法。

对于异方差形式为 $Var(\varepsilon_i) = \sigma^2 X_i$ 、 $Var(\varepsilon_i) = \sigma^2 X_i^2$ 的回归模型,可分别用 $W_i = \frac{1}{\sqrt{X_i}}$ 、 $W_i = \frac{1}{X_i}$ 作为权数进行 WLS 估计。一般地,权数可以是任一变化趋势与异方差的趋势相反的变量,例如可取 $W_i = \frac{1}{Y_i}$ 等。

SPSS 中进行 WLS 估计的步骤如下:

产生一个权数变量(如 W),打开线性回归分析主对话框后,单击[WLS]按钮,展开 WLS 赋权框,把权数变量选入[WLS Weight]框即可。

第九节 自相关

一、自相关的产生与后果

自相关(Autocorrelation)是对随机扰动项之间相互独立假定的违背,指扰动项序列相邻期之间不是随机独立而是存在相关关系,又称为序列相关。自相关主要表现在时间序列中。为明确起见,用 $t, t-1, t-2, \dots$ 表示观测数据的时期,作为扰动项的下标。因此对于线性回归模型:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} \cdots + \beta_p X_{pt} + \varepsilon_t \quad (8-99)$$

自相关可表示为

$$Cov(\varepsilon_t, \varepsilon_{t-i}) \neq 0 \quad (i = 1, 2, \dots, s)。$$

自相关的主要来源于以下几个方面:

1、经济惯性。自相关主要产生于时间序列样本。由于经济发展存在一定的趋势,形成惯性,许多经济变量前后期总是相互关联的。相邻观测值之间或多或少有一定的联系。例如,企业的第 t 期产量总是与第 $t-1$ 期,第 $t-2$ 期, ... 的产量密切相关。这样,在建立回归模型时,扰动项将是自相关的。

2、扰动项序列本身的自相关。一些通常认为是随机干扰因素,如自然灾害、战争、政策执行偏误等,其影响可能持续几个时期,从而形成扰动项的自相关。

3、遗漏重要变量时会导致序列的自相关。在回归分析的构模过程中,如果忽略了一个或几个重要的变量,而这些遗漏的重要变量在时间顺序上的影响是正相关的,回归模型中的误差项就会具有明显的正相关,这是因为主差包含了遗漏变量的影响。例如,根据有关数据建立家庭居民消费模型时,可支配收入是一个重要的变量,它对家庭的消费产生重要的影响,如果仅考虑家庭财富等其它变量而把这个重要变量漏掉了,就可能使得扰动项自相关。

4、模型的数学形式设定不当也可能引起自相关。例如,某经济问题的回归模型是

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

扰动项 ε 是无自相关的。但如果实际采用模型形式为

$$Y = \beta_0 + \beta_1 X + u$$

则扰动项 $u = \beta_2 X^2 + \varepsilon$ 随 X^2 系统变化，导致自相关。

5、数据处理造成自相关。在实际研究中，往往需要对原始数据进行处理。有时无自相关的原始数据经整理后反而产生自相关。例如在具有季节性时序资料的建模中，常常要对数据做修匀处理以达到消除季节性的影响，如果采用了不合适的差分变换，就会带来自相关问题。

当扰动项存在自相关时，就违背了标准线性回归模型的基本假定，如果仍直接用 OLS 法估计参数，将产生一系列后果。其主要影响后果有：

- 1、参数的估计量的方差增大，不再具有最小方差性。
- 2、导致常用的 F 检验和 t 检验失效。
- 3、如果不加处理地运用 OLS 法估计模型参数，用此模型进行预测时会带来较大的方差的错误的解释。

二、自相关的检验

用于检验扰动项是否存在自相关的方法主要有：

(一) 图示法

图示法是一种直观的检验方法，是把给定的回归模型直接用 OLS 法估计参数，求出残差项 e ， e 作为扰动项 ε 的真实值的估计值，再绘出残差 e 的散点图，根据 e 的相关性来判断扰动项 ε 的自相关性。残差 e 的散点图通常有两种绘制方式：绘制 $e_t - e_{t-1}$ 的散点图和按时间顺序绘制残差的散点图。在 $e_t - e_{t-1}$ 散点图中，如果大部分点落在第一、三象限，表明扰动项 ε 存在正自相关，如图 8-100(A)所示；如果大多数点落在第二、四象限，表明扰动项 ε 存在负自相关，如图 8-100(B)所示。在 $e_t - t$ 散点图中，如果 e_t 随时间变化并不频繁地改变符号，而是相几个正的 e_t 后面跟着几个负的 e_t ，表明扰动项存在正自相关，如图 8-100(C)所示；如果 e_t 随着时间的逐次变化频繁地改变符号，表明扰动项存在负自相关，如图 8-100(D)所示。

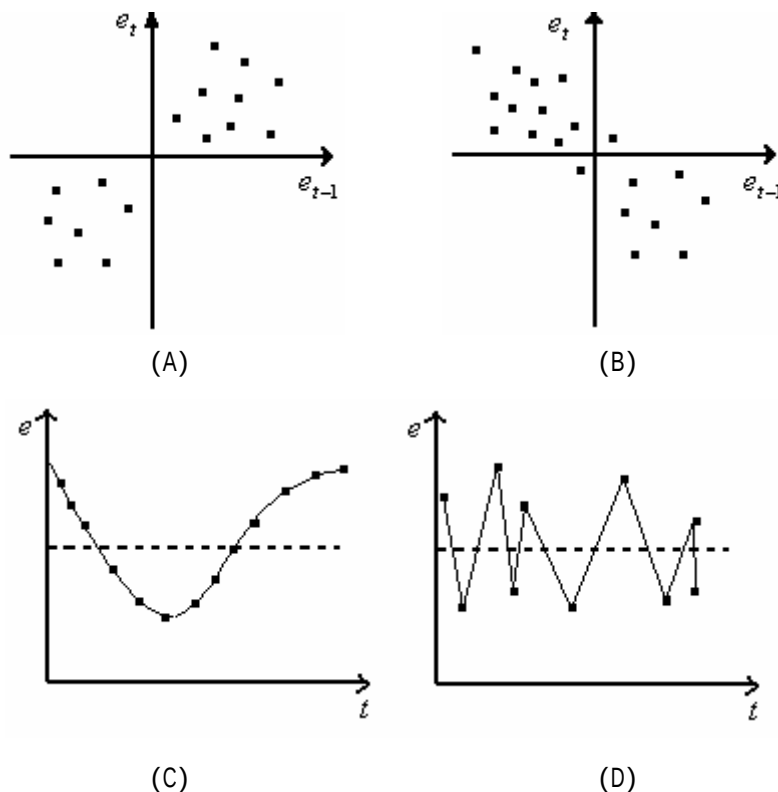


图 8-100 用于判断自相关的散点图

(二) D-W 检验 (Durbin-Watson 杜宾-瓦特森检验)

D-W 检验是 Durbin 和 Watson 于 1951 年提出的一种自相关检验方法。它只适用于扰动项的形式为

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (\rho \text{ 为自相关系数}) \quad (8-100)$$

的一阶自相关问题。这种方法是最常用的一阶自相关检验方法。其检验步骤如下：

1、提出假设。

$H_0: \rho = 0$ ，即扰动项不存在一阶自相关

$H_1: \rho \neq 0$ ，即扰动项存在一阶自相关

2、构造统计量。定义 DW 统计量为

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2} \quad (8-101)$$

对于大样本，可以把 DW 统计量写为

$$DW = 2(1 - \hat{\rho}) \quad (8-102)$$

其中 $\hat{\rho}$ 为自相关系数 ρ 的估计。因为 $|\hat{\rho}| \leq 1$ ，所以 DW 检验统计的值域为 $0 \leq DW \leq 4$ 。而且由式 (8-102) 可以看出：

当 $\hat{\rho} \rightarrow 0$ 时， $DW \rightarrow 2$ ；

当 $\hat{\rho} \rightarrow 1$ 时， $DW \rightarrow 0$ ；

当 $\hat{\rho} \rightarrow -1$ 时， $DW \rightarrow 4$ ；

3、判断。根据样本容量 n 和解释变量的数目 p 查 DW 分布表，得下临界值 D_L 和上临界值 D_U ，并依下列准则判断扰动项的自相关情形。

(1) 如果 $0 < DW < D_L$ ，则拒绝零假设，扰动项存在一阶正自相关。DW 越接近于 0，正自相关性越强。

(2) 如果 $D_L < DW < D_U$ ，则无法判断是否有自相关。

(3) 如果 $D_U < DW < 4 - D_U$ ，则接受零假设，扰动项不存在一阶正自相关。DW 越接近 2，判断无自相关性把握越大。

(4) 如果 $4 - D_U < DW < 4 - D_L$ ，则无法判断是否有自相关。

(5) 如果 $4 - D_L < DW < 4$ ，则拒绝零假设，扰动项存在一阶负自相关。DW 越接近于 4，负自相关性越强。

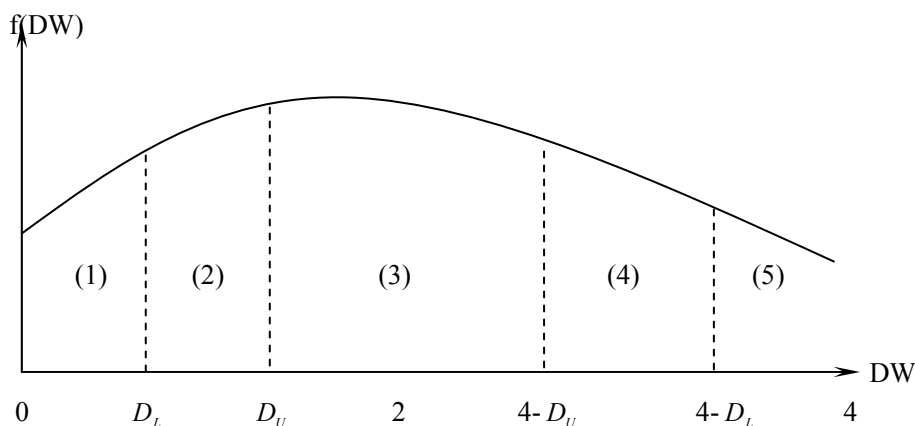


图 8-101

D-W 检验虽然是最常用的自相关检验方法，但在应用上有局限性。由于 DW 统计量的精确分布未知，Durbin 和 Watson 是在一定条件下用 Beta 分布近似。当这些条件不满足时，DW 检验就失效。这些局限性主要表现在以下几个方面：

- (1) D-W 检验只适合一阶自相关，不适合高阶自相关；
- (2) D-W 检验不适用于解释变量与扰动项相关的模型；
- (3) D-W 检验存在两个不能确定的区域，一旦 DW 值落入这个区域，就无法判断。这时只有增大样本容量或选取其他方法。

SPSS 可以输出 DW 值。在图 8-82 所示的对话框中单击[Statistics]按钮，打开[Linear Regression: Statistics]子对话框，在[Residuals]栏中选择[Durbin-Watson]选项，单击[Continue]按钮返回再单击主对话框[OK]按钮。这样在 SPSS 的输出结果中将包含 DW 值，根据该值就可以进行检验。

三、自相关的处理

产生自相关的原因有多种。如果自相关是由于遗漏重要变量，或者设定的模型形式错误，那么就应引入新变量或者修改模型形式。排除了上述原因后，经检验仍然存在自相关，就必须采用一定的方法来处理。其基本思想是通过差分变换，对原始数据进行整理，变自相关为无自相关。常见的方法有广义差分法等。

设线性回归模型

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \varepsilon_t \quad (8-103)$$

存在一阶自相关，

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

其中 u_t 满足基本假定。

将 (8-103) 滞后一期，并且两边同乘以自相关系数 ρ ，得

$$\rho Y_{t-1} = \rho \beta_0 + \beta_1 \rho X_{1(t-1)} + \beta_2 \rho X_{2(t-1)} + \cdots + \beta_p \rho X_{p(t-1)} + \rho \varepsilon_{t-1} \quad (8-104)$$

用式 (8-103) 减去式 (8-104)，可得广义差分模型为：

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_{1t} - \rho X_{1(t-1)}) + \cdots + \beta_p (X_{pt} - \rho X_{p(t-1)}) + \varepsilon_t - \rho \varepsilon_{t-1} \quad (-105)$$

记各变量的广义差分分别为：

$$\begin{aligned} Y_t^* &= Y_t - \rho Y_{t-1}, \quad t = 2, 3, \cdots, n \\ X_{it}^* &= X_{it} - \rho X_{i(t-1)}, \quad i = 1, 2, \cdots, p \\ u_t &= \varepsilon_t - \rho \varepsilon_{t-1}, \quad t = 2, 3, \cdots, n \end{aligned}$$

上述广义差分模型 (-105) 可改写为：

$$Y_t^* = \beta_0 (1 - \rho) + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \cdots + \beta_p X_{pt}^* + u_t \quad (8-106)$$

由 u_t 满足基本假定，所以对 (8-106) 可采用 OLS 法进行估计。

实践中，扰动项的自相关系数 ρ 往往是未知的，需要用某种方法对 ρ 进行估计。为此，可采用逐步迭代的方法进行，其步骤为：

- 1、应用 OLS 法对原模型 (8-103) 进行估计，计算出各残差 $e_t = Y_t - \hat{Y}_t$ ， $t=1, 2, \dots, n$ 。
- 2、利用公式 $\hat{\rho} = \frac{\sum_{t=2}^n (e_t e_{t-1})}{\sum_{t=2}^n e_t^2}$ 计算出自相关系数 ρ 的估计值。
- 3、利用估计出的自相关系数值 $\hat{\rho}$ 对被解释变量和解释变量的样本观测值进行广义差分

变换, 得出 Y_i^* 和 X_{it}^* , $i=1, 2, \dots, p, t=1, 2, \dots, n$, 并利用 OLS 法估计出广义差分回归模型 (8-106)。

4、将利用广义差分回归模型 (8-106) 估计出的参数代入原回归模型 (8-103), 重新计算残差, 并利用此残差重新估计 ρ 的值, 然后利用 ρ 的新估计值重新进行广义差分变换, 估计出新的广义差分模型。

此迭代过程, 可只迭代两次, 也可迭代多次, 直到 ρ 的估计值收敛为止。

第十节 回归模型的应用

回归模型的应用主要有预测、结构分析和政策评价等三个方面。系统的结构分析和政策评价方法需要较多的专业知识、数学知识和计算机知识, 因此这里仅介绍预测的过程。

根据样本对总体进行预测时, 通常有两类预测问题: 一是预测给定自变量 X 的某一特定值条件下因变量 Y_0 的点预测和区间预测; 二是在同一条件下回归方程估计值 $\hat{Y}_0$ 的点预测和区间预测。两种点预测的方法是完全相同的, 都是令自变量为某特定值, 并代入样本回归方程来计算其点预测值。但是两种区间预测是不同的, 因为在回归分析中后者是关于所有具备同一条件的个体的均值预测, 而前者只是关于某一个具备这一条件的个体的预测。由于多项正负误差之间相互抵消, 平均值预测总比单个值预测更为精确, 即后者比前者具有更窄的预测区间。通常, 我们称前者为特定条件下单个个体 Y_0 的区间预测, 称后者为特定条件下回归平面 $\hat{Y}_0$ 的区间预测。

一、点预测

Y_0 和 $\hat{Y}_0$ 的点预测过程是相同的, 都是特定条件的取值 $X_{01}, X_{02}, \dots, X_{0p}$, 由回归方程求出点预测值:

$$\hat{Y}_0 = b_0 + b_1 X_{01} + b_2 X_{02} + \dots + b_p X_{0p}$$

例如, 例 8-1 是关于消费支出 Y 对可支配收入 X 回归的模型, 其样本回归方程为:

$$\hat{Y} = 0.607 + 0.542X$$

假定我们需要预测可支配收入为 7 千元条件下 Y_0 和 $\hat{Y}_0$ 的点预测值。因为两者的点预测相同, 只要把两个给定值代入上面的方程, 便可得到它们的预测值:

$$\begin{aligned}\hat{Y} &= 0.607 + 0.542X \\ &= 0.607 + 0.542 \times 7 \\ &= 4.40\end{aligned}$$

二、区间预测

区间预测是在点预测的基础上, 计算相应抽样分布方差和标准误的估计, 并据此构造其预测区间。这时, 由于 Y_0 和 $\hat{Y}_0$ 的抽样分布方差不同, 所以估计公式也有所不同。

在给定显著性水平 α 条件下 Y_0 的预测区间为:

$$\left[\hat{Y}_0 - t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + 1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}}, \hat{Y}_0 + t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + 1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}} \right]$$

即

$$\left[\hat{Y}_0 - t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + 1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}}, \hat{Y}_0 + t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + 1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}} \right]$$

其中, C_{ij} 为矩阵 $C = (X'X)^{-1}$ 中第 i 行第 j 列的元素。

$\hat{Y}_0$ 的预测区间为:

$$\left[\hat{Y}_0 - t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}}, \hat{Y}_0 + t_{\frac{\alpha}{2}} \cdot S \cdot \sqrt{1/n + \sum_{i=1}^p \sum_{j=1}^p (X_{0i} - \bar{X}_i)(X_{0j} - \bar{X}_j)C_{ij}} \right]$$

在 SPSS 中, 预测是与估计检验同时进行的, 其操作步骤为:

- (1) 在原来的数据文件中变量 X 的下方输入给定的 $X_0=7$, 相应的变量 Y 将产生缺失值;
- (2) 选择主菜单[Analyze]⇒[Regression]⇒[Linear](如图 8-74所示), 打开[Linear Regression]主对话框(如图 8-75所示)。在左边列表框中选定变量 Y, 单击按钮, 使之进入[Dependent]框, 选定变量 X, 单击按钮使之进入[Independent(s)]框;
- (3) 单击主对话框中的[Save...], 打开[Linear Regression: Save]子对话框。选择[Predicted Values (预测值)]栏中的[Unstandardized (非标准化预测值)]选项把非标准化预测值保存在一个新变量中, 选择[Prediction Intervals (预测区间)]栏中的[Mean]和[Individual]选项, 并在下面的[Confidence Interval]框中输入置信度, 本例采用默认值 95%。选择了[Mean]和[Individual]选项将分别保存所输入的置信度条件下 $\hat{Y}_0$ 和 Y_0 在的预测区间上下限。
- (4) 单击[OK]按钮, 除了输出如表 8-39所示的分析结果外, 还将在数据文件中生成 pre_1、lmci_1、umci_1、lici_1 和 uici_1 等变量(如表 8-45所示)。其中, pre_1 变量保存预测值, lmci_1 和 umci_1 分别保存 $\hat{Y}_0$ 预测区间的下限和上限, lici_1 和 uici_1 分别保存 Y_0 预测区间的下限和上限。

表 8-45 预测结果输出

Y	X	pre_1	lmci_1	umci_1	lici_1	uici_1
1.60	2.00	1.69091	1.43100	1.95082	1.17798	2.20384
2.00	2.50	1.96182	1.74139	2.18225	1.46772	2.45592
2.30	3.00	2.23273	2.04734	2.41811	1.75323	2.71222
2.40	3.50	2.50364	2.34588	2.66139	2.03413	2.97314
3.00	4.00	2.77455	2.63261	2.91649	2.31012	3.23897
3.20	4.50	3.04545	2.90351	3.18739	2.58103	3.50988
3.10	5.00	3.31636	3.15861	3.47412	2.84686	3.78587
3.50	5.50	3.58727	3.40189	3.77266	3.10778	4.06677

3.60	6.00	3.85818	3.63775	4.07861	3.36408	4.35228
4.40	6.50	4.12909	3.86918	4.38900	3.61616	4.64202
.	7.00	4.40000	4.09792	4.70208	3.86446	4.93554

第十一节 案例 1:我国经济增长持续性的实证研究

建国 50 年来,我国的经济发展取得了长足的进展,尤其是改革开放以来的近 20 年,中国经济发展的成就更是举世瞩目,中国经济现已步入快速稳定发展的上升通道。但是,我国经济在高速增长的同时,经济增长中的一些问题也逐渐暴露出来,其中尤其突出的问题就是经济增长质量问题。在此,我们就经济增长持续性方面作一实证研究。

从大的经济社会环境上看,我国的经济增长大体上经历了两个不同的经济环境,一个是建国初期至改革开放,另一个则是改革开放至今。如果我们将国民经济某一绝对指标(如 GNP 或国民收入)的变动情况符合某一变动趋势这一现象称为“经济增长路径”的话,那么,用两个不同时期的经济增长路径及其相应的指标就可以考察我国经济增长的持续性特征。

1952~1977 年,我国国民收入按不变价格计算的变动情况如表 8-46 所示。

表 8-46 1952~1977 年我国国民收入(按 1952 年价)单位:亿元

年次 (t)	年份 (year)	国民收入 (Y)	年次 (t)	年份 (year)	国民收入 (Y)
1	1952	589.0	14	1965	1163.4
2	1953	671.5	15	1966	1361.2
3	1954	710.4	16	1967	1263.2
4	1955	755.9	17	1968	1181.1
5	1956	862.4	18	1969	1409.0
6	1957	901.3	19	1970	1737.3
7	1958	1099.5	20	1971	1859.0
8	1959	1189.7	21	1972	1912.9
9	1960	1173.0	22	1973	2071.6
10	1961	824.7	23	1974	2094.4
11	1962	771.0	24	1975	2268.3
12	1963	835.5	25	1976	2207.0
13	1964	994.4	26	1977	2379.2

资料来源:《中国统计年鉴》(1993),第 33、35 页。

利用表 8-46 的资料用 SPSS 作回归分析,结果输出如下:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.927	.860	.854	214.316

a. Predictors: (Constant), T

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6747125.263	1	6747125.263	146.896	.000
	Residual	1102354.324	24	45931.430		
	Total	7849479.587	25			

a Predictors: (Constant), T

b Dependent Variable: 国民收入Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	401.739	86.547		4.642	.000
	T	67.922	5.604	.927	12.120	.000

a Dependent Variable: 国民收入Y

从输出结果可得

$$\hat{Y} = 401.739 + 67.922 t \quad (8-107)$$

通过参数显著性检验。

回归方程(8-107)表明,建国到1977年间,我国的经济增长保持着一个斜率为67.922,截距为401.739的上升路径。意味着国民收入增加额保持在一个大约平均67.922亿元的水平上。经计算,这一运行轨迹的变异系数为15.51%。

改革开放以来,我国的经济发展进入了一个新的时期,1978—1995年,国民生产总值(即GNP)的变动情况如表8-47所示。

表 8-47 1978 ~ 1995 年我国 GNP (按 1978 年价) 单位: 亿元

年次 (t)	年份 (year)	GNP (Y)	年次 (t)	年份 (year)	GNP (Y)
1	1978	3624.1	10	1987	8481.8
2	1979	3899.5	11	1988	9440.2
3	1980	4204.0	12	1989	9836.7
4	1981	4392.4	13	1990	10249.9
5	1982	4776.6	14	1991	11182.6
6	1983	5273.1	15	1992	12759.3
7	1984	6048.6	16	1993	14430.8
8	1985	7012.6	17	1994	16249.1
9	1986	7607.0	18	1995	17711.5

资料来源:《中国统计年鉴》(1996),第42页。

利用所给的资料用 SPSS 进行回归,结果如下:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.971	.943	.940	1071.128

a Predictors: (Constant), T

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	305380978.687	1	305380978.687	266.170	.000
	Residual	18357056.151	16	1147316.009		

	Total	323738034.838	17			
--	-------	---------------	----	--	--	--

a Predictors: (Constant), T

b Dependent Variable: GNP Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1190.017	526.740		2.259	.038
	T	793.915	48.663	.971	16.315	.000

a Dependent Variable:GNP

从输出结果可得：

$$\hat{Y} = 1190.017 + 793.915 t \quad (8-108)$$

从回归方程 (8-108) 可以看出, 改革开放以来, 我国经济增长路径发生了重大变化, 它不仅表现在回归方程 (8-108) 的截距明显上移, 而且其斜率也明显增大, 斜率值 793.915 表明, 进入改革开放后, 我国的 GNP 平均增加额达到了 793.915 亿元的水平。回归方程截距的大幅上移和斜率的急速上升, 一方面说明我国的整体国力显著增强, 另一方面说明在整体国力增强的同时, 经济增长出现了加快的势头。经济增长自此进入一条新的增长路径, 并且经济增长路径的变异系数也有所下降, 其值为 12.63%, 表明经济增长持续性比第一个阶段有所改善。

比较 (8-107) 与 (8-108) 两个回归方程, 可以看出两者存在着重大的差别。如图 8-102 所示。

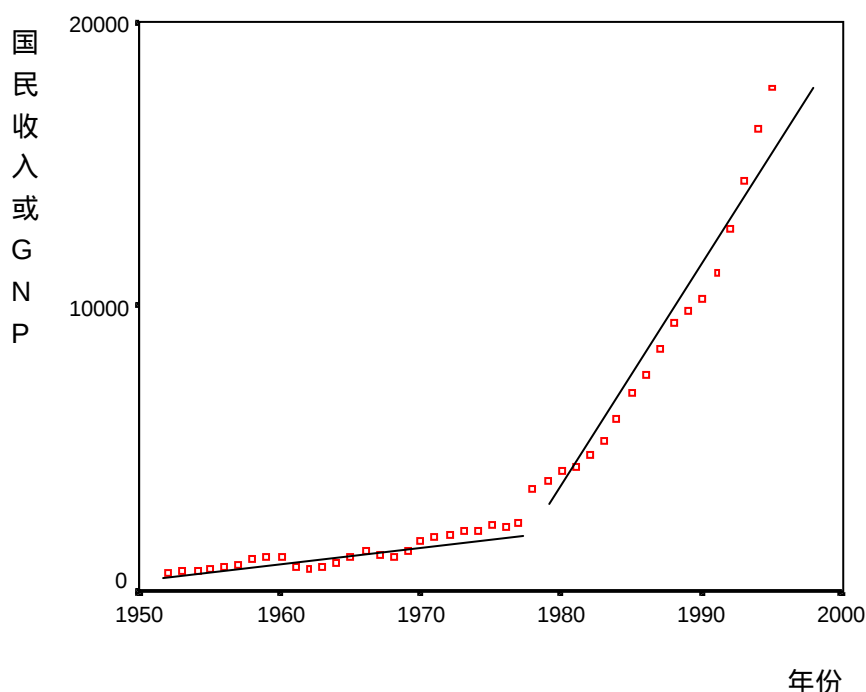


图 8-102

首先, 从形态上看, 显然第一个阶段的方程表现得较为平坦, 第二个阶段则表现得更为陡峭。如果这种形态不被人为地破坏, 经济增长的这一上升路径至少还可以保持若干年。对照罗斯托经济起飞阶段的标准, 可以说目前我国经济发展已经进入了起飞阶段。其次, 从拟

合优度 R^2 上看, 虽然两个回归方程的 R^2 之值都在 0.8 以上, 总体上说, 拟合情况都较好, 但是, 两者之间的差别也是明显的, 第一阶段 R^2 值为 0.86, 而第二阶段的 R^2 值则上升到 0.943, 两者相差 0.083 即 9.65%。这表明第二阶段方程总的拟合情况明显优于第一阶段, 也就是说, 国民收入或 GNP 的实际值总体对两条回归直线的偏离情况第二阶段要比第一阶段好得多, 即也意味着改革开放后经济增长的持续性明显强于改革开放以前。

值得注意的是, 虽然改革开放以来经济增长的持续性比改革开放前有了明显的改善, 但是, 实际 GNP 与 GNP 上升通道之间的绝对离差仍平均高达 222 亿元, 占 GNP 平均值的 2.68%, 意味着经济增长相邻年份的震幅平均可能会高达 5.36%, 未来我国的平均经济增长率即使按 10% 计算, 经济增长率的震幅也将超过经济增长率的 50%。因此, 当前我国经济增长状况并不特别乐观。我国经济增长尚处在粗放型阶段, 经济增长方式的转变已成为迫在眉睫的当务之急。必须加快经济体制转变的步伐, 使市场机制真正成为发挥在资源配置中基础地位的作用, 否则粗放型的增长方式难以改变。

第十二节 案例 2：中德人口老龄化水平之比较

1992 年, 第 47 届联合国大会通过了《老龄问题宣言》, 正式确定 1999 年为“国际老人年”。人口老龄化已成为全球化的趋势和全世界瞩目的焦点。在西欧诸发达国家中, 人口老龄化现象已有近百年的历史。德国自十九世纪末就初见端倪而历经几十年的老龄化进程, 为正处于人口转型期的我国提供了宝贵的经验和教训。

我国对人口老龄化的研究正在逐步走向系统化。但不难发现, 我国的人口老龄化研究还有以下不足之处: 其一, 对于老龄化现象的成因研究尚以定性分析为主, 较少使用统计模型对各个因素对老龄化的影响程度进行定量考察。其二, 老龄化的研究多注重从专业化套路化的人口学角度, 而人口老龄化问题应该说并非只是人口学领域的研究范畴, 政治学、经济学、社会学、医学等多学科都将在此问题上与人口学形成交融和撞击。其三, 未将中国和其它西方国家的影响人口老龄化的经济学、社会学、医学等诸因素及特征作过系统地数量分析和比较, 也未深入地研讨过德国等发达国家的养老保障制度的内核对于我国的实际借鉴意义。

下面从中德两国的人口老龄化水平入手, 运用统计模型分析两国的老龄化现象成因, 进而对两国老龄化的特征作比较。在此基础上, 以德国的社会养老保障为鉴, 对我国的养老模式之选择和养老保障制度之变革提出一孔之见。

一、人口老龄化的内涵及中德两国人口老龄化的历史发展评述

联合国曾于 1956 年把 65 岁作为老年人口起始年龄对人口年龄结构类型提出划分标准, 如表 8-48 所示:

表 8-48 人口年龄结构类型划分标准

人口年龄结构类型	年轻型	成年型	老年型
65 岁以上人口占总人口的比重(%)	< 4	4~7	≥ 7

依照此标准, 对一个国家而言, 只要其人口年龄结构中 65 岁以上人口超过总人口的 7% 时, 就可认为其已进入老龄化社会。

德国在 19 世纪后半叶就开始了其漫长的人口老龄化进程, 1930 年其 65 岁以上人口已占总人口数的 7%, 这标志着德国已率先迈入了老年型社会。此后老年人的比重不断上升, 在 1930 年至 1975 年长达 45 年的历程中, 德国 65 岁以上人口比重已从 7% 跃升至 14%。据

参见杨绮、米红:《中德两国人口老龄化水平与社会养老保障制度的比较研究》,《人口学刊》,1999 年第 6 期。

《中国人口统计年鉴》(1995)所载,德国统一前西德 1990 年 65 岁以上人口比重达 15.4%,统一后全德 1991 年 65 岁以上人口比重达 15.9%。据联合国中等方案预测,到 2025 年,德国 65 岁以上人口将达总人口数的 21.77%,到 2050 年,将达 29.16%。

我国的人口老龄化起步较晚,正处于人口转变的中期。建国后死亡率的下降和出生率的稳定曾导致人口一度出现年轻化,70 年代的计划生育政策所造成的出生率的下降为人口老龄化的出现奠定了潜在的基础。1982 年 65 岁以上人口仅占总人口的 4.88%,到 1996 年已达到 6.94%。根据联合国中等预测方案,到 2005 年我国的 65 岁以上人口比重将达 7.28%,即已成为典型的老年型社会。在 2025 年和 2050 年,预计我国人口年龄结构中 65 岁以上人口将达 12.21%和 19.23%。

从现状和未来的发展看,我国的人口老龄化程度大大小于德国的程度。然而中国将不可避免地由成年型过渡为老年型社会,正如德国所经历过的一样。因而,借鉴和研究德国的老龄化水平和养老保障对于我国的老龄工作有较强的实践意义。

二、中德两国人口老龄化现象和社会养老保障制度的特点对比

中德两国人口老龄化现象各有其特点:

(一) 老年人口绝对数规模的对比

就现状而言,我国的老年人口占总人口的比重远小于早已进入老龄化社会的德国。但从老年人口的绝对数规模上来看,形势恰恰相反。如表 8-49 所示,我国的老年人口总数远多于德国。我国人口基数大,每年仍以大于 10%的自然增长率递增,人口的膨胀与人口的老龄化相结合的结果,就出现了巨大的老年人群体与较低的老龄化水平的强烈反差。而纵观德国,总人口数少,甚至还出现了负增长,与我国相比,自然它就出现了“高老龄化程度,低老年人总数”。据报道,我国的老年人口约占全世界的 50%。

表 8-49 中德老年人口绝对数规模的对比

国 家	中国	德国	中国	德国
年 份	1982	1980	1990	1987
65 岁以上老人占总人口的比重(%)	4.88	15.46	5.60	15.30
65 岁以上老人总人口数(人)	49275549	9534600	63232361	9347709
自然增长率(‰)	15.68	-1.5	14.39	-1.1

资料来源:《中国统计年鉴》1987;《国际统计年鉴》1996;United nations: The sex and age distribution of the world population, the 1996 revision;网络资料: U.S. bureau of the census, International Data Base

(二) 人口老龄化发展速度的对比

出生率的变化将显著影响人口老龄化的发展速度。在我国,由于 20 多年来计划生育政策的普遍推行,出生率显著下降。自 80 年代进入成年型社会,我国 1990 年第四次人口普查时 65 岁以上人口的比重已高达 5.6%,到 96 年人口变动抽查时就发现 65 岁以上人口的比重已达 6.94%,可以预期我国进入老年型社会前后不足 20 年,其来势之猛,发展之快,可见一斑。相对于我国人口老龄化的突发性,德国则显得“从容不迫”。从 19 世纪后半叶开始老龄化进程,德国约 70 年才进入老年型社会。

(三) 经济状况对人口老龄化的承受力的对比

老龄化水平越高,老年人口越多,退休费、医疗费、福利费的支出自然呈剧增趋势,这客观上要求有坚强的经济后盾。德国能成功地将高水平的老龄化程度对社会各领域的影响冲击程度降到最低,很大程度得力于其雄厚的物质基础。而如前述,我国的人口老龄化与经济发展不同步,经济水平的滞后性大大削弱了其对老龄化的承受能力,使我国老龄化形势雪上

加霜。

（四）人口老龄化水平对经济状况的反作用的对比

老年人的普遍特征是体力下降,但仍拥有丰富的智力资源。我国的科技水平不发达,企业以劳动密集型产业为主,对劳动者的体力能力的需求远胜于对脑力能力的需求。况且由于教育水平的落后,在老年人中受过高等教育,有较高文化素质的人并非多数。即使我国企业多为技术密集型,这些老年人也不可能胜任企业的“智囊”角色。这样,老龄化水平的提高无疑降低了劳动生产率,严重影响了经济水平。而德国科技水平较发达,人口的老龄化与较高的劳动力素质的互补性降低了老龄化对劳动生产率的提高的威胁程度,因而其老龄化对经济的消极影响不如我国显著。

中德两国的人口老龄化特点迥异,适应于人口老龄化发展趋势的社会养老保障制度也因其带有各自不同的政治、经济和人口特征而各成一系。

（一）筹资模式的对比

德国的人口老龄化绝对数规模小,发展速度慢。与此相适应,其养老保险体制的筹资模式采用现收现支模式。德国的养老保险的资金筹措曾先后采用资本积累法(1881—1957年)、分段式收支抵偿法(1957-1969年)、靠拢现收现支法(1969-1977年)、纯粹现收现支法(1977年至今)。1992年养老保险改革中将现收现支法写入了社会法典中,1999年改革中仍坚持法定养老金部分的现收现支模式,并为避免现收现支模式下保险费率的频繁变动,允许波动储备金介于一个月至一个半月支出的区间内。德国以现收现支作为其主要的筹资模式,灵活地适应了其现状。因为德国的养老金水平较高,若转为资本积累筹资模式,考虑已并入的东德地区,又须维持高福利水平,就必然会面临10万亿马克数量级资金的管理问题。且不说这一巨额资本的增值问题,单为规避未来通货膨胀将导致的巨额贬值风险就须苦心经营,可见转为资本积累模式的“机会成本”太高。而现收现支模式虽然无法抵御人口老龄化浪潮冲击的缺陷,但是在人口老龄化绝对数规模不大,发展速度不快的德国暂不会掀起狂澜。正如1999年改革方案中所提出的加大资本积累模式的份额,可以在企业补充保险和个人人寿保险中采用部分资本积累模式,以弥补现收现支模式的不足之处。

我国的养老金筹资模式原为现收现支法,但我国人口老龄化绝对数规模大,发展极为迅速,退休人员日益增多,养老金负担日益沉重,随时会出现下一代人养活上一代人的负担系数过大而入不敷出的现象。人口老龄化的严峻形势决定了我国不能象德国那样乐观地采用现收现支法。而使用资本积累模式除会出现巨额资本的保值增值问题外,也不切合我国的实际经济状况。因为这种方式将造成对已缴纳养老保险者和企业的双重负担。对于收入尚低的劳动者和面临残酷市场竞争的企业不啻是加重了负担。因而,只有采用部分积累模式,对已退休劳动者实行现收现付,对未来劳动者提取一部分养老基金,才能适应我国人口老龄化趋势和养老保障制度的现状。

（二）养老金水平的对比

如前述,德国的经济状况对人口老龄化的承受能力较强,因而其养老保险费率和养老金水平较高。据统计,德国1997年和1998年的法定养老保险的保险费率高达20.3%,工资的附带成本较高。其雄厚的经济基础使其有资本维持较高的养老金支出。但是,人口老龄化向纵深方向的发展和人口寿命的增加,必将导致养老金领取人数和领取年份的增加,即使在当前的经济状况下,能否继续维持原有的高福利水平尚值得怀疑。如果经济状况恶化,高福利的刚性使养老金水平难以大幅度下调,则势同骑虎难下。德国在1999年养老金改革法中为降低过高的养老金水平,在养老金计算公式中加入人口发展因子,同时以“养老金水平保障条款”来保证标准养老金水平。

我国的经济发展水平落后,社会保障水平处于中下等收入国家的上线水平,养老金水平亟待提高。而德国的养老金改革也说明了必须重视福利刚性问题,养老金水平有个“度”的

限制。“由简入奢易，由奢入简难”，一旦过高的养老金水平在经济不景气时无以为继，要降低又会引起养老金享受者的抵触情绪，就必然陷入两难的境地。高福利的政策是不适用于我国的现状的。

三、人口老龄化的成因分析与统计模型的运用

人口老龄化是与人类社会文明化的历程相伴相生的。中国抑或德国，其人口老龄化现象各有其特征，但其正在经历或已经发生的人口老龄化都是近现代社会发展过程所出现的必然人口现象。透过现象看本质，中德两国的老龄化现象的成因不乏相似之处，经济的发展是人口老龄化的原动力，出生率和死亡率的下降是直接原因，而医疗条件的改善和医疗水平的提高也可能在一定程度上推动了人口老龄化的进程。

为了证实上述几个因素对人口老龄化的影响，本文以 65 岁以上人口数占总人口的比重代表老龄化水平，选择人均国内生产总值、每个医生负担的人口数、出生率三个变量，分析它们与老龄化水平之间的关系，借以对中德两国的人口老龄化现象作定量分析。

表 8-50 德国人口老龄化水平及影响因素分析表

年 份 Year	老龄化水平(万 人均国内生产 分之一) Y	总值(美元) X1	出生率(‰) X2	每个医生平均 负担人口数 (人) X3
1950	941	487	**18.10	824
1961	1107	1531	**19.61	670
1970	1317	3158	13.40	561
1980	1546	13216.5	10.10	442
1984	1468	10084.3	9.50	380
1985	1470	10190.0	9.60	377
1987	1530	18130.9	10.50	*364.95
1990	1540	23738.7	11.50	370
1991	1590	19852.5	10.40	*317.06

注：1、资料来源：《国际统计年鉴》(1996)；《世界经济年鉴》(1989、1993、95)；《世界经济统计简编》(1982、1987)；《国外经济统计年鉴》(1949~1976)；《中国人口统计年鉴》(1990—1996)；网络资料：U.S. bureau of the Census, International Data Base

2、*系通过建立“每个医生平均负担人口数”与“年份”之间的线性关系后估计而得（每个医生平均负担人口数 = 年份 * (-11.973) + 24155.3）

**系通过建立“出生率”与“自然增长率”之间的线性关系后估计而得（出生率 = 自然增长率 * 1.076 + 11.753）；“自然增长率”见表 8-51

3、1991 年前系西德资料，1991 年系全德统一后的资料

表 8-51 德国的自然增长率与出生率关系表

年 份	出生率(‰)	自然增长率(‰)
1950	18.10	5.9
1961	19.61	7.3
1970	13.40	1.3
1980	10.10	-1.5
1984	9.50	-1.8
1985	9.60	-2.4
1987	10.50	-1.1
1990	11.50	0

1991	10.40	-1.0
------	-------	------

资料来源：同表 8-50。

表 8-52 中国人口老龄化水平及影响因素分析表

年 份 Year	老龄化程度 (万分之一) Y	人均国内生产 总值(美元)* X1	出生率(‰) X2	每个医生平均 负担人口数 (人)**X3
1982	488	265.52	22.28	778
1985	518	266.56	21.04	749
1988	550	364.25	22.37	687
1989	579	320.34	21.58	656
1990	560	313.03	21.06	649
1991	599	346.04	19.68	651
1992	607	397.74	18.24	648
1994	636	455.10	17.70	637
1995	669	581.32	17.12	631
1996	694	677.98	16.98	631

注：1、本表资料来源：《中国统计年鉴》(1987—1997)；《中国人口统计年鉴》(1992、1993、1995、1996)；《中国人口年鉴》(1990)；United Nations: The sex and age distribution of the world population, the 1996 revision

2.*系按照当年汇率将当年人均国内生产总值人民币数折算为美元数；

**系以当年总人口数除以当年的医生总数

经过 SPSS 软件分析，可得出以下两个方程：

$$\text{德国：} \hat{Y} = 1910.016 + 0.0044X_1 - 14.150X_2 - 0.819X_3 \quad (-109)$$

t 值： 13.823 1.030 -1.168 -2.339

p 值： 0.000 0.350 0.295 0.066

容限： 0.277 0.166 0.097

VIF： 3.605 6.034 10.265

$R^2=0.945$ $F=46.60, p=0.000$

其中：Y——德国老龄化水平

X_1 ——德国人均国内生产总值

X_2 ——德国出生率

X_3 ——德国每个医生平均负担人口数

$$\text{中国：} \hat{Y} = 991.081 + 0.212X_1 - 8.077X_2 - 0.485X_3 \quad (-110)$$

t 值： 12.666 5.022 -2.883 -5.909

p 值： 0.000 0.002 0.028 0.001

容限： 0.273 0.253 0.509

VIF： 3.669 3.954 1.966

$R^2=0.981$ $F=152.419$ $p=0.000$

其中：Y——中国老龄化水平

X_1 ——中国人均国内生产总值

X_2 ——中国出生率

X_3 ——中国每个医生平均负担人口数

从回归方程(-109)的结果看, X_1 、 X_2 和 X_3 的整体对 Y 具有显著的影响作用, 但是其中的 X_1 和 X_2 在 $\alpha=0.10$ 条件下均无法通过显著性检验, X_3 的容限度和 VIF 值分别为 0.097、10.265, 说明解释变量间存在多重共线性问题。 X_3 与 X_1 、 X_2 之间存在很强的线性关系, X_3 变动的 90.3%(1-0.097)可由 X_1 、 X_2 来解释。回归方程 (-110)不存在类似问题。为了便于比较, 我们剔除变量 X_3 , 重新估计模型, 得:

$$\text{德国: } \hat{Y} = 1734.532 + 0.011X_1 - 37.349X_2 \quad (-111)$$

t 值: 11.314 2.645 -4.064

p 值: 0.000 0.038 0.007

容限: 0.503 0.503

VIF: 1.990 1.990

$R^2=0.928$ $F=38.479, p=0.000$

$$\text{中国: } \hat{Y} = 753.190 + 0.261X_1 - 13.488X_2 \quad (-112)$$

t 值: 4.643 2.598 -2.107

p 值: 0.002 0.036 0.073

容限: 0.283 0.283

VIF: 3.532 3.532

$R^2=0.912$ $F=36.122, p=0.000$

两方程在显著性水平 $\alpha=0.10$ 左右均通过显著性检验, 且修正了多重共线性问题。

从以上两个回归方程, 可见两国的人口老龄化水平都与经济水平、出生率有较强的相关性。人均国内生产总值与老龄化水平呈正相关, 而出生率都与老龄化水平呈负相关。这也就从量化角度证实了本文前述的观点。

对比回归方程(-111)(-112), 不难得出以下结论:

1. 中国的人均国内生产总值对其老龄化水平的影响程度大于德国。假设双方的人均国内生产总值均增加 1000 美元, 其它因素不变, 中国的人口老龄化水平将平均增加 2.61%, 而德国的人口老龄化水平将平均增加 0.11%。

2. 中国的出生率对其老龄化水平的影响程度小于德国。假设其它因素不变条件下, 双方的人口出生率都降低 1‰, 中国的人口老龄化水平将平均增加万分之 13.488, 然而德国的人口老龄化水平将平均增加万分之 37.349。

从表 8-50、表 8-52中可以看出, 我国目前的人口老龄化水平尚低于德国 1950 年水平。运用回归方程(-112), 可以作一项控制预测。若我国的人口老龄化水平、出生率均达到德国 1980 年水平时, 人均国内生产总值将为 3559.54 美元。这一数据大大低于德国 1980 年的人均国内生产总值 13216.5 美元。即使我国的出生率仍维持当前水平, 而人口老龄化水平达到德国 1980 年水平, 我们的人均国内生产总值仅达到 3915.08 美元, 仍大大低于德国。从这一项预测中可见, 我国相对于德国而言, 经济水平严重滞后于人口老龄化的发展水平。

四、对家庭养老与健康老龄化的新思考

诚然, 在我国应实行基本养老保险, 企业补充养老保险和个人储蓄性养老保险相结合的多层次的养老保险体系。但必须明确的是, 对老年人基本生活的需求的满足绝不能取代对心理需要的满足。在德国等发达西方国家中, 经济的发达并不能掩盖人情的淡漠, 老年人在退休之后, 精神上的空虚和寂寞无人关怀, 随之, “恐老症”患者剧增, 老年人的自杀与犯罪现象屡见不鲜。然而作为亚洲国家的典范, 中国在其几千年的历史长河中, “孝悌”始终被视为传统美德而倍受推崇。正如孟子所云: “谨庠序之教, 申之以孝悌之义, 颁白者不负戴

于道路矣。”虽然当前由于“四·二·一”家庭结构的出现,青年一代独立性的加强和依赖性的减弱及人们价值观的转变等种种因素已使得单一的家庭养老模式出现了危机,但西方忽视家庭养老所带来的负效应已给我们敲醒了警钟:社会养老绝不能完全取代家庭养老,在养老模式的选择上,必须重视家庭养老的重要性,借以寻求社会养老和家庭养老的最佳结合点。

在“老有所养,老有所为”的基础上,必须追求“老有所为,老有所学,老有所乐”的境界。换言之,就是要营造一个健康老龄化的氛围,实现老年人的文化养老,提高老年人的生活质量。老人公寓、托老所、老人俱乐部、老人活动中心、老人大学、老人进修学院、老人疗养院及老人专家开设的老龄问题咨询处等社会性老年福利设施在西方已层出不穷,前西德还创办了“老年知识交易所”,请老专家为青少年传授各种知识,深受欢迎。这些都是我国值得借鉴的。

联合国制定的《2001 年全球解决人口老龄化问题方面的奋斗目标》中强调:“开展健康老龄化运动旨在从整体上促进老年人的健康,从而使老年人在体力方面、才能方面、社会方面、感情方面、脑力和精神方面得到平衡发展。”在我国,人口老龄化的趋势是不可避免的,但借鉴于德国老龄工作的经验于我国,健康的人口老龄化是可以实现的。我们期盼着这一天的早日到来。

附: SPSS 操作步骤:

1、根据表 8-52 和表 8-50 资料,按如图 8-103 所示的格式录入并保存为 SPSS 数据文件。

	year	y	x1	x2	x3	group
1	1950	941	487.0	18.10	824.00	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	1991	1590	19853	10.40	317.06	1
10	1982	488	265.5	22.28	778.00	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	1996	694	678.0	16.98	631.00	2

图 8-103 数据文件格式

其中, group 为一标识变量(1——德国, 2——中国)。

2、估计模型。选择主菜单[Analyze]⇒[Regression]⇒[Linear], 在显示的主对话框中选择因变量和自变量。由于数据按图 8-103 的格式排列, 把 group 变量选入[Selection Variable]框中作为选择变量(如图 8-104 所示), 并单击[Rule]按钮设置 group=1, 计算过程中将只使用中国的资料(group=1)。返回主对话框后单击[OK]即可得到中国老龄化水平的回归方程。类似可得出德国老龄化水平的回归方程。

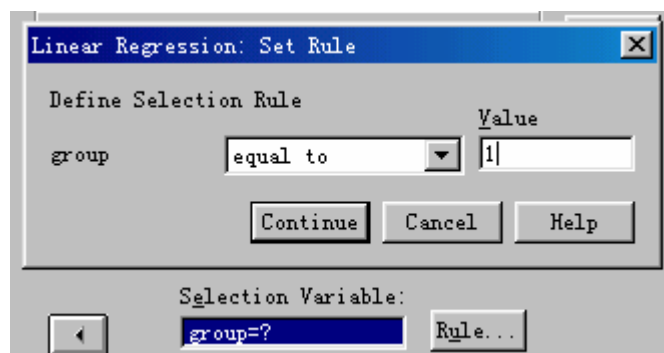


图 8-104

第九章 含虚拟自变量的回归分析

第一节 虚拟变量回归模型的基本概念

上一章讨论的回归模型中,因变量和自变量都是可以直接用数字计量的,即可以获得其实际观测值(如收入,支出,产量,物价水平等等),这类变量称作数量变量、定量变量或数量因素。然而,在实际问题的研究中,影响被解释变量的不仅有量的因素,还有有质的因素(如性别,民族,职业,文化程度,地区等等)即定性变量。例如,当我们要建立饮料的需求模型时,除了要考虑收入和价格这两个量的因素之外,还必须将"季节"这个质的因素,作为一个重要解释变量;建立粮食产量预测方程就应考虑到正常年与受灾年的不同影响。本节将讨论自变量含有定性变量的回归模型,因变量含有定性变量的回归分析——Logistic 分析留待下一节介绍。

由于受到质的因素影响,回归模型的参数不再是固定不变的。例如在饮料需求函数中,收入、价格与饮料需求量的关系是随着季节变化而改变的,也就是说,在不同的季节回归模型的参数也会有所不同;再如,我国居民的消费行为在改革开放前后大不相同,因此消费函数的参数也会发生变化。显然,如果我们忽略质的因素,仍把模型中的参数看作是固定不变的,得到的参数估计量就不能正确描述经济变量之间的关系。例如,如果我们在建立消费模型时没有考虑改革开放这一质的因素,估计结果就既不能代表改革前居民的消费行为,也不能正确描述改革开放后居民的消费行为。

质的因素通常表明某种"品质"或"属性"是否存在,所以将这类品质或属性量化的方法之一就是构造取值为"1"或"0"的人工变量。"1"表示这种属性存在,"0"则表示这种属性不存在。例如"1"可以表示改革开放以后的时期,"0"则表示改革开放以前的时期。再如,用"1"表示某人是男性,"0"表示某人是女性。这种取值为1和0的变量称为虚拟变量(Dummy Variable),又称为哑变量、二进制变量。

需要指出的是,虚拟变量主要是用来代表质的因素,但在有些情况下也可以用来代表数量因素。例如,在建立储蓄模型时,"年龄"显然是一个重要解释变量,虽然"年龄"是一个数量因素,但为了方便也可以用虚拟变量表示。例如,可以把居民分为两个年龄组:

第一组:20岁—35岁的居民

第二组:35岁—60岁的居民

用"1"表示第一年龄组,"0"表示第二年龄组,就可以估计年龄对储蓄的影响。

第二节 包含一个质因素的虚拟变量模型

一、运用虚拟变量改变回归直线的截距

如果回归模型中只包含一个质的因素,且这个因素仅有两种特征,则回归模型中只需引入一个虚拟变量。

图 9-105 表示两种情况下,中国通货膨胀率的变化情况。

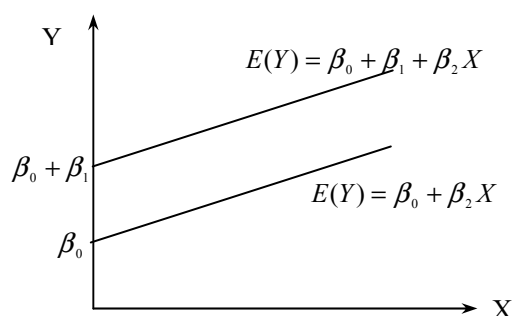


图 9-105

较低的直线描述了一般情况下通货膨胀率（Y）和工业生产增长速度（X）的关系；较高的直线则描述了 1988 年这一特殊年份的变化情况。

从这两条直线形状来看，它们趋势都相同，即这是两条平行但截距不同的直线，截距不同是因为预期基点不同。

定义

$$D = \begin{cases} 1 & (1988\text{年}) \\ 0 & (\text{其它}) \end{cases}$$

建立线性回归模型：

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon$$

这样，我们可以用上面这个模型表示以下两种情况：

$$\begin{cases} Y = \beta_0 + \beta_1 + \beta_2 X + \varepsilon & (1988\text{年}) \\ Y = \beta_0 + \beta_2 X + \varepsilon & (\text{其它}) \end{cases}$$

从上式我们可以看到 1988 年时直线的截距为 $\beta_0 + \beta_1$ ，其他年份时，直线的截距为 β_0 ，因此，用一个方程就可以表示截距不同的两条直线。

[例 10-20]通过设定回归模型考察采取某项保险革新措施的速度与保险公司的规模及其类型之间的关系。这里因变量 Y 是第一个公司采纳这项革新和给定公司采纳这项革新在时间上先后间隔的月数。公司的规模用其总资产额来计量，作为第一个自变量 X；公司的类型由股份公司和互助公司两种类型组成，引入虚拟变量 D 来计量。数据资料如表 9-53 所示。

表 9-53 保险公司数据资料

所需时间 Y	公司规模 X(100 万美元)	公司类型 D
17	151	0
26	92	0
21	175	0
30	31	0
22	104	0
0	277	0
12	210	0
19	120	0
4	290	0
16	238	0
28	164	1
15	272	1

11	295	1
38	68	1
31	85	1
21	224	1
20	166	1
13	305	1
30	124	1
14	246	1

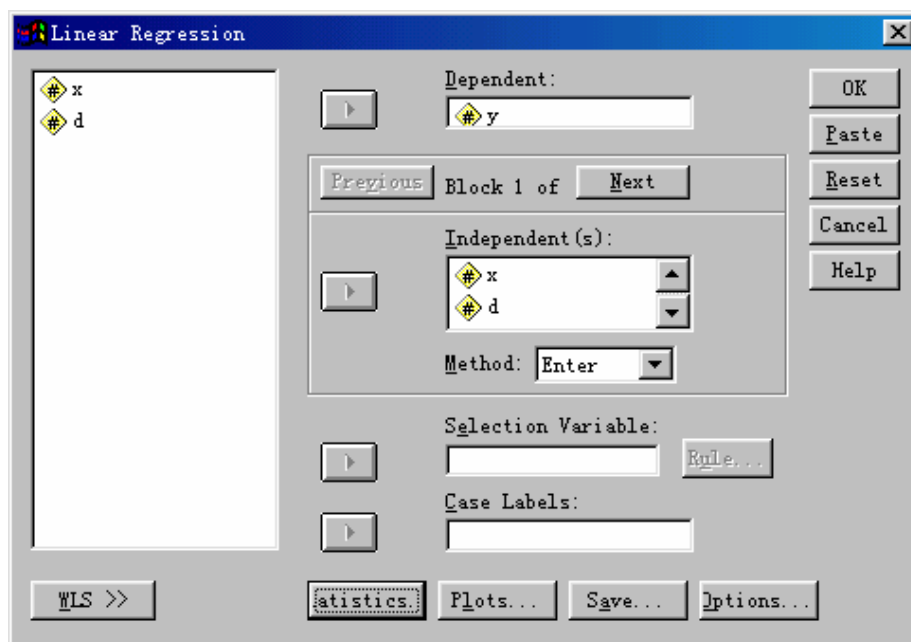
建立回归模型

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon$$

$$D = \begin{cases} 1 & \text{股份公司} \\ 0 & \text{互助公司} \end{cases}$$

在 SPSS 中进行回归分析的步骤：

- (1) 创建 SPSS 数据文件并录入数据（变量分别为 Y、X、D）；
- (2) 选择主菜单[Analyze]>[Regression]>[Linear]，在回归分析主对话框中，把 Y 选入[Dependent]列表框，把 X、D 选入[Independents]框。单击[OK]采用默认设置进行计算。结果如下：



Model Summary

Model	R(复相关系数)	R Square (R^2)	Adjusted R Square(调整过的 R^2)	Std. Error of the Estimate	Durbin-Watson (DW)
1	.946	.895	.883	3.2211	1.971

ANOVA (方差分析表)

Model		Sum of Squares	df	Mean Square	F	Sig(p值)
1	Regression	1504.413	2	752.207	72.497	.000
	Residual	176.387	17	10.376		
	Total	1680.800	19			

Coefficients (系数)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	33.874	1.814		18.675	.000
	X	-.102	.009	-.911	-11.443	.000
	D	8.055	1.459	.439	5.521	.000

从输出结果可以得到回归方程

$$\hat{Y} = 33.874 - 0.102X + 8.055D$$

本例中研究者感兴趣的是公司类型 D 对采纳革新措施所需时间 Y 的影响, 所以对公司类型的影响效应进行显著性检验, 在 $\alpha=0.01$ 条件下, 将拒绝公司类型没有影响效应的零假设 (p 值 $=0.000 < \alpha=0.01$), 即公司类型有显著的影响效应。

可以看出, 在 SPSS 中对含有虚拟变量的回归模型进行参数估计时, 其操作步骤与一般线性回归模型是一致的。

二、运用虚拟变量改变回归直线的斜率

仍然研究通货膨胀率和工业生产增长率之间的相互关系, 这一回假设 1988 年与普通年份的预期基点相同, 即截距相同, 但变化幅度不同, 也就是斜率不同。这是很常见的事, 通货膨胀率在某一特定年份异乎寻常地高涨, 剧烈上升, 与普遍年份的变化趋势大不相同, 如图 9-106 所示。这就是通货预期不变, 但工业生产增长率的上升会伴有非常高的通货膨胀率的情况, 仍定义

$$D = \begin{cases} 1 & (1988 \text{年}) \\ 0 & (\text{其它}) \end{cases}$$

这次所要使用的线性回归模型要复杂些了, 其形式如下:

$$Y = \beta_0 + \beta_1 DX + \beta_2 X + \varepsilon$$

这样, 我们可以用上面的公式表示以下两种情况:

$$\begin{cases} Y = \beta_0 + (\beta_1 + \beta_2)X + \varepsilon & (1988 \text{年}) \\ Y = \beta_0 + \beta_2 X + \varepsilon & (\text{其它}) \end{cases}$$

由上式可见, 直线的斜率变了, 1988 年时直线的斜率是 $\beta_1 + \beta_2$, 其余年份则是 β_2 。所以, 我们同样运用含有虚拟变量的一个模型就表示了斜率不同的两种情况。

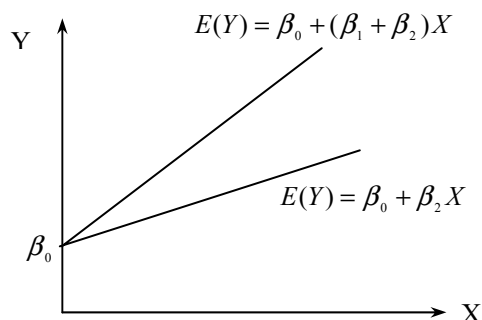


图 9-106

三、运用虚拟变量同时改变回归直线的斜率和截距

实际生活中，1988 年通货膨胀率的变化与普通年份的变化根本不同。这种不同表现在通货膨胀预期的基点和趋势都不同。图 9-107 的情况更符合实际情形。

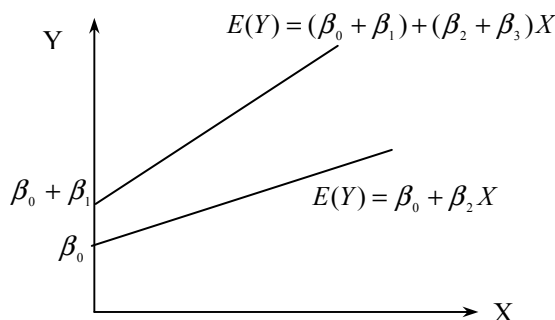


图 9-107

要用一个模型表示出这样一种复杂的情形，就可以通过改变斜率、截距的复合形式构造出一个新模型来。仍定义：

$$D = \begin{cases} 1 & (1988\text{年}) \\ 0 & (\text{其它}) \end{cases}$$

所构造的新模型具体形式如下：

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 DX + \varepsilon$$

上面这个方程可以表示如下两种情况：

$$\begin{cases} Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X + \varepsilon & (1988\text{年}) \\ Y = \beta_0 + \beta_2 X + \varepsilon & (\text{其它}) \end{cases}$$

这样一来，回归直线的斜率和截距都发生了变化。

到了这里，也许会有小小的疑问：为什么非要用虚拟变量，而不直接对下面的两个方程：

$$Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X + \varepsilon$$

和

$$Y = \beta_0 + \beta_2 X + \varepsilon$$

进行回归呢？它与对 $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 DX + \varepsilon$ 进行回归之后再写成分裂式有什么区别呢？其区别在于随机扰动项 ε 。如果直接对两个方程进行回归，这是两个单独方程，所以相当于进行了两次回归。现在的问题是我们不能保证两次回归中两个随机扰动项是同样的，这就意味着，我们是在不同的随机影响下分别研究 1988 年的情况和其它年份情况。而 $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 DX + \varepsilon$ 则仅需一次回归， ε 是相同的，意味着无论是高通胀还是其它时间，随机因素的影响都假设是一样的。

四、设置虚拟变量的原则

如果在模型中引入多个虚拟变量时，虚拟变量的个数应按下列原则来确定：对于包含一个具有 m 种特征或状态的质因素的回归模型，如果回归模型不带常数项，则中需引入 m 个虚拟变量；如果有常数项，则只需引入 $m-1$ 个虚拟变量。

第三节 包含多个质的因素的虚拟变量模型

上一节讨论的虚拟变量回归模型中只包含一个质的因素。在实际问题中,经常出现有多个质因素影响截距和斜率的情形。例如,居民消费行为的影响因素中还可能包括性别、教育程度、地理区域等质因素。又如,饮料需求量除了受收入水平的影响外,还会受到季节和地区等质因素的影响。这些质的因素不仅改变模型的截距和斜率,质因素之间也往往有交互影响。

现建立一个简单的饮料需求模型:

$$Y_t = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + \beta_4 X_t + \varepsilon$$

其中, Y_t 和 X_t 分别表示某季度的饮料需求量和人均收入, D_1 和 D_2 是虚拟变量, 分别表示季节因素和地区因素:

$$D_1 = \begin{cases} 1 & \text{夏季} \\ 0 & \text{冬季} \end{cases} \quad D_2 = \begin{cases} 1 & \text{城市} \\ 0 & \text{农村} \end{cases}$$

模型中的虚拟变量及虚拟变量的乘积, 考虑了截距项各种可能的变化:

$D_1 = 0, D_2 = 0$, 截距为 β_0

$D_1 = 1, D_2 = 0$, 截距为 $\beta_0 + \beta_1$

$D_1 = 0, D_2 = 1$, 截距为 $\beta_0 + \beta_2$

$D_1 = 1, D_2 = 1$, 截距为 $\beta_0 + \beta_1 + \beta_2 + \beta_3$

其中 β_3 表示季节因素和地区因素的交互影响。

第四节 案例：虚拟变量在新股上市模型中的应用

对于股票二级市场来说, 认购新股是一种风险小而收益较高的重要投资手段。但新股价格的市场定位受多种复杂因素的影响, 具有相当的“不可预测性”。若能较准确地预测出新股上市的价位, 将有助于市场投资者理智地控制自己对上市新股的投资行为, 最大限度地规避市场风险。对于券商及机构投资者, 由于资金量巨大, 要求充分考虑资金的时间价值和机会成本, 对申购新股收益率的测算就显得尤为重要, 而收益率的测算实质是新股上市价位的预测。同时, 预测新股上市价位能为其提供有关买卖时机的有益参考, 有利于提高大资金在市场中的可操作性。从上市公司的角度来看, 预测公司股票上市价位有助于判断股票发行价的制定是否合理。就证券主管部门而言, 预测新股上市价位有利于其对整个证券市场的宏观调控。

以下尝试以深市 1997 年及 1998 年 1—4 月上市新股为样本, 运用统计分析方法进行数据处理, 最终建立起一个数学模型, 借以描述影响新股上市后市场价位的诸因素及它们之间的复杂联系, 预测其上市价位。

(一) 模型设计

1、新股上市价格影响因素的定性分析

参考邱冬阳、黄娇敏:《深市新股上市价格预测模型》,《预测》1999.1。

根据证券投资理论的一般原理,新股上市价格的影响因素包括政治因素,经济周期、通胀、利率、财政金融政策等宏观经济状况,行业的寿命周期因素,上市公司本身的财务状况,市场的期望水平,一级二级市场资金的供求状况。

2、变量设置

根据影响因素分析和资料搜集的可行性,我们把上述因素设置如下变量进行模拟。 X_1 为冻结天数,指发行期申购资金冻结天数; X_2 为等待天数,指申购资金解冻至上市的时间间隔; X_3 为中签率; X_4 为股市指数; X_5 为每股收益,指新股发行前每股收益(未摊薄); X_6 为成长性,用新股发行前两年的年均收益增长率近似; X_7 为发行价; X_9 为发行后每股资产净值; X_{10} 为市净率, $X_{10} = X_7 / X_9$; X_{11} 为发行前股本; X_{12} 为新股发行数; X_{14} 为发行后总股本, $X_{14} = X_{11} + X_{12}$; X_{15} 为发行前市盈率, $X_{15} = X_7 / X_5$; X_{16} 为发行后市盈率, $X_{16} = (X_7 / X_5) X_{14} / X_{11}$; X_{17} 为新股发行值, $X_{17} = X_7 X_{12}$; X_8 为上市日平均价,指新股上市日市场平均价格; y 涨幅, $y = (X_8 - X_7) / X_7$ (y 为因变量,其余为自变量)。

设置虚拟变量。市场大势、行业板块、市场偏好、经济金融政策等也是影响新股上市价格的重要因素。但这些因素无具体的变量来代替,且不直接表现为一定数值,为了在模型中充分反映这些因素的作用,通过设定如下虚拟变量来解决。

D 为市场大势,是主观评定的市场强度,同大盘走势和市场人气密切相关。我国股市当前没有做空机制,好的大势环境对新股助涨明显,但疲弱的行情对新股上市当天涨幅的负面影响则相对较小,可以讲并不产生真正意义上的负面推动力。加之政府主管部门对新股发行时机的把握也较有尺度,疲软大势状况对新股上市价格定位的影响被进一步削弱。考虑这些因素,故将大势分为刺激性涨势和其他两种情况,分别赋值为 1 和 0。

S_i 期望系数,该变量内涵较广,包含行业属性、市场偏好、公司形象宣传等对新股上市价位有着重要影响的非数量性因素。期望系数的内涵中,对个股利好的方面包括:朝阳产业或优势行业、政策性利好、市场偏好或近期热点(如重庆实业的直辖市概念等特殊题材)形象宣传较佳等;对个股不利的方面有:夕阳产业或劣势行业、政策性利空、非市场热点、公司形象不佳等。在众多因素中,公司的行业属性是最重要的影响因素,因此,首先将上市公司按行业属性分为优势行业(含高科技、电子产业、医药行业及某些公用事业股)、一般性行业、劣势行业(含钢铁、纺织及重化工等),然后根据影响个股涨幅的上述本质性因素的有无及多寡,设置虚拟变量为: $S_i = 1$ 或 $S_i = 0$ ($i=1,2,3,4$)。

这两个虚拟变量在实际的预测运用当中是具有现实可操作性的。一只股票上市,我们可以根据其行业属性、有无特殊的本质、市场热点或当时大势有无突发性转折(受政策、消息等刺激)来进行综合评价,判断其属性,并赋予相应的模拟数值。

3、原始数据

样本数据来自 1997 年 1 月—1998 年 4 月 10 日在深交所上市的采用“上网发行”和“网下即退”方式发行的新股,有效数据共 85 组。

4、模型的初步构想

建立一个多元线性模型,其具有最高的拟合优度,模型中应尽量避免因素间的自相关以及多重共线性。

(二) 因素分析及模型检验

1、单因素的相关性检验及取舍

对变量进行单因素分析,首先定性分析其相关性的方向,然后从定量的角度分析每一个自变量与因变量的直接关系,为多元分析及对自变量的筛选作准备。

鉴于期望系数的特殊性及对 y 影响的重要性,我们没有将虚拟变量 S 进行单因素分析,而直接引入多元回归分析中。从计算结果来看,以下变量与 y 的相关性较强: X_{12} , X_{14} , X_{17} , X_{11} , X_{15} , D , X_3 , X_{16} , X_{40} 。其次依次为 X_{45} , X_{60} , X_{100} , X_{70} , X_{20} , X_{90} , X_{50} , X

10

2、多因素分析

采用逐步回归法,按 1 中各变量的排列次序依次加入变量,逐次记录计算结果,观察每一个因素对系统的影响,包括 3 方面: 对拟合优度的影响; 对调整后的拟合优度的影响;

参照 T 检验值考察自身的零系数概率。在变量筛选过程中,我们通过细致地观察各变量在不断变换的方程模式中的具体表现,以数据为基础,将定量分析和定性分析有机地结合起来,最终确定模型的最适变量。最终确定留用的变量为: X_3 、 X_4 、 X_5 、 X_7 、 X_{10} 、 X_{12} 、 X_{16} 、 D 、 S_1 、 S_2 、 S_3 、 S_4 ,基本涵盖了影响股票价格的主要因素,无论从经济意义上分析还是计算结果显示都比较满意。

3、模型的显著性检验

用以上变量构造模型并进行估计检验(其操作步骤与一般线性回归模型是一致的),得

$$y=0.9228-0.01961X_3+0.0003X_4+0.4909X_5-0.1246X_7-0.1055X_{10}-0.000037X_{12}-0.0106X_{16}+0.51193D+1.3062S_1+0.6177S_2-0.3119S_3-0.5969S_4$$

$$R^2=0.94282 \quad DW=2.1659 \quad F=98.9350$$

各变量对应的参数均通过显著检验。

(三)模型的应用及说明

从现实的角度来看,该模型的意义在于有助于合理地推测新股上市价格,为市场操作提供某种依据,并可以判断新股上市后的走势。

(1)模型的原始数据采自 1997 年 1 月至 1998 年 4 月 10 日在深市上市的新股,但模型建立的思路及方法同样适用于其他时期或者上海股市。

(2)市场大势、行业属性、市场偏好、企业形象等是影响新股上市价格定位的重要因素,我们通过设置虚拟变量成功地在模型中反映出它们的重要影响。但在应用该模型来进行预测时,对变量 D 、 S_n 的评价不可避免地带有一定的主观因素,因此较难把握,预测的准确程度将取决于应用者对市场形势、行业前景、上市公司潜质、国家宏观经济政策等因素的综合分析和判断能力。

(3)总体上讲,这个模型并不是一个封闭的模型,随着市场状况发生变化和样本的扩大,模型各变量、系数可能会有一定改变。一些变量的作用可能会削弱,一些变量的作用可能会增强,各项变量的系数也会随之进行一定的调整,变量的组合可能会发生一定变化。但本文所给出的模型的建立与调试方法却是较严谨的,就一般的线性模型来讲是科学的。通过不同时段、不同市场模型的差异可以更深入地分析产生差异的原因,更有利于我们把握股市的运动、变化规律。

第十章 Logistic 回归分析

第一节 Logistic 回归基本概念

线性回归模型的一个局限性是要求因变量是定量变量(定距变量、定比变量)而不能是定性变量(定序变量、定类变量)。但是在许多实际问题中,经常出现因变量是定性变量(分类变量)的情况。可用于处理分类因变量的统计分析方法有:判别分析(Discriminant analysis)、Probit 分析、Logistic 回归分析和对数线性模型等。在社会科学中,应用最多的是 Logistic 回归分析。Logistic 回归分析根据因变量取值类别不同,又可以分为 Binary Logistic 回归分析和 Multinomial Logistic 回归分析, Binary Logistic 回归模型中因变量只能取两个值 1 和 0 (虚拟因变量),而 Multinomial Logistic 回归模型中因变量可以取多个值。本章将只讨论 Binary Logistic 回归,并简称 Logistic 回归。

因变量只取两个值,表示一种决策、一种结果的两种可能性。例如,某个人能否拥有房子,受到多种因素的影响,如家庭情况、工龄、收入情况等,但最终的可能性只有两个,要么拥有住房,要么没有住房。我们把 $Y=1$ 定义为拥有住房, $Y=0$ 定义为其它情况,即

$$Y = \begin{cases} 1 & \text{拥有住房} \\ 0 & \text{其它情况} \end{cases}$$

从模型角度出发,不妨把事件发生的情况定义为 $Y=1$,事件未发生的情况定义为 $Y=0$,这样取值为 0、1 的因变量可以写为下式:

$$Y = \begin{cases} 1 & \text{事件发生} \\ 0 & \text{事件未发生} \end{cases}$$

我们可以采用多种方法对取值为 0、1 的因变量进行分析。通常以 p 表示事件发生的概率(事件未发生的概率为 $1-p$),并把 p 看作自变量 X_i 的线性函数,即

$$p = P(y=1) = F(\beta_i X_i) \quad i=1,2,\dots,k$$

不同形式的 $F(\cdot)$,就有不同形式的模型,最简单的莫过于使 $F(\cdot)$ 为一线性函数,即

$$p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (10-113)$$

我们可能会认为可用普通最小二乘法对上式进行估计,但因 p 的值一定在区间 $[0,1]$ 内,而且当 p 接近于 0 或 1 时,自变量即使有很大变化 p 的值也不可能变化很大,所以对上式直接用普通最小二乘法进行估计是行不通的。

从数学上看,函数 p 对 X_i 的变化在 $p=0$ 或 $p=1$ 的附近是不敏感的、缓慢的,且非线性程度较高。于是寻求一个 p 的函数 $\theta(p)$,使得它在 $p=0$ 或 $p=1$ 附近时变化幅度较大,而函数的形式又不是很复杂。因此,我们引入 p 的 Logistic 变换(或称为 p 的 Logit 变换),即

$$\theta(p) = \log it(p) = \ln\left(\frac{p}{1-p}\right) \quad (10-114)$$

特别指出,本章介绍的 Logistic 回归,应与第八章的 Logistic 曲线模型(即 S 或倒 S 形曲线)相区别。

与第八章的符号表示不同,本章中 p 表示事件发生的概率,而用 k 表示自变量个数。

(11-1) 是一个线性概率模型,可用 WLS 进行估计,但仍存在许多问题。

其中, $p/(1-p)$; $\text{logit}(p)$ 是因变量 $Y=1$ 的差异比(odds ratio)或似然比(likelihood ratio)的自然对数, 称为对数差异比(log odds ratio)、对数似然比(log likelihood ratio)或分对数。

很明显, $\theta(p)$ 以 $\text{logit}(0.5)=0$ 为中心对称 (如表 10-54 所示), $\theta(p)$ 在 $p=0$ 和 $p=1$ 的附近变化幅度很大, 而且当 p 从 0 变化 1 时, $\theta(p)$ 从 $-\infty$ 变到 $+\infty$ 。用 $\theta(p)$ 代替式(10-113)中的 p 就克服了前面指出的两点困难。如果 p 对 X_i 不是线性的关系, $\theta(p)$ 对 X_i 就可以是线性的关系了。用 $\theta(p)$ 代替式(10-113)中的 p , 得

$$\theta(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (10-115)$$

由式(10-114), 将 p 由 θ 来表示, 得

$$p = \frac{e^\theta}{1+e^\theta} = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon}}$$

表 10-54 p 和 $\text{logit}(p)$ 之间的关系 (一部分)

p	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
$\text{Logit}(p)$	-0.847	-0.405	0.0	0.405	0.847	1.386	2.197	2.944	4.595

第二节 Logistic 回归模型的估计与检验

对于模型(10-115), 采用最大似然估计法(Maximum likelihood estimation, MLE)进行估计, 它与用于估计一般线性回归模型参数的普通最小二乘法 (OLS) 形成对比。OLS 通过使得样本观测数据的残差平方和最小来选择参数, 而最大似然估计法通过最大化对数似然值(log likelihood)估计参数。最大似然估计法是一种迭代算法, 它以一个预测估计值作为参数的初始值, 根据算法确定能增大对数似然值的参数的方向和变动。估计了该初始函数后, 对残差进行检验并用改进的函数进行重新估计, 直到收敛为止 (即对数似然不再显著变化)。

[例 10-1]设有住房及收入情况的统计资料如表 10-55 所示。

表 10-55 住房及收入数据

住房 Y	收入 X	住房 Y	收入 X	住房 Y	收入 X
0	10	0	10	0	11
1	17	1	17	0	8
1	18	0	13	1	17
0	14	1	21	1	16
0	12	1	16	0	7
1	9	0	12	1	17
1	20	0	11	1	15
0	13	1	16	1	10
0	9	0	11	1	25
1	19	1	20	0	15
0	12	1	18	0	12
0	4	1	16	1	17
1	14	0	10	0	17
1	20	0	8	1	16
0	6	0	18	1	18

1	19	1	22	0	11
0	11	1	20		

建立 Logistic 回归模型

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

其中, $p_i = P(Y_i=1)$, $Y_i = \begin{cases} 1 & \text{有住房} \\ 0 & \text{无住房} \end{cases}$ 。

在 SPSS 中估计参数步骤如下:

(1) 在 SPSS 中录入表 10-55 中数据 (变量为 Y 和 X), 并保存数据文件; 在主菜单中选择 [Analyze] => [Regression] => [Binary Logistic] (如图 10-108 所示)。

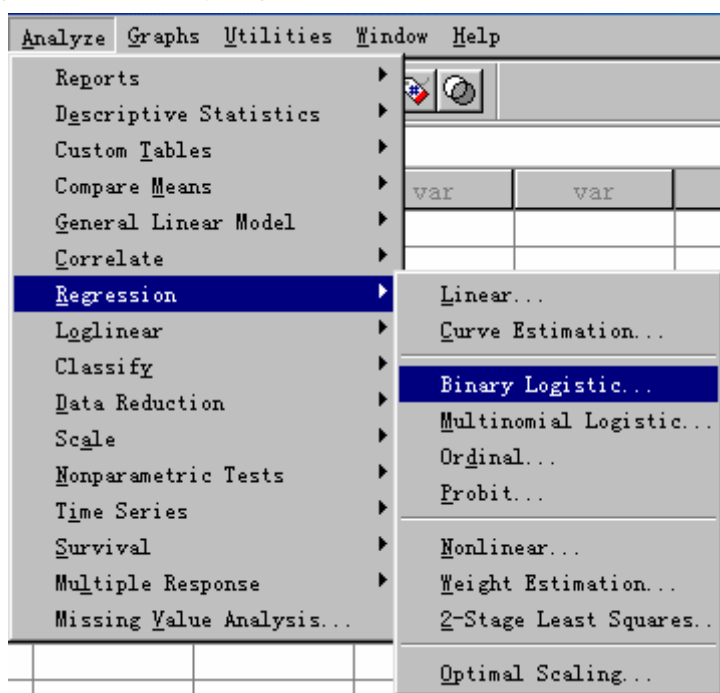


图 10-108

(2) 在 [Logistic Regression] 对话框中, 选择 Y 进入 [Dependent] 框作为因变量, 选择 X 进入 [Covariates] 作为自变量 (如图 10-109 所示)。单击 [Method] 的下拉菜单, SPSS 提供了 7 种方法:

- [Enter]: 所有自变量强制进入回归方程;
- [Forward: Conditional]: 以假定参数为基础作似然比检验, 向前逐步选择自变量;
- [Forward: LR]: 以最大局部似然为基础作似然比检验, 向前逐步选择自变量;
- [Forward: Wald]: 作 Wald 概率统计法, 向前逐步选择自变量;
- [Backward: Conditional]: 以假定参数为基础作似然比检验, 向后逐步选择自变量;
- [Backward: LR]: 以最大局部似然为基础作似然比检验, 向后逐步选择自变量;
- [Backward: Wald]: 作 Wald 概率统计法, 向后逐步选择自变量。

本例选默认项 [Enter] 方法。

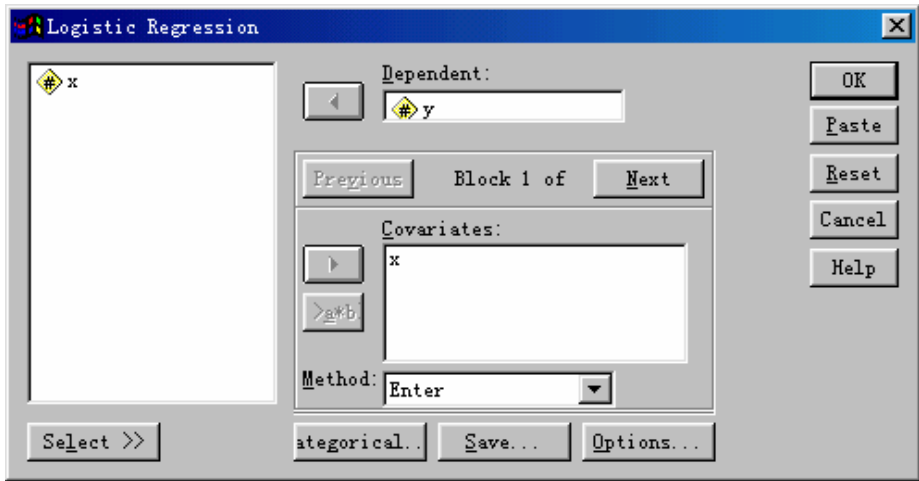


图 10-109

(3) 单击[Logistic Regression]对话框中的[Options]按钮，在显示的子对话框中选择[Classification plots]和[Hosmer-Lemeshow goodness-of-fit]等选项（如图 10-110所示），并单击[Continue]返回主对话框。

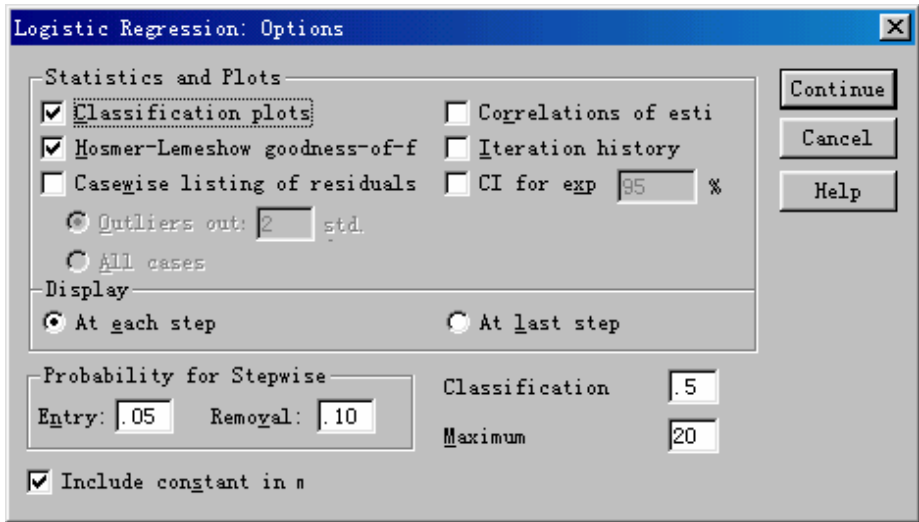


图 10-110

(4) 单击主对话框中[OK]按钮，输出结果如下：

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	32.379	1	.000
	Block	32.379	1	.000
	Model	32.379	1	.000

Model Summary			
Step	-2 Log likelihood (-2对数似然值)	Cox & Snell R Square (Cox & Snell的R ²)	Nagelkerke R Square (Nagelkerke的R ²)
1	36.856	.477	.636

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.

1	11.266	7	.127
---	--------	---	------

Hosmer and Lemeshow Goodness-of-Fit Test

Group	Y= 0		Y=1		Total
	Observed	Expected	Observed	Expected	
1	5.000	4.909	.000	.091	5.000
2	4.000	5.548	2.000	.452	6.000
3	5.000	4.281	.000	.719	5.000
4	6.000	4.406	.000	1.594	6.000
5	2.000	1.816	2.000	2.184	4.000
6	.000	1.313	5.000	3.687	5.000
7	1.000	1.011	5.000	4.989	6.000
8	1.000	.537	5.000	5.463	6.000
9	.000	.179	7.000	6.821	7.000

Classification Table for Y

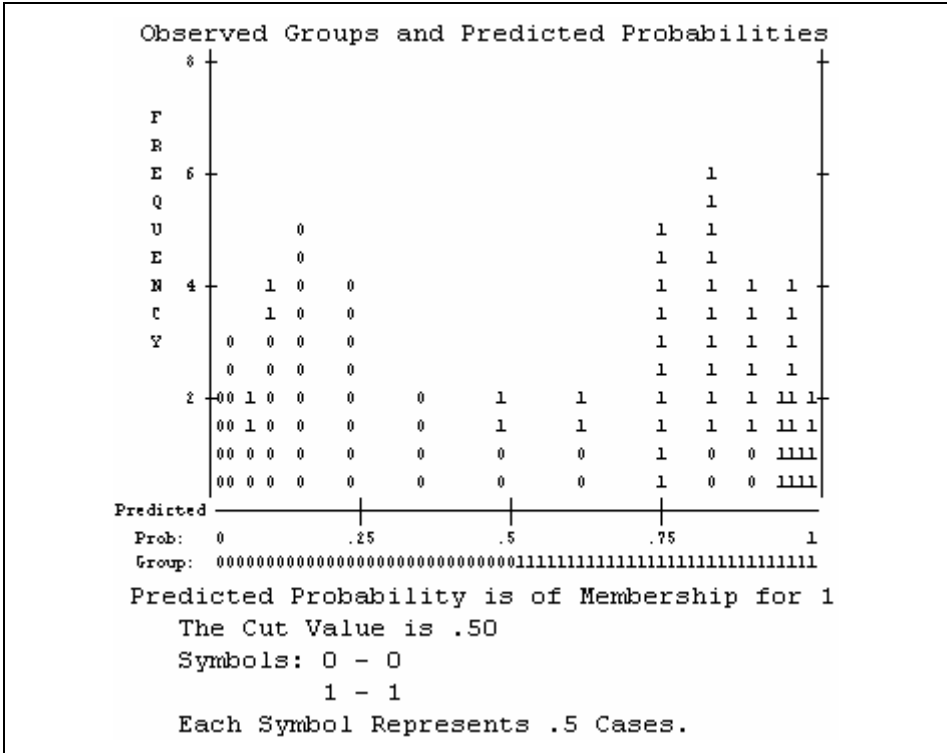
		Predicted (预测值)		
		0	1	Percent Correct
Observed (观测值)	0	21	3	87.50%
	1	3	23	88.46%
	Overall(总计)			88.00%

The Cut Value is .50

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
X	.563	.145	15.005	1	.000	1.757
Constant	-7.981	2.129	14.046	1	.000	.000

a Variable(s) entered on step 1: X.



模型参数估计后，必须进行检验。下面解释一些常用的检验统计量：

(一) -2 对数似然值(-2 log likelihood , -2LL)

似然(likelihood)即概率，特别是由自变量观测值预测因变量观测值的概率。与任何概率一样，似然的取值范围在 0、1 之间。对数似然值(log likelihood, LL)是它的自然对数形式，由于取值范围在[0,1]之间的数的对数值负数，所以对数似然值的取值范围在 0 至 $-\infty$ 之间。对数似然值通过最大似然估计的迭代算法计算而得。因为 $-2LL$ 近似服从卡方分布且在数学上更为方便，所以 $-2LL$ 可用于检验 Logistic 回归的显著性。 $-2LL$ 反映了在模型中包括了所有自变量后的误差，用于处理因变量无法解释的变动部分的显著性问题，又称为拟合劣度卡方统计量(Badness-of-fit Chi-square)。当 $-2LL$ 的实际显著性水平大于给定的显著性水平 α 时，因变量的变动中无法解释的部分是不显著的，意味着回归方程的拟合程度越好。 $-2LL$ 的计算公式为：

$$-2LL = -2 \sum_{i=1}^n (y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i))$$

其中, y_i 为各样本观测值, $\hat{p}_i$ 为第 i 个样本的预测概率, 即

$$\hat{p}_i = \exp(\hat{\theta}_i) / [1 + \exp(\hat{\theta}_i)] = \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}) / [1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik})].$$

（二）拟合优度(Goodness of Fit)统计量

Logistic 回归的拟合优度统计量计算公式为

$$\sum_{i=1}^n \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}$$

实际应用中，经常采用分类表(Classification Table)来反映拟合效果。分类表如下所示。

		预测值		
		0	1	正确分类比例
观测值	0	n_{00}	n_{01}	f_0

	1	n_{10}	n_{11}	f_1
	总计			f

其中, $n_{ij} (i=1,0, j=1,0)$ 表示样本中因变量实际观测值为 i 被预测为 j 的样品数,

$$f_0 = \frac{n_{00}}{n_{00} + n_{01}} \times 100\%, \quad f_1 = \frac{n_{11}}{n_{10} + n_{11}} \times 100\%,$$

$$f = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \times 100\%$$

另外, 还经常采用分类图(Classification Plot)来反映拟合效果。它也可以用于发现奇异值, 也可以用于判断是否应按其它规则来划分两个预测组。

(三) Cox 和 Snell 的 R^2 (Cox & Snell's R-Square)

Cox 和 Snell 的 R^2 试图在似然值基础上模仿线性回归模型的 R^2 解释 Logistic 回归模型, 但它的最大值一般小于 1, 解释时有困难。其计算公式为:

$$R_{CS}^2 = 1 - \left(\frac{l(0)}{l(\hat{\beta})} \right)^2$$

其中, $l(\hat{\beta})$ 表示当前模型的似然值(likelihood), $l(0)$ 表示初始模型的似然值。

(四) Nagelkerke 的 R^2 (Nagelkerke's R-Square)

为了对 Cox 和 Snell 的 R^2 进一步调整, 使得取值范围在 0 和 1 之间, Nagelkerke 把 Cox 和 Snell 的 R^2 除以它的最大值, 即:

$$R_N^2 = R_{CS}^2 / \max(R_{CS}^2)$$

这里的 $\max(R_{CS}^2) = 1 - [l(0)]^2$

(五) 伪 R^2 (Pseudo-R-square)

伪 R^2 与线性回归模型的 R^2 相对应, 其解释与相似, 但它的最大值小于 1。

(六) Hosmer 和 Lemeshow 的拟合优度检验统计量 (Hosmer and Lemeshow's Goodness of Fit Test Statistic)

与一般拟合优度检验不同, Hosmer 和 Lemeshow 的拟合优度检验通常把样本数据根据预测概率分为十组, 然后根据观测频数和期望频数构造卡方统计量 (即 Hosmer 和 Lemeshow 的拟合优度检验统计量, 简称 H-L 拟合优度检验统计量), 最后根据自由度为 8 的卡方分布计算其 p 值并对 Logistic 模型进行检验。如果该 p 值小于给定的显著性水平 α (如 $\alpha=0.05$), 则拒绝因变量的观测值与模型预测值不存在差异的零假设, 表明模型的预测值与观测值存在显著差异。如果 p 值大于 α , 我们没有充分的理由拒绝零假设, 表明在可接受的水平上模型的估计拟合了数据。

(七) Wald 统计量

同线性回归方程的参数显著性检验类似, Wald 统计量用于判断一个变量是否应该包含在模型中, 检验步骤为:

(1) 提出假设:

$$H_0: \beta_i = 0 \quad (i=1,2,\dots,k)$$

$$H_1: \beta_i \neq 0$$

(2) 构造检验统计量——Wald 统计量

如果自变量 X_i 不是分类变量, 那么 Wald 统计量为

$$Wald_i = \frac{b_i^2}{Var(b_i)}$$

如果自变量 X_i 是分类变量，Wald 统计量的公式较繁，这里从略。

Wald 统计量近似服从于自由度等于参数个数的卡方分布。

(3) 作出统计判断。

第三节 案例：审计意见预测模型的构建

本案例意在用会计变量、股市变量和公司变量等来构建审计意见的预测类型。研究过程中主要分以下两个步骤：

(1) 将获得无保留审计意见的上市公司与获得非无保留审计意见的上市公司作为研究对象，比较双方共同的属性特征，找出决定审计意见签发类型的关键因素；

(2) 在此基础上构建审计意见预测模型，以助于注册会计师审计意见的签发，也可作为法律诉讼中的客观依据。

通过对 1995 年至 1996 年我国两类上市公司的财务报告及股市行情的描述性统计分析（详见表 10-56）得出：分别获得无保留审计意见和非无保留审计意见的两类公司的一些指标在以下五个方面存在显著的差异：该公司的财务杠杆、应收帐款占用率、总资产报酬率、所属哪个证交所以及前年度是否接受过非无保留审计意见。另外，当年亏损与公司规模这两个变量的差异性则次之。上表中各变量的计算公式或定义方式及其预期符号的确定详见表 10-56。

由于以上十一个变量并非正态分布，无法采用多元判别分析的方法，所以这里抽取了两类样本：估计样本与验证样本，运用 Logistic 回归模型进行分析。其中，估计样本 55 对，用于估计型参数；验证样本 25 对，用于检验所构建模型的适用性。

在进行 Logistic 模型分析过程中，选用了五个可供选择的模型（详见表 10-58）。从表中可以得出，审计意见的类型与以下四个变量显著相关：财务杠杆（F）、当年亏损（C）、总资产报酬率（R）及所属证交所（S）。

估计样本与验证样本对于两类审计意见的总正确分类率列于表 10-59。由表中数据可得，估计样本的审计意见总正确分类率为 70%—90%；验证样本的审计意见总正确分类率为 64%—85%。

虽然，仅从验证样本的审计意见全部正确分类率来看，本模型构建较为理想，但与其他国家在此方面的研究相比，样本量仍显太少。这主要是因为与发达国家相比，我国证券市场才刚起步，上市公司中可用于分析研究的公司数量过少。另外，这里仅讨论了十一个变量，而实际上影响注册会计师审计意见类型的因素还有很多。最后，注册会计师自身的超然独立性也是一个十分需要的关键因素。总之，有关此方面的问题仍有待进一步的研究与探讨。

表 10-56 自变量的描述统计结果

		平均数				中位数			
自变量	预期符号	无保留审计意见	非无保留审计意见	T 检验值	<i>p</i> 值	无保留审计意见	非无保留审计意见	Wilcoxon 检验值	<i>p</i> 值
会计变量									
财务杠杆	+	0.4130	0.5120	-3.8020	0.0002	0.4220	0.5275	798	0.0021
存货占用率	+	0.1882	0.1702	0.7486	0.4552	0.1454	0.1600	-71	0.7358
应收帐款占用率	+	0.0901	0.1140	-1.8225	0.0703	0.0648	0.0935	438	0.0348
公司规模	—	20.4637	20.5023	-0.3102	0.7568	20.3744	20.5049	333	0.1107
当年亏损	+	0.0750	0.1500	-1.5024	0.1350	0.0000	0.0000	25.5	0.2101
总资产报酬率	—	0.0497	0.0201	-5.0021	0.0001	0.0486	0.0198	400	0.0427
股市变量									
股市（已调整年价格回报率）	—	-0.1731	-0.2489	0.8273	0.4093	-0.1538	-0.1351	-204	0.3310
市场模型 Beta 系数	+	1.0160	1.0260	-0.5007	0.6173	1.0191	1.0303	83	0.6932
公司变量									
注册会计师类型	+	0.5125	0.5125	0.0000	1.0000	1.0000	1.0000	0	1.0000
所属证交所	?	0.6750	0.5000	2.2703	0.0245	1.0000	0.5555	-136.5	0.2210
以前年度接受非无保留审计意见	+	0.0000	0.1625	-3.9151	0.0001	0.0000	0.0000	45.5	0.0002

表 10-57 变量的计算公式或定义及其预期符号

变 量	计算公式或定义	预期符号	备 注
一、因变量			
审计意见类型	0=无保留审计意见 1=非无保留审计意见		
二、自变量			
1、会计变量			
财务杠杆	负债/资产	+	
存货占用率	存货/资产	+	
应收帐款占用率	应收帐款/资产	+	
公司规模	总资产帐面价值的对数	—	
当年亏损	1=当年亏损 0=其他	+	
总资产报酬率	净利润/总资产	—	
2、股市变量			
股市-已调整年价格回报率	年股值回报率-年股价回报率	—	
市场模型 Beta 系数	市场模型的溢出系数（根据当年日价格回报率）	+	
3、公司变量			
注册会计师类型	1=国际性大所的注册会计师 0=其他	+	
所属证交所	1=上交所 0=深交所	?	对虚拟变量不予确定预期符号
以前年度接受非无保留审计意见	1=以前年度接受过非无保留审计意见 0=其他	+	

表 10-58 Logistic 回归分析结果

	预期 符号	会计变量			市场变量			公司变量			全部变量			显著变量		
自变量		系数	Wald	p 值	系数	Wald	p 值	系数	Wald	p 值	系数	Wald	p 值	系数	Wald	p 值
常数		3.9287	0.3236	0.5695	1.7827	1.2125	0.2708	0.5182	1.7106	0.0909	6.7771	0.5307	0.4663	0.4528	0.3129	0.5759
会计变量																
财务杠杆	+	2.6161	2.2146	0.1059							2.7704	2.3953	0.1217	2.4189	3.1157	0.0775
存货占用率	+	3.0459	1.1857	0.2762							3.2917	1.2542	0.2628			
应收帐款占用率	+	-2.2971	1.8206	0.1772							-1.9629	1.0761	0.2996			
公司规模	—	-0.1940	0.2962	0.5863							-0.2488	0.3469	0.5559			
当年亏损	+	-2.1223	4.9868	0.0255							-1.9510	3.5864	0.0583	-1.7917	3.4473	0.0634
总资产报酬率	—	-23.4140	10.7840	0.0010							-24.7628	8.3088	0.0039	-20.3635	8.3257	0.0039
市场变量																
股市(已调整年价格 回报率)	—				-0.3857	0.8127	0.3673				0.5864	0.9472	0.3304			
市场模型 Beta 系数	+				1.7111	1.1444	0.2847				-1.3551	0.3759	0.5398			
公司变量																
注册会计师类型	+							0.2178	0.2669	0.6054	0.2598	0.2728	0.6014			
所属证交所	?							-1.1902	7.0774	0.0078	-1.0895	4.3425	0.0372	-0.9126	3.9718	0.0463
以前年度接受非无 保留审计意见	+							14.3966	0.0012	0.9724	14.4771	0.0013	0.9714			
伪 R^2		0.2547			0.0269			0.2132			0.3969			0.2622		
-2 对数似然值统计量		23.3240			2.2430			19.1660			38.8690			24.0840		

表 10-59 估计样本和验证样本的总正确分类率

	总的正确分类率
一、估计样本	
1、会计变量	85.8%
2、市场变量	70.0%
3、公司变量	79.5%
4、全部变量	90.4%
5、显著变量	86.2%
二、验证样本	
1、会计变量	78.9%
2、市场变量	64.2%
3、公司变量	70.8%
4、全部变量	86.2%
5、显著变量	79.5%

第十一章 非参数检验

第一节 非参数检验基本概念

在统计学的发展过程中，最先出现的推断统计方法都对样本所属总体的性质作出若干假设，即对总体的分布形状作某些限定。例如，第五章中的 Z 检验和 t 检验，假设样本的总体是正态分布的，或者假设两个样本都取自具有相同方差的总体，等等。这类推断统计方法对总体分布形状加以某些限定，把所要推断的总体数字特征看作未知的“参数”进行推断，这种推断统计方法称为参数统计方法（parameter statistical methods）或限定分布统计方法（distribution-specified statistical methods），其所做的假设检验则称为参数检验（parametric test）。前面我们讨论过的统计检验如 t 检验、Z 检验、F 检验等都是参数检验。这种对所要推断的总体分布事先作出某些限定或假设，在应用上存在一定的局限性，它只有在关于总体分布假设成立时，所得出的结论才是正确的。因此，它在很多场合下不便应用。因此，一些统计学家发展了许多对总体不作太多的或严格的限定的推断统计方法，这些方法一般不涉及总体参数，为了同上述的参数检验相区别，通常称为非参数统计（nonparametric statistics）或自由分布统计方法（distribution-free statistical methods），其所做的假设检验则称为非参数检验（nonparametric test）或自由分布统计检验（distribution-free statistical test）。

与参数检验方法对比，非参数检验方法具有许多特点。

首先，检验条件比较宽松，适应性强。参数检验假定总体分布正态或近似正态或以正态分布总体为基础构造 t 分布或 χ^2 分布来检验总体均值或方差是否发生显著性变化，这些条件是相当严格的。如果这些条件不存在，很可能检验结果产生方向性的错误。非参数检验不受这些条件的限制，大大填补了参数检验的不足。例如非正态的、方差不等的以及分布形状未知的资料都可适用，所以它的适应性强。

其次，检验的方法比较灵活，用途更广泛。非参数检验不但可以应用于定距、定比变量的检验而且也适用于定类、定序变量的检验，对于那些

不能直接进行加减乘除四则运算的定类数据和定序数据，运用符号检验、符秩检验都能起到比较好的效果，所以非参数检验的用途是更加广泛的。

再次，非参数检验计算相对简单，易于理解。由于非参数检验不用计量的方法，而用计数的方法，其过程及其结果都可以被直观地理解，为使用者所接受。

非参数检验的缺点也是明显的。非参数检验方法对总体分布的假定不多，适应性强，但方法也就缺乏针对性，其功效就不如参数检验。非参数检验用的是等级或符秩，而不是实际数值，方法简单，又会失去许多信息，因而检验的有效性也就比较差。例如对于一批适用于 t 检验的配对资料，如果采用符秩检验处理，其功效将低于 t 参数检验，如果用符号检验处理则效率更低，因为它对信息利用更不充分。当然如果假定的分布不成立，那么非参数检验就是更值得信赖的。

几乎对于每一种参数检验方法，都有一种或几种相应的非参数检验方法（如表 11-60 所示）。本章只选择其中的几种加以讨论。

表 11-60 参数检验与非参数检验方法对应表

参数检验方法	非参数检验方法
t 检验法	两个独立样本的中位数检验
t 检验法	两个独立样本的秩和检验
t 检验法（配对样本）	成对比较、单样本正负号检验
t 检验法（配对样本）	成对比较、单样本符号秩检验
单因素方差分析	K 个独立样本的 H 检验法
多因素方差分析	Friedman 检验法
相关系数	Spearman 秩相关系数

第二节 非参数检验方法

一、卡方检验(Chi-square test)

假设一个定性变量 Y 具有 k 个可能取值或有 k 种分类（标为 $1, 2, \dots, k$ ）， Y 的概率分布自然地由概率函数 $P(Y=i)(i=1, 2, \dots, k)$ 所确定。现在要考查已观察到的一组样本（容量为 n ）与某确定的分布 G 拟合的程度，相当于研究 $P(Y=i)(i=1, 2, \dots, k)$ 与 G 之间的差异，看这个差异是否属于偶然变异，根据原假设认为差异是偶然变异所致这样的原则，卡方检验的步骤如下：

1、提出假设

$H_0: P(Y=i)=G_i \quad (i=1, 2, \dots, k, \quad G_i \text{ 为 } G \text{ 分布})$

$H_1: P(Y=i) \neq G_i$

2、构造统计量

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

其中， O_i 为观测频数，期望频数 $E_i = n/k$ 。

3、作出判断

如果 $\chi^2 > \chi^2_{\alpha}(k-1)$ 或 $p < \alpha$ ，则拒绝零假设。

[例 11-21] 掷一颗六面体 300 次，结果如所示，试问这颗六面体是否均匀？
($\alpha=0.05$)

表 11-61 掷一颗六面体的结果

点数 i	1	2	3	4	5	6
观测频数 O_i	43	49	56	45	66	41

解：(1) 定义变量名为 Y，取值为 1、2、3、4、5、6，分别代表六面体的六个点，在 SPSS 中输入数据。

(2) 选择主菜单 [Analyze] => [Nonparametric Tests] => [Chi-square] (如图 11-111 所示)。

(3) 在显示的 [Chi-square Test (卡方检验)] 主对话框中，把 Y 选入 [Test Variable] 作为检验变量 (如图 11-112 所示)。

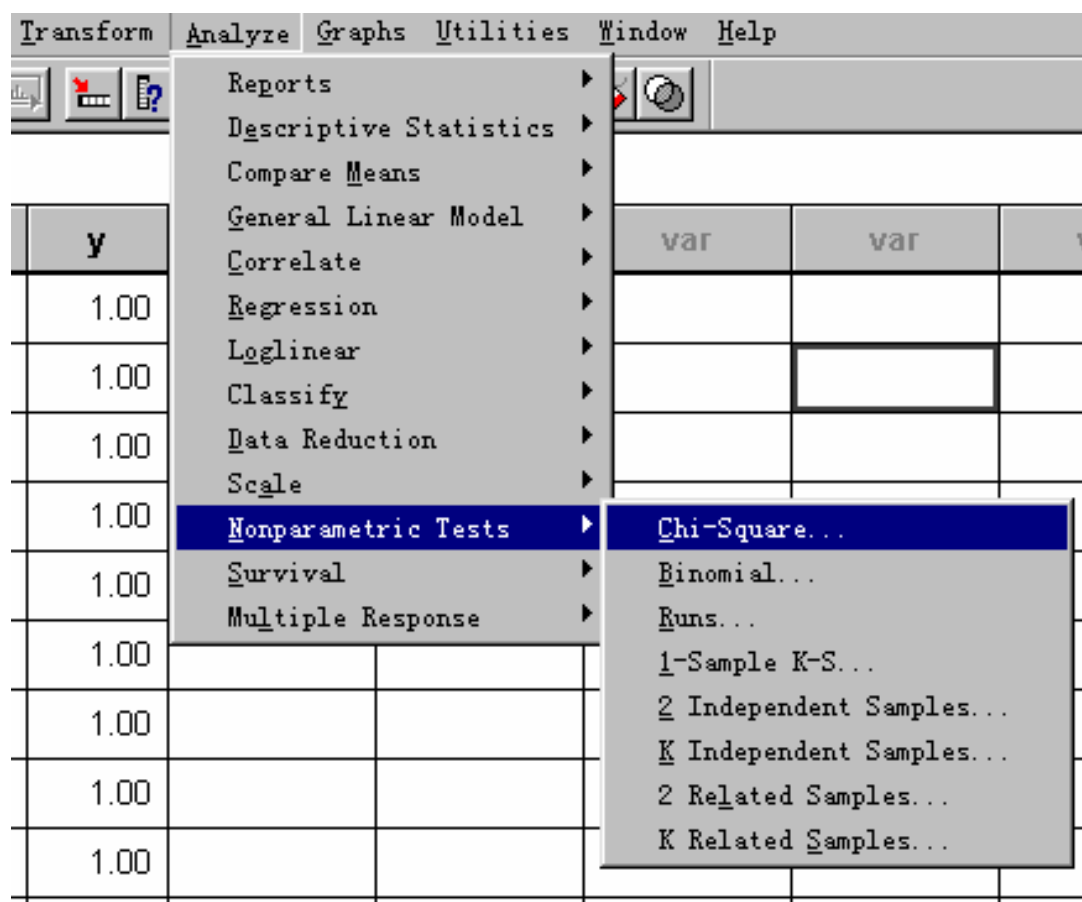


图 11-111

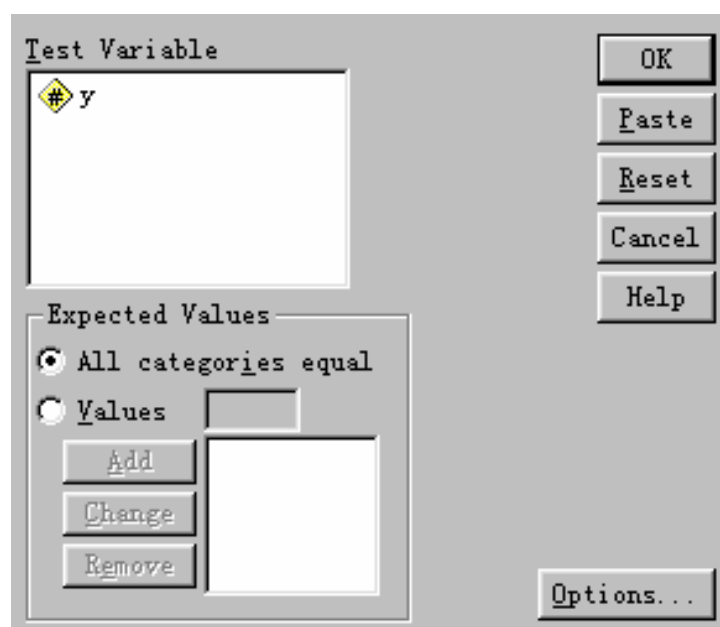


图 11-112

(4) 单击[OK]后，输出结果如下：

Y			
	Observed N	Expected N	Residual
1.00	43	50.0	-7.0
2.00	49	50.0	-1.0
3.00	56	50.0	6.0
4.00	45	50.0	-5.0
5.00	66	50.0	16.0
6.00	41	50.0	-9.0
Total	300		

Test Statistics (检验统计量)	
	Y
Chi-Square	8.960
Df (自由度)	5
Asymp. Sig. (渐近显著性水平)	.111

这里的 Asymp. Sig. (The significance level based on the asymptotic distribution of a test statistic)是基于卡方统计量的渐近分布的实际显著性水平 (渐近 p 值), 它以数据集为一个大大样本的假设为基础。因为 $p=0.111>\alpha=0.05$, 所以认为该六面体是均匀的。

二、二项分布检验(Binomial Test)

实际问题中,有许多总体是由二项式组成的。例如,是与非、男与女、正面与背面、正确与错误等等。这种总体通常就称为二项总体。对于一个二项总体,如果其中的一类所占所占比重为 P , 则另一类的比重一定是 $Q=1-P$ 。在既定总体中, P 是一个定值。然而,从该总体中任意抽取一个随机样本,所得到的样本比率 P , 却是一个随机变量。因为样本仅是总体的一小部分,基于样本得到的信息 P , 不会刚好等于总体的 P , 二者之间难免出现误差,这种误差称为抽样误差。理论上已经证明,二者之间出现较小误差的概率比较大,而出现较大误差的概率相对来说就比较小,这就是通常所说的“小概率不可能出现”的原理。当研究对象属于二项总体时,可以用二项分布来检验假设,判断所抽取的样本是否来自具有既定值的总体。其检验步骤如下:

1、提出假设

$$H_0: P=P_t \quad (0 \leq P_t \leq 1)$$

$$H_1: P \neq P_t$$

2、计算统计量值和 p 值

设 n_1 和 n_2 分别为第 1 类和第 2 类的观测值个数, $N = n_1 + n_2$, $m = \min(n_1, n_2)$,

$$p_t^* = \begin{cases} p_t & \text{若 } m = n_1 \\ 1 - p_t & \text{若 } m = n_2 \end{cases},$$

$$\text{双侧准确概率 } p = 2 \left(\sum_{i=0}^m C_N^i p_t^{*i} (1 - p_t^*)^{N-i} \right) - C_N^m p_t^{*m} (1 - p_t^*)^{N-m}$$

双侧近似概率计算如下：

$$Z_1 = \frac{n_1 + 0.5 - Np_t}{\sqrt{Np_t(1 - p_t)}}, \quad Z_2 = \frac{n_1 - 0.5 - Np_t}{\sqrt{Np_t(1 - p_t)}}$$

$F(Z_i)$ 为标准正态分布的累积概率, 那么

$$P(X = n_1) = F(Z_2) - F(Z_1)$$

$$P(X \leq n_1) = F(Z_1)$$

$$\text{双侧近似概率 } p = 2P(X \leq n_1) - P(X = n_1)$$

3、根据 p 值作出统计判断。

[例 11-22] 掷一枚球类比赛用的挑边器 40 次, 出现 A 面和 B 面在上的次数如表 11-62 所示, 试问这枚挑边器是否均匀?

表 11-62 掷一枚挑边器的结果

1 1 0 1 1 0 1 1 1 1 0 1 1 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 0 1 1 0

其中: 0 表示 A 面向上, 1 表示 B 面向上。

解: (1) 在 SPSS 中输入表 11-62 中的数据 (变量名为 Y)。选择主菜单的 [Analyze] => [Nonparametric Tests] => [Binomial Test]。

(2) 显示如图 11-113 所示的 [Binomial Test (二项检验)] 主对话框, 把 Y 选入 [Test Variable], 其它选项采用默认值。

(3) 单击主对话框中的 [OK] 按钮, 输出结果如下:

Binomial Test

	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
Y	Group 1	1	28	.70	.018
	Group 2	0	12	.30	
	Total		40	1.00	

a Based on Z Approximation.

从结果可以看出， $p=0.018 < \alpha=0.05$ ，认为该挑边器不是均匀的。

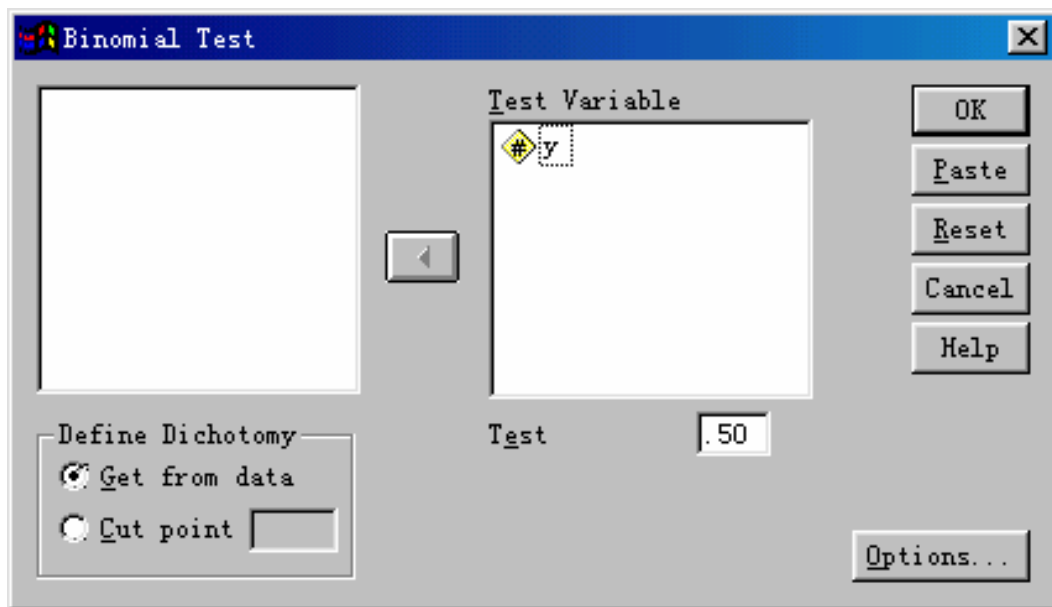


图 11-113

三、游程检验(Run Test)

游程检验是一种利用游程的总个数来判断样本随机性的统计检验方法。所谓游程，就是指在样本单位的抽取序列中，某一类型的单位被另一类型单位在其前后隔开所形成的一个连续串。例如，令 $x_1, x_2, \dots, x_n$ 为样本容量 n 的一个随机样本的观察值，假设它存在两种不同类型的单位，一类记为 A，另一类记为 B。这样，当将其按任何顺序排列时，可以得到一个由 A 和 B 两种元素组成的序列。形成的序列有如下几种可能的典型方式(假设 A 的单位数为 $n_1 = 8$ ，B 的单位数为 $n_2 = 7$)：

第一种情况：AAAAAAAAABBBBBBBB；

第二种情况：AAAABBBBAAAABBB；

第三种情况：ABBAAABABBBABAA；

第四种情况：ABABABABABABABA。

在第一种情况中，A 的游程数为 $R_1=1$ ，B 的游程数为 $R_2=1$ ；

在第二种情况中，A 的游程数为 $R_1=2$ ，B 的游程数为 $R_2=2$ ；

在第三种情况中，A 的游程数为 $R_1=5$ ，B 的游程数为 $R_2=4$ ；

在第四种情况中，A 的游程数为 $R_1=8$ ，B 的游程数为 $R_2=7$ 。

设 R 为总游程数， $R=R_1+R_2$ 。在第一种情况中， $R=1+1=2$ ；第二种情况下， $R=2+2=4$ ；第三种情况中， $R=5+4=9$ ；第四种情况中， $R=8+7=15$ 。显然， R 的最小值为 2，最大值在 $n_1 \neq n_2$ 时为 $2 \min(n_1, n_2)+1$ ，在 $n_1 = n_2$ 时为 $n_1 + n_2$ 。

游程检验的基本原理是这样的：如果我们希望从总体的一个样本所包含的信息中得出关于该总体的某些结论，或是要判别两个样本是否来自同一个总体，那么所采用的样本必须是随机样本。游程检验法使得我们能够检验“样本是随机的”这一假设。在任一既定大小的样本中，游程总数标志着样本是否是随机样本。如果游程总数太少，例如上述的第一、第二两种情况，它意味着样本中包含着某种主观的带有倾向性的因素，缺乏独立性，因此，肯定不是随机的样本。同理，如果游程总数太多，达到最大值，例如上述的第四种情况，也同样有理由认为这是由于有系统的短周期波动影响着观察的结果。也就是说，游程总数太少或太多的样本序列绝对不是随机的序列。为了知道 R 是否太少或太多，即检验样本序列的随机性，必须了解游程总数 R 的概率分布。实际检验步骤如下：

(1) 提出假设

H_0 ：样本是随机的；

H_1 ：样本不是随机的。

(2) 构造统计量并计算 p 值

用于把样本数据分成两类（A 和 B）的分割点可以是指定的某个具体数值，也可以是均值、中位数、众数等。当 $x_i >$ 分割点时设为 A 类，否则为 B 类，其相应的单位数分别为 n_1 和 n_2 。在大样本情况下，游程总数 R 的分布接近于正态分布，其数学期望和方差分别为：

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

因此， Z 统计量为 $Z = \frac{R - E(R)}{\sqrt{Var(R)}}$ 。

由于在游程检验中将一个样本各单位归属于两种类别之中，所以样本各单位的分布成为二项分布。在样本容量足够大时， R 的分布接近正态分布。但是当样本容量不太大时，应当作连续性校正，这时 Z 的统计量公式为：

$$Z = \begin{cases} (R - E(R) + 0.5) / \sqrt{Var(R)} & \text{当 } R - E(R) \leq 0.5 \text{ 时} \\ (R - E(R) - 0.5) / \sqrt{Var(R)} & \text{当 } R - E(R) > 0.5 \text{ 时} \\ 0 & \text{当 } |R - E(R)| < 0.5 \text{ 时} \end{cases}$$

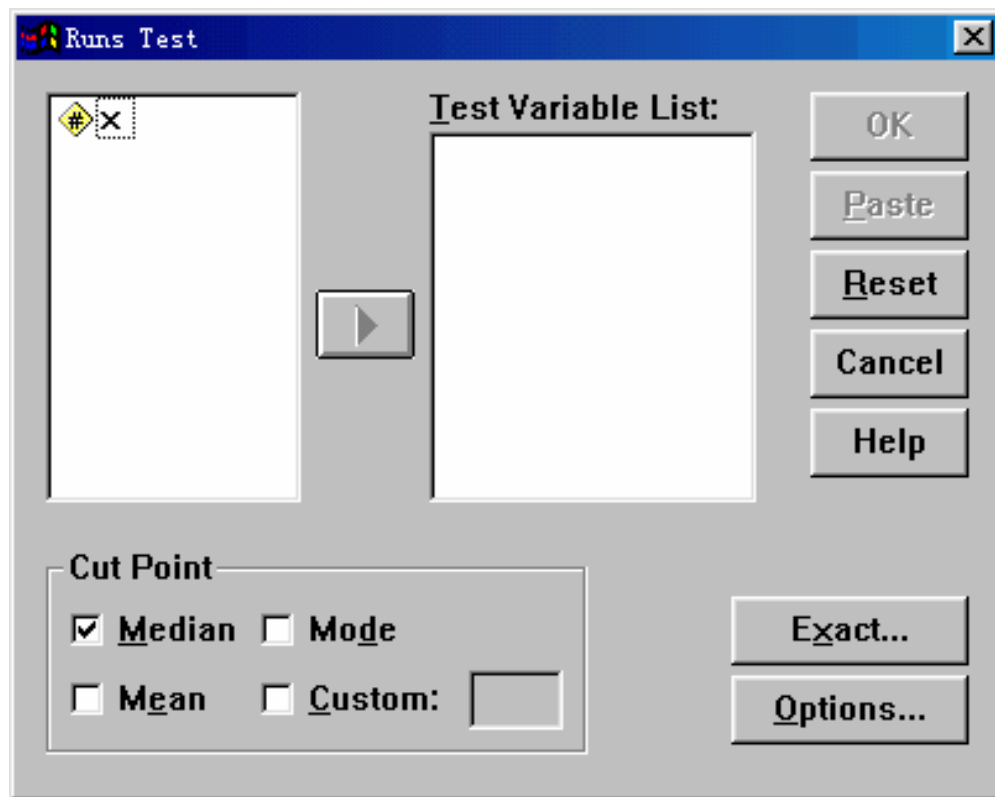
(3) 作出判断

[例 11-23] 假设从总体中抽取一个样本，记录其先后出现的样本值如下，试利用游程检验法来检验样本序列的随机性。（ $\alpha=0.05$ ）

31 23 36 43 51 44 12 26 43 75 2 3
15 18 78 24 13 27 86 61 13 7 6 8

解：(1) 在 SPSS 中输入数据（变量名为 X ），然后选择主菜单 [Analyze]⇒[Nonparametric Tests]⇒[Runs]；

(2) 在显示的 [Runs Test (游程检验)] 主对话框中，把变量 X 选择入 [Test Variable (检验变量)] 列表框中，并采用默认的分割点 (Cut point)：中位数 (Median)。



(3) 单击主对话框中 [OK] 按钮，输出结果如下：

Runs Test	
	X
Test Value (检验值 : 中位数)	25.0000
Cases < Test Value	12
Cases >= Test Value	12
Total Cases	24
Number of Runs (总游程数 R)	10
Z	-1.044
Asymp. Sig. (2-tailed) (p 值)	.297

根据输出结果, $p=0.297>\alpha=0.05$, 所以接受零假设, 即样本是随机的。

四、单样本柯尔莫哥诺夫—斯米尔诺夫检验(One-sample K-S test)

柯尔莫哥诺夫—斯米尔诺夫检验(Kolmogorov-Smirnov Test, 简称 K-S 检验)用于检验一组样本观测结果的经验分布同某一指定的理论分布(如正态分布、均匀分布、泊松分布、指数分布)之间是否一致。K-S 检验的基本思路为: 将顺序分类数据的理论累积频率分布同观测的经验累积频率分布加以比较, 求出它们最大的偏离值, 然后在给定的显著性水平上检验这种偏离值是否的偶然出现的。

设理论累积频数分布为 $F(x)$, n 次观测的随机样本的经验分布函数 $F_n(x)$, K-S 检验的步骤如下:

(1) 零假设 H_0 : 经验分布与理论分布没有显著差别。

(2) 把样本观测值从小到大排列为: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, 计算经验累积分布函数

$$F_n(x) = \begin{cases} 0 & -\infty < x < x_{(1)} \\ i/n & x_{(i)} \leq x < x_{(i+1)} \\ 1 & x_{(n)} \leq x < +\infty \end{cases} \quad i = 1, 2, \dots, n-1$$

和理论累积分布函数 $F(x)$ 。

记 $D = \max |F_n(x_i) - F(x_i)|$ ($i = 1, 2, \dots, n$), 则检验统计量为

$$Z = D\sqrt{n}$$

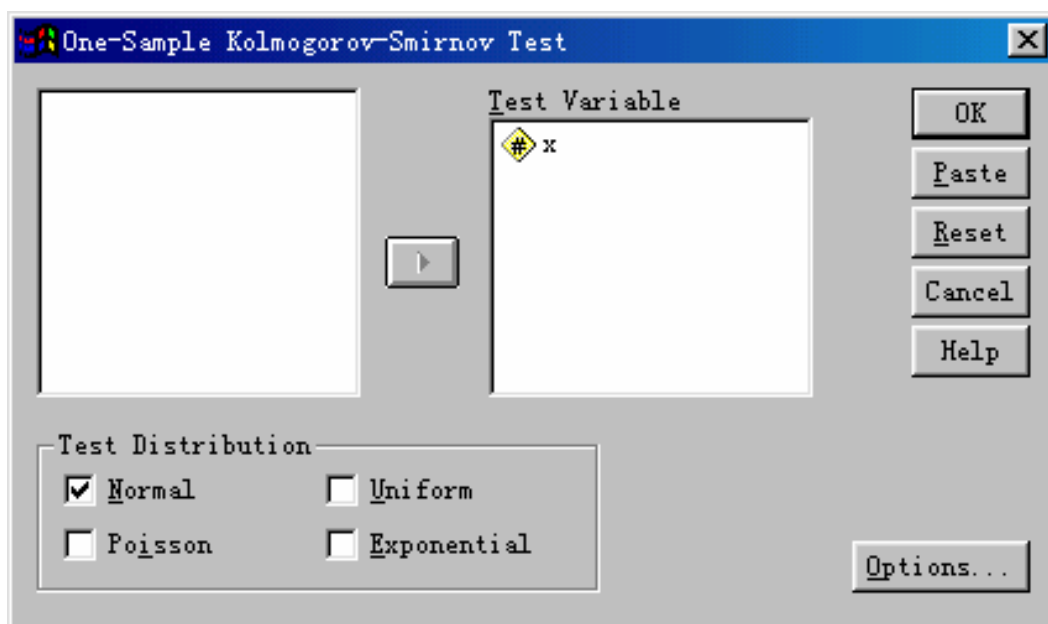
双侧显著性水平 (p 值) 根据 Smirnov 提出的公式计算, 这里从略。

(3) 作出判断。

[例 11-24] 检验[例 11-23]的样本数据是否来自正态总体。

解: (1) 在 SPSS 中输入数据(变量名为 X), 选择[Analyze]⇒[Nonparametric Tests]⇒[1-Sample K-S]。

(2) 在[One-Sample Kolmogorov-Smirnov Test (单样本 K-S 检验)]主对话框中, 把变量 X 选入[Test Variable]列表框中, 并选择[Test Distribution(检验分布)]中的[Normal(正态分布)]。



(3) 单击主对话框中的[OK]按钮, 输出结果如下:

One-Sample Kolmogorov-Smirnov Test		
		X
N		24
Normal Parameters ^{a,b}	Mean	31.0417
	Std. Deviation	24.5790
Most Extreme Differences	Absolute	.149
	Positive	.149
	Negative	-.119
Kolmogorov-Smirnov Z		.728
Asymp. Sig. (2-tailed)		.664

a Test distribution is Normal.

b Calculated from data.

由结果： $p=0.664>\alpha$ ，所以认为样本来自正态分布总体。

五、两个独立样本检验

虽然有时样本所属的总体的分布类型往往是不明的，但我们还是想知道在这种情况下两个独立样本是否来自相同分布的总体，Mann-Whitney U 检验、Kolmogorov-Smirnov Z 检验、Moses Extreme Reactions 检验和 Wald-Wolfowitz 游程检验等就是用于处理此类问题的有效方法。其中 Mann-Whitney U 检验是处理该问题中最常用的方法。这些方法的基本假设有：(1) 随机抽样；(2) 两个样本是独立的；(3) 数据变量为定序变量或更高层次的变量。

Mann-Whitney U 检验又称为秩和 U 检验，用于检验两个独立样本是否来自相同的总体（与 t 检验类似）；Kolmogorov-Smirnov Z 检验，用于推测两个样本是否来自具有相同分布的总体；Moses extreme reactions 检验两个独立样本之观察值的散布范围是否有差异存在，以检验两个样本是否来自具有同一分布的总体；Wald-Wolfowitz 游程检验考察两个独立样本是否来自具有相同分布的总体。这些方法的检验步骤为：

(1) 提出假设

H_0 ：两个独立样本来自相同的总体

H_1 ：两个独立样本来自不同的总体

(2) 计算相应检验统计量值或 p 值

(3) 作出判断

若 $p>\alpha$ ，接受 H_0 ，认为两个样本来自相同的总体；否则，拒绝 H_0 ，认为两个样本来自不同的总体。

[例 11-25]设有甲、乙两种安眠药，要比较它们的治疗效果。现独立观察 20 个失眠者（其中 10 人服用甲药，另 10 人服用乙药），服用安眠药后睡眠时间延长的时数如所示。现延长的睡眠时数的分布情况不明，试问这两种药物的疗效有无显著性差异？

表 11-63 服用安眠药后延长时数表

序号	1	2	3	4	5	6	7	8	9	10
A	1.9	0.8	1.1	0.1	0.1	4.4	5.5	1.6	4.6	3.4
B	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

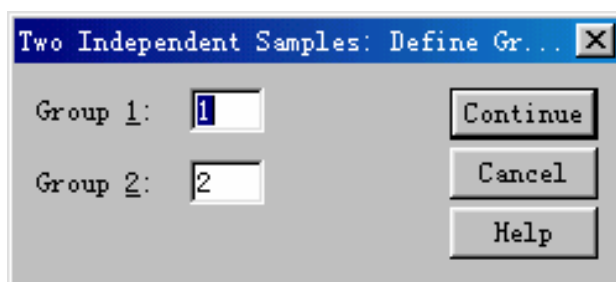
A 表示失眠者服用甲药后睡眠时间延长的时数；B 表示失眠者服用乙药后睡眠时间延长的时数。

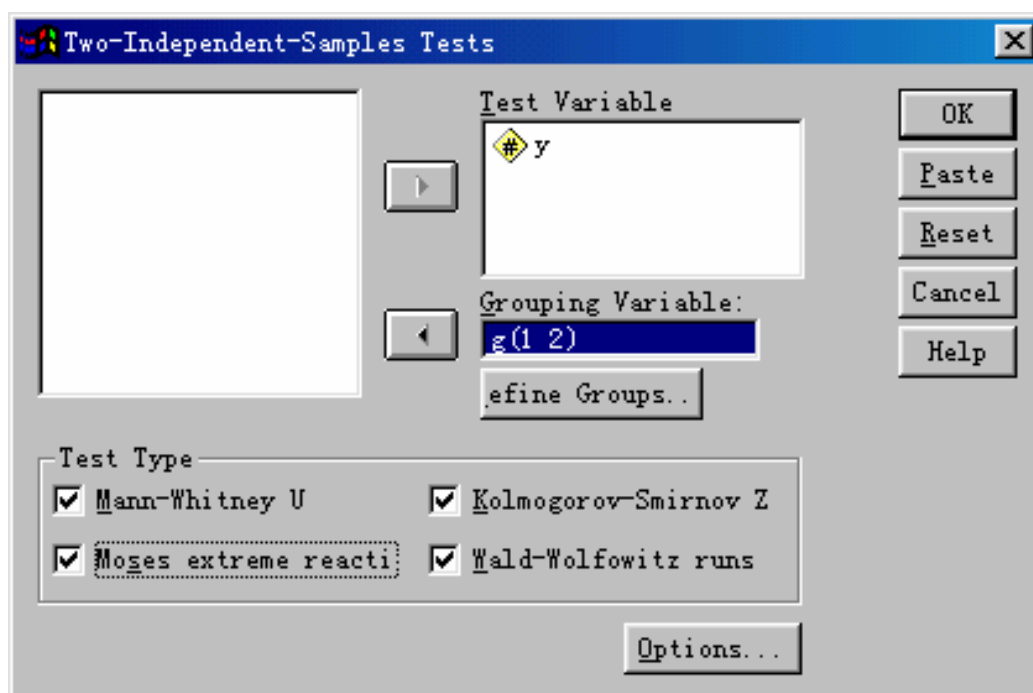
操作步骤：

(1) 录入数据。服用安眠药后时间延长的变量为 Y，用变量 G 表示所对应的实验组，G=1 表示失眠者服用甲药组别，G=2 表示失眠者服用乙药组别（如下图所示）。

	y	g
1	1.90	1
2	.80	1
...	...	...
9	4.60	1
10	3.40	1
11	.70	2
12	-1.60	2
...	...	...
19	.00	2
20	2.00	2

(2) 选择主菜单 [Analyze] => [Nonparametric Tests] => [2 Independent Samples]。在[Test Type(检验类型)]中选择四种检验方法。把 Y 选入[Test Variable]列表框，把 G 选入[Grouping Variable]并单击[Define Groups(定义组)]按钮。在定义组对话框中[Group 1]的右框中输入 1，在[Group 2]的右框中输入 2，并单击[Continue]返回主对话框。





(3) 单击主对话框中的[OK]按钮，输出结果如下：

Mann-Whitney Test

	Y
Mann-Whitney U	24.000
Wilcoxon W	79.000
Z	-1.968
Asymp. Sig. (2-tailed)	.049
Exact Sig. [2*(1-tailed Sig.)]	.052

Moses Test

		Y
Observed Control Group Span		15
	Sig. (1-tailed)	.070
Trimmed Control Group Span		14
	Sig. (1-tailed)	.500
Outliers Trimmed from each End		1

Two-Sample Kolmogorov-Smirnov Test

		Y
--	--	---

Most Extreme Differences	Absolute	.500
	Positive	.000
	Negative	-.500
Kolmogorov-Smirnov Z		1.118
Asymp. Sig. (2-tailed)		.164

a. Grouping Variable: G

Wald-Wolfowitz Test

		Number of Runs	Z	Exact Sig. (1-tailed)
Y	Minimum Possible	8 [*]	-1.149	.128
	Maximum Possible	10 [*]	-.230	.414

*There are 2 inter-group ties involving 4 cases.

因四种方法计算出来的 p 值均大于 0.05，所以可以认为这两种药物的疗效无显著性的差异。

六、多个独立样本检验

多个独立样本检验方法主要有：Kruskal-Wallis H 检验、中位数 (Median) 检验和 Jonckheere-Terpstra 检验。Kruskal-Wallis H 检验为单向方差分析，检验多个样本在中位数上是否有差异；中位数检验法用于检验多个样本是否来自具有相同中位数的总体；Jonckheere-Terpstra 检验法用于检验多个独立样本是否来自相同总体，它适用于定量数据和定序分类数据，当要检验的多个总体是定序变量时，Jonckheere-Terpstra 检验法比 Kruskal-Wallis H 检验法更为有效。

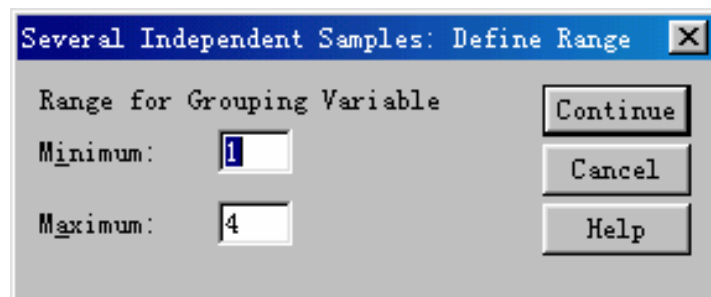
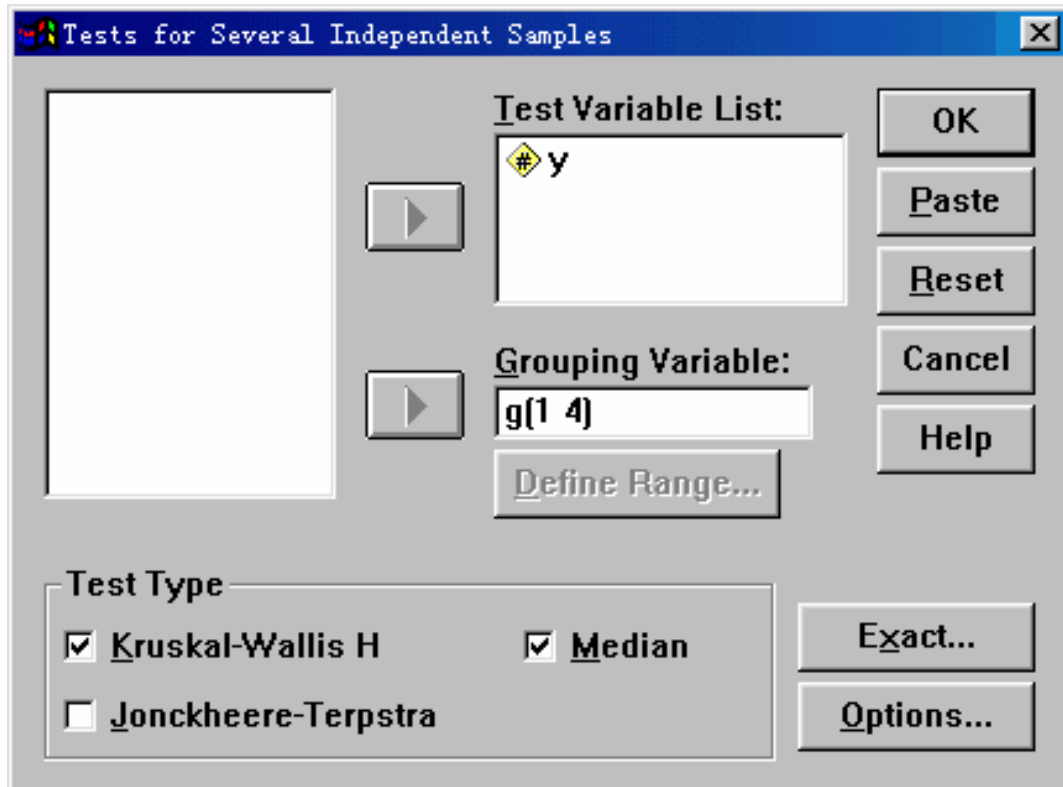
[例 11-26] 消费者协会采用 1 到 20 分来评价四家冷藏食品公司的油炸鸡。他们相求出这些公司的鸡在质量上是否有所不同。表 11-64 给出了四家公司的评价。($\alpha=0.05$)

表 11-64 四家冷藏食品公司油炸鸡的评价

公司	评分 Y	G
A	2 2 5 6 10	1
B	18 19 16 20 12 18	2
C	18 15 17 12 14 12 11	3
D	4 1 3 8 7 8 9	4

解：(1) 变量 Y 表示评分，G 表示相应的公司。在 SPSS 中录入数据。

(2) 选择[Analyze]>[Nonparametric Tests]>[K Independent Samples]。在对话框中,在[Test Type]中选择[Kruskal-Wallis H]和[Median] ,把 Y 选入[Test Variable] ;把 G 选入[Grouping Variable(分类变量)]并单击[Define Range(定义范围)],在定义范围对话框的[Minimum]的右框中输入 1 ,在[Maximum]的右框中输入 4 ,单击[Continue]返回主对话框。



(3) 单击[OK] , 输出结果如下 :

Kruskal-Wallis Test

Ranks

	G	N	Mean Rank
Y	1	6	7.83
	2	6	22.00
	3	7	17.71
	4	7	6.86
	Total	26	

Test Statistics

	Y
Chi-Square	18.171
df	3
Asymp. Sig.	.000

a Kruskal Wallis Test

b Grouping Variable: G

Median Test

Frequencies

		G			
		1	2	3	4
Y	> Median	1	6	6	0
	<= Median	5	0	1	7

Test Statistics

	Y
N	26
Median	11.5000
Chi-Square	19.238
df	3
Asymp. Sig.	.000

a 8 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.0.

b Grouping Variable: G

从结果可以看出，两种检验方法的 p 值均小于 0.05，所以拒绝零假设，认为四家公司的产品之间有显著性的差异。

七、两个相关样本检验

两个相关样本检验的方法主要有：Wilcoxon 检验、Sign (符号) 检验、McNemar 检验和 Marginal Homogeneity 检验等。Wilcoxon 检验用于检验两个相关样本是否来自相同的总体，但对总体分布形式没有限制；Sign 检验通过计算两个样本的正负符号的个数来检验两个样本是否来自相同总体；McNemar 检验用于两个相关二分变量的检验；Marginal Homogeneity 检验用于两个相关定序变量的检验，是 McNemar 检验的扩展。

[例 11-27]为研究长跑运动对增强普通高校学生的心功能效果，对某院 15 名男生进行实验，经过 5 个月的长跑锻炼后看其晨脉是否减少。锻炼前后的晨脉数据如表 11-65 所示。

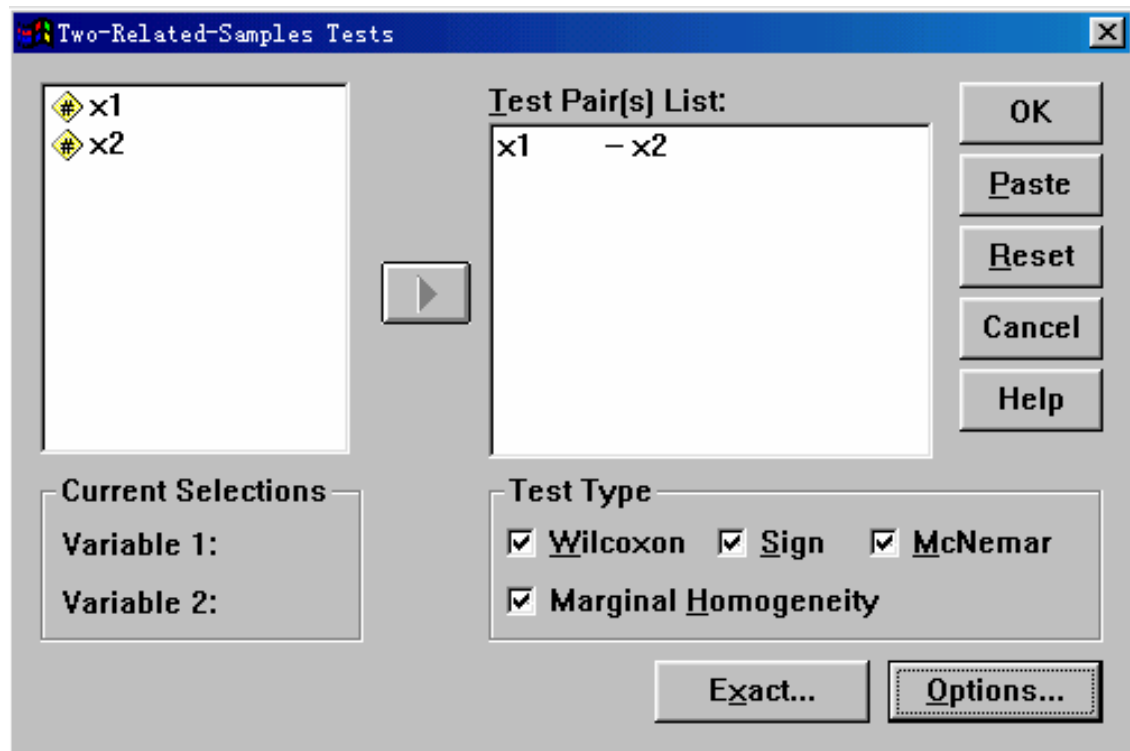
表 11-65 高校学生的晨脉数据

锻炼前	70	76	56	63	63	56	58	60	65	65	75	66	56	59	70
锻炼后	48	54	60	64	48	55	54	45	51	48	56	48	64	50	54

SPSS 操作步骤如下：

(1) 输入数据，变量 X1 表示锻炼前晨脉数据，变量 X2 表示锻炼后晨脉数据。

(2) 选择[Analyze]>[Nonparametric Tests]>[2 Related Samples]。在显示的[Two-Related-Samples Test]先后单击变量 X1 和 X2，在[Current Selections]框中的[Variable 1]和[Variable 2]中依次出现所选择的两个相关变量，然后单击右边一个右箭头按钮，变量名被选入[Test Variable List]列表框中；选择[Test Type]框中的[Wilcoxon]、[Sign]、[McNemar]和[Marginal Homogeneity]检验方法。



(3) 单击[OK]按钮，输出结果如下：

Warnings

The McNemar Test for X1 & X2 is not performed because both variables are not dichotomous with the same values.

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
X2 - X1	Negative Ranks	12	9.17	110.00
	Positive Ranks	3	3.33	10.00
	Ties	0		
	Total	15		

- a $X2 < X1$
- b $X2 > X1$
- c $X1 = X2$

Test Statistics

	X2 - X1
Z	-2.842
Asymp. Sig. (2-tailed)	.004

- a Based on positive ranks.
- b Wilcoxon Signed Ranks Test

Sign Test

Frequencies

		N
X2 - X1	Negative Differences	12
	Positive Differences	3
	Ties	0
	Total	15

- a $X2 < X1$
- b $X2 > X1$
- c $X1 = X2$

Test Statistics

	X2 - X1
Exact Sig. (2-tailed)	.035

- a Binomial distribution used.
- b Sign Test

Marginal Homogeneity Test

	X1 & X2
Distinct Values	17
Off-Diagonal Cases	15
Observed MH Statistic	799.000
Mean MH Statistic	878.500
Std. Deviation of MH Statistic	27.491
Std. MH Statistic	-2.892
Asymp. Sig. (2-tailed)	.004

从输出结果可以看出, $p < 0.05$, 说明经过 5 个月的长跑锻炼后学生的晨脉减少了。

八、多个相关样本检验

多个相关样本的检验方法有:Friedman 检验、Kendall W 检验和 Cochran Q 检验等。Friedman 检验为双向方差分析, 考察多个相关样本是否来自同一总体; Cochran Q 检验作为两相关样本 McNemar 检验的多样本推广, 特别适用于定性变量和二分字符变量; Kendall W 检验, 通过计算 Kendall 和谐系数 W, 以检验多个相关样本是否来自同一分布的总体。

[例 11-28]某商店想了解顾客对几种款式不同的衬衣的喜爱程度。某日询问了 9 名顾客, 请它们对 3 种款式的衬衣按喜爱程度排次序(最喜爱的给秩 1, 其次的给秩 2, 再次的给秩 3, 结果如表 11-66 所示, 试问顾客对 3 种款式的衬衣的喜爱程度是否相同?

表 11-66 顾客对衬衣的喜爱程度

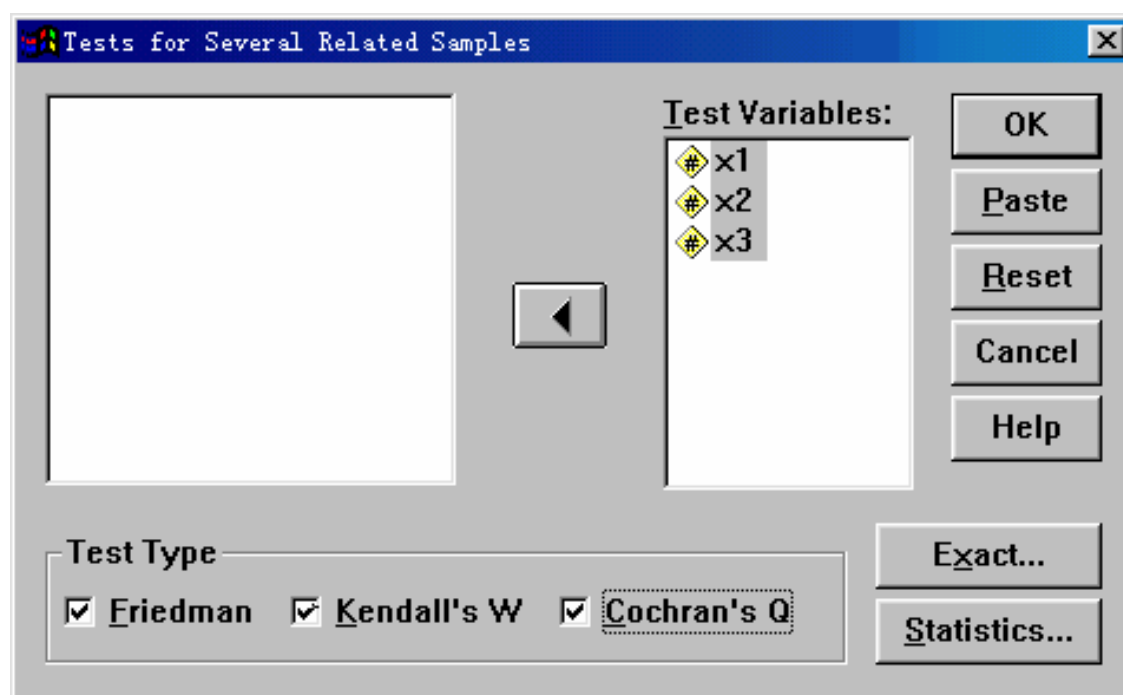
顾客号	1	2	3	4	5	6	7	8	9
款式 1	1	2	2	1	3	1	2	1	1
款式 2	3	1	3	3	2	2	3	3	3
款式 3	2	3	1	2	1	3	1	2	2

解:(1) 在 SPSS 按左图方式输入数据(变量名分别为 X1、X2、X3)。

	x1	x2	x3
1	1	3	2
2	2	1	3
3	2	3	1
4	1	3	2
5	3	2	1
6	1	2	3
7	2	3	1
8	1	3	2
9	1	3	2

(2) 选择[Analyze]=>[Nonparametric Tests]=>[K Related Samples]。在显示的主对话框中, 选择[Test Type]栏中的[Friedman]、[Kendall's W]和

[Cochran's Q]。单击[OK]按钮。



(3) 输出结果如所示：

Warnings

The Cochran Test for X1 X2 X3 is not performed because all variables are not dichotomous with the same values.

Friedman Test			
Ranks		Test Statistics	
	Mean Rank	N	9
X1	1.56	Chi-Square	4.667
X2	2.56	df	2
X3	1.89	Asymp. Sig.	.097

Kendall's W Test

Ranks

	Mean Rank
X1	1.56
X2	2.56
X3	1.89

Test Statistics

N	9
Kendall's W	.259
Chi-Square	4.667
df	2

Asymp. Sig.	.097
-------------	------

a Kendall's Coefficient of
Concordance

本例无法进行 Cochran 检验 ,Friedman 和 Kendall W 检验的 p 值均大于 0.05 ,
所以可认为顾客对 3 种款式的衬衣的喜爱程度是不相同的。

第十二章 聚类分析

第一节 聚类分析概述

一、聚类分析的概念

俗话说：“物以类聚，人以群分”。分类是人们认识世界的基础。在社会、经济及自然现象的研究中，存在着大量分类研究的问题。例如，为了研究不同地区农民家庭不同收入的分布规律，需要对不同地区、不同农民家庭、不同收入进行分类；在制订农业发展区划时，需要根据不同地区的气候条件、土壤类型、粮食产量水平、灌溉水平、经济物质条件等对各地区进行分类；等等。尽管传统的分类方法起源很早，但利用数学和计算机手段对复杂对象进行定量分类的方法还只有几十年的历史。过去人们主要靠经验和专业知识进行定性分类处理，致使许多分类带有主观性和任意性，不能很好地提示客观事物内在的本质差别与联系，特别是对于多因素、多指标的分类问题。为了克服定性分类的不足，有必要引入数学方法，形成了数值分类法。

数值分类一般有两种情况：一是已知研究对象的分类情况，需将某些未知个体正确地归属于其中某一类，是一种有师分类；二是研究对象不存在事前分类的情况，而将数据进行结构性分类，是一种无师分类。对于前者，属判别分析(Discriminant Analysis)的内容；而后者则属于聚类分析的内容。聚类分析是研究“物以类聚”的一种多元统计分析方法。

聚类分析的基本思想是根据对象间的相关程度进行类别的聚合。在进行聚类分析之前，这些类别是隐蔽的，能分为多少种类别事先也是不知道的。聚类分析的原则是同一类中的个体有较大的相似性，不同类中的个体差异很大。例如，有 A、B、C、D、E 五个地区（即样本单位或样品）的农民家庭收入（指标或变量 X_1 ）分别为 X_{11} 、 X_{21} 、 X_{31} 、 X_{41} 、 X_{51} （ X_1 的取值），农民家庭人口（指标或变量 X_2 ）分别为 X_{12} 、 X_{22} 、 X_{32} 、 X_{42} 、 X_{52} （ X_2 的取值），而的数值高低不一，参差不齐，不易综合判断五个地区的家民家庭收入水平状态，为此，可以运用一定的方法将相似程度较大的数

据或单位划为一类，划类时关系密切的聚合为一小类，关系疏远的聚合为一大类，直到把所有的数据或单位聚合为唯一的类别。这种分类就是最常用最基本的一种聚类分析方法——系统聚类分析（或称为分层聚类分析）的内涵。此外还有动态聚类法、模糊聚类法、有序聚类法等等。

系统聚类法的具体聚类过程是：聚类开始时，样本中的各个样品（或变量）自成一类；通过计算样品（或变量）间的相似性测度，把其中最相似的两个样品（或变量）进行合并，合并后，类的数目就减少一个；重新计算类与类之间的相似性测度，再选择其中最相似的两类进行合并，……，这种计算、合并的过程重复进行，直至所有的样品（或变量）归为一类。整个聚类过程可以用聚类图（树图）形象地描绘出来。

根据分类对象的不同可把聚类分析分为样品聚类和变量聚类，若我们把样本数据按二维表形式（如表 15-67 所示）排列，则相当于分别对表 15-67 的行和列进行聚类。实际中应用较多的是样品聚类分析。

表 15-67 聚类分析样本数据表

变量 样品	变量(指标)1	变量(指标)2		变量(指标) p
样品(单位)1	X_{11}	X_{12}		X_{1p}
样品(单位)2	X_{21}	X_{22}		X_{2p}
.....				
样品(单位) n	X_{n1}	X_{n2}		X_{np}

1、样品聚类

样品聚类又称为 Q 型聚类，就是对样本单位的观测量进行聚类，是根据被观测的对象的特征，即反映被观测对象的特征各变量值进行分类。不同的分析目的选用不同的指标（变量）作为分类的依据。

2、变量聚类

变量聚类又称为 R 型聚类。反映同一事物特点的变量有很多，我们往往根据所研究的问题选择部分变量对事物的某一方面进行研究。由于人类对客观事物的认识是有限的，往往难以找出彼此独立的有代表性的变量，而影响对问题的进一步认识和研究。例如，在回归分析中由于自变量的共线性导致偏回归系数不能真正反映自变量对因变量的影响等。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息。在生产活动中也不乏需要进行变量聚类的实例。如制衣业制定衣服

本书对判别分析不予以介绍，代之以第十章的 Logistic 回归分析。

型号就是根据人体各部分尺寸数据找出最有代表性的指标如身长、胸围和裤长、腰围作为上衣和裤子的代表性指标。

二、聚类分析的一般步骤

不管是 Q 型聚类分析还是 R 型聚类分析，一般来讲，分析过程可以分为三个步骤：

1、数据变换处理。在聚类分析过程中，需要对各个原始数据进行一些相互比较运算，而各个原始数据往往由于计量单位不同而影响这种比较和运算。因此，需要对原始数据进行必要的变换处理，以消除不同计量单位对数据值大小的影响。

2、计算聚类统计量。聚类统计量是根据变换以后的数据计算得到的一个新数据。它用于表明各样品或变量间的关系密切程度。常用的统计量有距离和相似系数两大类。

3、选择聚类方法。根据聚类统计量，运用一定的聚类方法，将关系密切的样品或变量聚为一类，将关系不密切的样品或变量加以区分。选择聚类方法是聚类分析最终的、也是最重要的一步。

第二节 数据变换处理

为了克服原始数据由于计量单位的不同对聚类分析结果产生不合理的影响。在聚类分析过程中，首先应对原始数据进行数据变换处理。

设原始数据（如表 15-67 所示）构成如下数据矩阵：

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad (15-116)$$

其中： n 为样品数， p 为原始变量个数， X_{ij} 表示第 i 个单位在第 j 个变量上的数据值。

所谓数据变换，就是将原始数据矩阵中的每个元素，按照某种特定的运算，把它变为一个新值，而且数值的变化不依赖于原始数据集合中其它数据的新值。对原始数据(15-116)进行变换的方法主要有把数值变换为 Z 分数（标准化变换）、变换到 $0 \sim 1$ 范围内（规格化变换）、变换到 $-1 \sim +1$

范围内、变换到最大值为 1、变换到均值为 1 或标准差为 1 等等，下面仅介绍前两种方法。

一、标准化变换

标准化变换把原始数据转换为标准 Z 分数(Z score)变换方法。其变换公式为：

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \quad (i=1,2,\dots,n, \quad j=1,2,\dots,p) \quad (15-117)$$

其中： X'_{ij} 表示标准化数据、 $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ 表示变量 j 的均值， S_j 表示变量 j 的标准差即：

$$S_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}$$

将式 (4-29) 用矩阵表示，则有：

$$\mathbf{X}' = \begin{bmatrix} \frac{X_{11} - \bar{X}_1}{S_1} & \frac{X_{12} - \bar{X}_2}{S_2} & \dots & \frac{X_{1p} - \bar{X}_p}{S_p} \\ \frac{X_{21} - \bar{X}_1}{S_1} & \frac{X_{22} - \bar{X}_2}{S_2} & \dots & \frac{X_{2p} - \bar{X}_p}{S_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{X_{n1} - \bar{X}_1}{S_1} & \frac{X_{n2} - \bar{X}_2}{S_2} & \dots & \frac{X_{np} - \bar{X}_p}{S_p} \end{bmatrix} \quad (15-118)$$

不难看出，经过标准化变换后的数据矩阵式(15-118)的每列数据的平均值为 0，方差为 1。使用标准化变换处理后，消除了数据计量单位不同的影响，便于数据的直接比较。因此标准化变换方法在实际中应用最多。

二、规格化变换

规格化变换又称为极差正规比变换。它是从数据矩阵中的每一个变量中找出其最大值和最小值，并用最大值减去最小值得出极差(Range)。然后以每一个原始数据减去该变量中的最小值，再除以极差，即得规格化数据。

设原始数据矩阵仍为(15-116)，规格化数据为 X'_{ij} ，则规格化数据的计算公式如下：

$$X'_{ij} = \frac{X_{ij} - \min_{1 \leq i \leq n} \{X_{ij}\}}{\max_{1 \leq i \leq n} \{X_{ij}\} - \min_{1 \leq i \leq n} \{X_{ij}\}} \quad (i=1,2,\dots,n; j=1,2,\dots,p) \quad (15-119)$$

经过规格化变换后，将每列的最大数据变为 1，最小数据变为 0，其余数据取值在 0~1 之间。规格化变换后的数据也消除了计量单位的影响。

上面两种变换方法都是通过变量进行变换处理，也可以通过样品进行变换处理，例如标准化变换公式（4-29）中的均值和标准差采用样品的均值和标准差而不是变量的均值和标准。实际中应用较多的是通过变量进行变换。

第三节 聚类统计量

研究样品或变量疏密程度的数量指标有两大类，一类是距离，另一类是相似系数。这两大类指标就是用于反映各样品或各变量间差别大小的统计量。变量的测量尺度不同，所采用的统计量也就不同。

为了方便起见，我们仍用原始数据矩阵的符号来表示变换处理后的数据矩阵：

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad (15-120)$$

式(15-120)中， $\mathbf{X}$ 表示变换处理后的数据矩阵， n 为样品数， p 为变量的个数， X_{ij} 表示第 i 个样品在第 j 个变量上的变换后的数据值。那么任意两个样品之间的相似性大小，都可以通过计算任意两行或两列之间的距离和相似系数来反映（下面以样品为例）。

一、定距(Interval)、定比(Ratio)变量的聚类统计量

定距、定比变量的聚类统计量可以分为两类：距离和相似系数。距离通常用于样品聚类分析，而相似系数用于变量聚类分析。

（一）距离(Distance)

距离的计算方法多种多样，但常用方法主要有三种，即欧氏距离、明

考斯基距离、绝对值距离、切比雪夫距离等等。

欧氏距离 (Euclidean distance) 是聚类分析中用得最广泛的距离。如果仍根据式(15-120)的变换数据矩阵计算第 i 行和第 k 行的欧氏距离, 则有欧氏距离公式为:

$$d_{ik} = \sqrt{\sum_{j=1}^p (X_{ij} - X_{kj})^2} \quad (15-121)$$

将所有行之间的欧氏距离都算出, 同样可以得到一个 $n \times n$ 的欧氏距离矩阵:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \quad (15-122)$$

其中 d_{ij} ($i=1,2,\cdots,n; j=1,2,\cdots,n$) 表示式(15-120)中第 i 行和第 j 行的欧氏距离。

由欧氏距离的计算可知, 距离是把每个单位看成是 p 维 (p 是变量的个数) 空间的一个点, 在 p 维坐标系中计算的点与点之间的某种距离。不同距离之间的差别在于对距离的定义不同, 这里把各种距离定义公式统一列在表 15-68 中, 而不再逐一介绍。有了距离, 则可以根据点与点之间的距离进行分类, 即将距离较近的点归为一类, 而将距离较远的点归为不同的类。

表 15-68 各种距离计算公式

欧氏距离 (Euclidean distance)	第 i 个样品与第 k 个样品之间的欧氏距离为 $d_{ik} = \sqrt{\sum_{j=1}^p (X_{ij} - X_{kj})^2}$ 即两样品之间的距离是每个变量值之差的平方和之平方根。
欧氏距离平方 (Squared Euclidean distance)	是欧氏距离的平方, 即样品之间的距离是每个变量值之差的平方和。

切比雪夫距离 (Chebychev)	$d_{ik} = \max \{ X_{ij} - X_{kj} \}, 1 \leq j \leq p$, 即任意一个变量值之差的最大绝对值。
明考斯基距离 (Minkowski)	$d_{ik} = [\sum_{j=1}^p X_{ij} - X_{kj} ^q]^{1/q}$, 是欧氏距离的扩展, 每个变量值之差的 q 次方值的绝对值之和的 q 次方根。
块距离(绝对值距离) (Block)	$d_{ik} = \sum_{j=1}^p X_{ij} - X_{kj} $ 即每个变量值之差的绝对值总和。
自定义距离 (Customized)	$d_{ik}(q_1, q_2) = [\sum_{j=1}^p X_{ij} - X_{kj} ^{q_1}]^{1/q_2}$ 在 SPSS 中由用户指定指数 q_1 和开方次数 q_2 (q_1 、 q_2 可取 1 至 4 之间的不同值) 的距离。

注：表中的距离公式是样品间的距离，同样适用于变量间的距离。

(二) 相似系数(similarity)

根据变换后的数据矩阵计算任意两个样品（即两行）或任意两个变量（即两列）之间的相似程度，除距离外，还有相似系数。对于变量之间的相似程度以相似系数来测定尤为广泛。因此，关于行、列的相似系数将一并予以介绍。无论是行、列，相似系数的计算一般有两种方法：一种是夹角余弦；另一种是相关系数。现分述于下：

1、夹角余弦。在 p 维空间中，如果以 $\cos \theta_{ik}$ 表示第 i 行和第 k 行数据值的夹角余弦，则有：

$$\cos \theta_{ik} = \frac{\sum_{j=1}^p X_{ij} \cdot X_{kj}}{\sum_{j=1}^p X_{ij}^2 \cdot X_{kj}^2} \quad (i, k = 1, 2, \dots, n) \quad (15-123)$$

如果将所有行之间的夹角余弦都算出来，则构成一个 $n \times n$ 的夹角余弦矩阵：

$$\cos \theta = \begin{bmatrix} \cos \theta_{11} & \cos \theta_{12} & \cdots & \cos \theta_{1n} \\ \cos \theta_{21} & \cos \theta_{22} & \cdots & \cos \theta_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \cos \theta_{n1} & \cos \theta_{n2} & \cdots & \cos \theta_{nn} \end{bmatrix} \quad (15-124)$$

从式(15-123)可知，如果 X_i 和 X_k 比较相似，则它们的夹角接近于 0，

从而 $\cos \theta_{ik}$ 接近于 1。

在 n 维空间中, 向量 $X_i = (X_{1i}, X_{2i}, \dots, X_{ni})'$ 与 $X_j = (X_{1j}, X_{2j}, \dots, X_{nj})'$ 的

夹角如果记作 α_{ij} , 则变量第 i 列和第 j 列的数据余弦为:

$$\cos \alpha_{ij} = \frac{X_i' X_j}{\sqrt{X_i' X_i} \sqrt{X_j' X_j}} = \frac{\sum_{k=1}^n X_{ki} X_{kj}}{\sqrt{\sum_{k=1}^n X_{ki}^2} \sqrt{\sum_{k=1}^n X_{kj}^2}} \quad (i, j = 1, 2, \dots, p) \quad (15-125)$$

如果将所有列之间的夹角余弦都算出来, 则构成一个 $p \times p$ 的夹角余弦矩阵:

$$\mathbf{cos\alpha} = \begin{bmatrix} \cos \alpha_{11} & \cos \alpha_{12} & \cdots & \cos \alpha_{1p} \\ \cos \alpha_{21} & \cos \alpha_{22} & \cdots & \cos \alpha_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \cos \alpha_{p1} & \cos \alpha_{p2} & \cdots & \cos \alpha_{pp} \end{bmatrix} \quad (15-126)$$

如果 X_i 与 $X_j (i, j = 1, 2, \dots, p)$ 比较相似, 则它们的夹角接近于 0, 从而

$\cos \alpha_{ij}$ 接近于 1。

2、相关系数。在 p 维空间中, 如果以 r_{ik} 表示第 i 行和第 k 行数据的相关系数, 则有:

$$r_{ik} = \frac{\sum_{j=1}^p (X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k)}{\sqrt{\sum_{j=1}^p (X_{ij} - \bar{X}_i)^2 \cdot \sum_{j=1}^p (X_{kj} - \bar{X}_k)^2}} \quad (i, k = 1, 2, \dots, n) \quad (15-127)$$

如果将所有行之间的相关系数都算出来, 就构成一个 $n \times n$ 的相关系数矩阵:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \quad (15-128)$$

由式(15-127)可知, 如果样品之间 (即行之间) 越相近, 它们的相关系

数就越接近 1 或-1；而彼此无关的样品，它们的相关系数就越接近于 0。

在 n 维空间中，如果以 r_{ij} 表示第 i 列和第 j 列数据值的相关系数，则有：

$$r_{ij} = \frac{\sum_{k=1}^n (X'_{ki} - \bar{X}'_i)(X'_{kj} - \bar{X}'_j)}{\sqrt{\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \cdot \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2}} \quad (i, j = 1, 2, \dots, p) \quad (15-129)$$

如果将所有列之间的相关系数都算出来，就构成一个 $p \times p$ 的相关系数矩阵：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad (15-130)$$

与行与行之间的相关系数一样，如果变量与变量之间（即列与列之间）越相近，它们的相关系数就越近于 1 或 -1；而彼此无关的变量，它们的相关系数就越接近于 0。而且可以证明，距离与相似系数有这样的关系： $d_{ij}^2 + r_{ij}^2 = 1$ 。

由相似系数的计算可知，越相近的样品或变量，它们的相似系数越接近于 1 或-1；而彼此关系越疏远的样品或变量，它们的相似系数则越接近于 0。这样，就可以根据样品或变量的相似系数大小，把比较相似的样品或变量归为一类，把不相似的样品或变量归为不同的类。

二、计数变量(Count)（离散变量）的聚类统计量

对于计数变量或离散变量，可用于度量样品（或变量）之间的相似性或不相似性程度的统计量主要有卡方测度（Chi-square measure）和 Phi 方测度（Phi-square measure），如所表 15-69 示。

表 15-69 计数变量的聚类统计量

Chi-square measure	用卡方值测量不相似性。该测度的大小取决于被进行近似计算的两个观测量或变量的总频数期望值。测试产生的值是卡方值的平方根。	
--------------------	---	--

Phi-square measure	该测度试图考虑减少样本量对测度值的实际预测频率减少的影响。该测度把卡方除以合并的频率平方根，使不相似性的卡方测度规范化。其值是 F 平方统计量的平方根。	
--------------------	--	--

三、二值(Binary)变量的聚类统计量

二值变量的聚类统计量有 Euclidean distance (欧氏距离)、squared Euclidean distance (欧氏距离平方)、size difference、pattern difference、variance、dispersion、shape、simple matching、phi 4-point correlation、lambda、Anderberg D、dice、Hamann、Jaccard、Kulczynski 1、Kulczynski 2、Lance and Williams、Ochiai、Rogers and Tanimoto、Russel and Rao、Sokal and Sneath 1、Sokal and Sneath 2、Sokal and Sneath 3、Sokal and Sneath 4、Sokal and Sneath 5、Yule's Y、Yule's Q.等等。

第四节 聚类方法

确定了样品或变量间的相似性或不相似性统计量后，就要对样品或变量进行分类。样品聚类和变量聚类的方法很多，但考虑到使用的广泛性，本节仅阐述样品聚类中的系统聚类法。

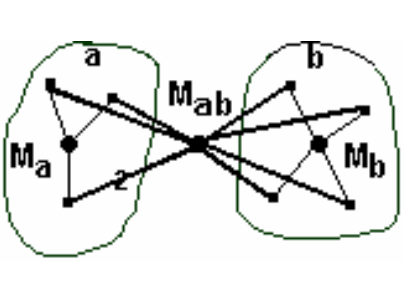
系统聚类法是目前应用最多的一种聚类方法。正如前面所说，该方法的基本思想是，首先将每个样品各自看成一类，选择距离最小的两类合并成一新类（如果样品间关系采用相似系数，则应选择相似系数绝对值最大的两类合并成一新类），然后计算该新类与其他类之间的距离，再将距离最小的两类进行合并，如此继续，这样每次合并后都减少一类，直到所有的样品都聚为一类为止。

类与类之间的距离有多种计算方法，如既可以计算两类单位之间的最近距离以表示两类之间的距离大小，也可以计算两类单位之间的最远距离以表示两类之间的距离大小等。正因为类与类之间距离的不同计算，就产生了系统聚类的不同方法。系统聚类法常用的方法如表 15-70所示。

表 15-70 常用的聚类方法

聚类方法	说明	备注
1、组间连接法 (between-groups linkage)	合并两类的结果使所有的两两样品之间的平均距离最小。样品对的两个单位分别属于不同的类。(SPSS 默认方法)	
2、组内连接法 (within-groups linkage)	合并后的类中的所有样品之间的平均距离最小。两类间的距离即是合并的类中所有可能的样品对之间的距离平方。	
3、最短距离法 (nearest neighbor)	首先合并最近的或最相似的两类，用两类间最近点间的距离代表两类间的距离。	
4、最长距离法 (furthest neighbor)	用两类之间最远点的距离代表两类之间的距离。	
5、重心法 (centroid clustering)	先求出各类的重心点，以重心点的距离作为类间相似性的测度。要求样品间距离为欧氏距离平方。	
6、中位数法 (median clustering)	用两类的中位数之间的距离作为测度。要求样品间距离为欧氏距离平方。	

详见有关参考文献。

7、Ward 法 (离差平方和法) (Ward's method)	由 Ward 提出来的,其思想来源于方差分析。如果类分得正确,同类样品的离差平方和应当较小,类与类之间的离差平方和应当较大。要求样品间距离为欧氏距离。	
--------------------------------------	---	--

在实际中应用较多的聚类方法有组间连接法(SPSS 默认聚类方法)和 Ward 离差平方和法。上述七种系统聚类方法,聚类原则与步骤完全一致,不同的只是类与类之间的距离有不同的计算,从而得到不同的递推公式。为了统一递推公式,1969 年 Wishart 首先提出了统一公式的模型,这为编制统一的计算程序提供了极大方便。Wishart 的统一递推公式为:

$$D_{rs}^2 = d_i D_{si}^2 + d_j D_{sj}^2 + \beta D_{ij}^2 + \gamma |D_{si}^2 - D_{sj}^2|$$

其中, D_{ij}^2 表示类 G_i 和类 G_j 之间的距离平方,类 G_i 和类 G_j 合并成新类 G_r ,

D_{rs}^2 表示类 G_r 与任一类 G_s 的距离递推式, D_{si}^2 和 D_{sj}^2 分别表示类 G_i 和类 G_j 与

类 G_s 的距离平方, d_i 、 d_j 、 β 和 γ 对不同的聚类方法有不同的取值,其具体值如所示。

表 15-71 聚类方法参数表

聚类方法	d_i	d_j	β	γ
组间连接法	n_i / n_r	n_j / n_r	0	0
最短距离法	1/2	1/2	0	-1/2
最长距离法	1/2	1/2	0	1/2
重心法	n_i / n_r	n_j / n_r	$-n_i n_j / n_r^2$	0
中位数法	1/2	1/2	-1/4	0
离差平方和法	$\frac{n_s + n_i}{n_s + n_r}$	$\frac{n_s + n_j}{n_s + n_r}$	$-\frac{n_s}{n_s + n_r}$	0

综上所述,聚类分析的最终结果,不是一个或几个具体的新数据(或指标),而只是一个或一些定性的类别,如哪几个单位同属一类等,因此,

对于多因素综合的复杂现象进行定性判别有很大的作用。

[例 15-29]我国各地区三次产业产值如所示,试根据三次产业产值进行聚类分析。

表 15-72 我国各地区三次产业产值(单位:亿元)

地区 Region	第一产业 产值 X1	第二产业 产值 X2	第三产 业产值 X3	地区 Region	第一产业 产值 X1	第二产业 产值 X2	第三产业 产值 X3
北 京	86.56	786.85	1137.9	湖 北	748.22	1752.91	1203.08
天 津	74.03	660	602.35	湖 南	828.31	1294.17	1088.92
河 北	790.6	2084.33	1381.08	广 东	1004.92	3991.97	2922.23
山 西	207.26	856.13	537.72	广 西	574.25	678.19	650.6
内蒙古	341.62	479.53	371.14	海 南	164	90.63	184.29
辽 宁	531.46	1855.22	1495.05	重 庆	298.67	585.38	545.21
吉 林	429.5	597.29	530.99	四 川	941.24	1527.07	1111.95
黑龙江	463.05	1506.76	863.03	贵 州	264.89	326.03	250.96
上 海	78.5	1847.2	1762.5	云 南	408.43	828.37	557.1
江 苏	1016.27	3640.1	2543.58	西 藏	31.31	20.24	39.63
浙 江	631.31	2709.08	1647.11	陕 西	283.49	567.66	530.38
安 徽	739.7	1253.53	812.22	甘 肃	202.21	382	285.54
福 建	610.04	1444.73	1275.41	青 海	41.63	88.42	90.11
江 西	450.44	740.33	661.21	宁 夏	48.69	94.01	84.76
山 东	1215.81	3457.03	2489.36	新 疆	291.05	430.73	394.89
河 南	1071.39	2012.74	1272.47				

解: SPSS 操作步骤如下:

(1) 在 SPSS 中录入数据。

(2) 选择[Statistics]=>[Classify]=>[Hierarchical Cluster], 打开分层聚类对话框。

(3) 把变量 X1、X2、X3 选入[Variable]框, 把变量 region 选入[Label Cases], 系统默认为样品聚类。

(4) 单击[Statistics]按钮选择要输出的统计量, 统计量对话框中各选项如下:

Agglomeration schedule 凝聚状态表	显示聚类过程的每一步合并的类或样品、被合并的类或样品之间的距离以及样品或变量加入到一类的
---------------------------------	--

	类水平。
Proximity matrix 相似矩阵	给出各类之间的距离或相似测度值。
Cluster Membership 类成员	显示每个样品被分配到的类或显示若干步凝聚过程。具体内容有三个选项： None:不显示类成员表，是默认值； Single solution:要求列出聚为一定类数的各样品所属的类。 Range of solutions:要求列出某个范围中每一步各样品所属的类。

本例均使用默认设置。

(5) 单击[Plots]按钮选择统计图表，统计图表对话框各选项如下：

Dendrogram 树形图	树形图表明每一步中被合并的类及其系数值，把各类之间的距离转换成 1 至 25 之间的数值。
Icicle 冰柱图	冰柱图把聚类信息综合到一张图上。 纵向冰柱图(Vertical)：参与聚类的个体各占一列，标以样品(或变量)号或标签；聚类过程中的每一步占一行，标以步的顺序号。 横向冰柱图(Honrizontal)：参与聚类的样品(或变量)各占一行，聚类的每一步各占一列。如果不加限定的选择项，则显示聚类的全过程。

本例仅选择树形图，其它选项不变。

(6) 单击[Method]按钮选择聚类方法，其对话框中各选项如下：

Cluster 聚类方法选择	见表 15-70列出的各种聚类方法
Measure 对距离和相似系数的 不同测量方法	见第三节的各种距离和相似系数
Transform Values 转换数值的方法，标 准化方法	如果参与聚类的变量的量纲不同会导致错误的聚类结果。在聚类之间必须先标准化数据，以消除量纲的影响。 如果参与聚的变量量纲相同，可以使用系统默认值 None，即不进行标准化处理。 标准化处理方法有：(参见第二节) Z scores：把数值减去均值后再除以其标准差；

	Range -1 to 1 : 标准化到-1 到+1 之间 ; Range 0 to 1 : 标准化到 0 到 1 之间 ; Maximum magnitude of 1 : 标准化到最大值为 1 ; Mean of 1 : 标准化在一个均值范围内 ; Standard deviation of 1 : 标准化到单位标准差。
Transform Measures 测度的转换方法	对距离测度数值进行转换, 有三种方法 : Absolute Values : 把距离值标准化 ; Change sign : 把相似性值变为不相似值, 或相反 ; Rescale to 0~1 range : 首先去掉最小值然后除以范围把距离标准化。对于已经按某种换算方法标准化了的测度一般不再使用此方法进行转换。

本例使用默认选项。

(7) 单击[SAVE]按钮, 显示保存新变量对话框, 选项如下:

None	不建立新变量
Single solution	单一结果
Range of solutions	指定范围内的结果

对于本例, 使用默认值。

(8) 设置完各种选项后, 单击[OK]按钮, 输出聚类结果。

Case Processing Summary

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
31	100.0	0	.0	31	100.0

a Squared Euclidean Distance used (使用测量方法: 欧氏距离平方)

b Average Linkage (Between Groups) (使用聚类方法: 组间连接)

Average Linkage (Between Groups) (组间连接)

Agglomeration Schedule (凝聚过程表)(略)

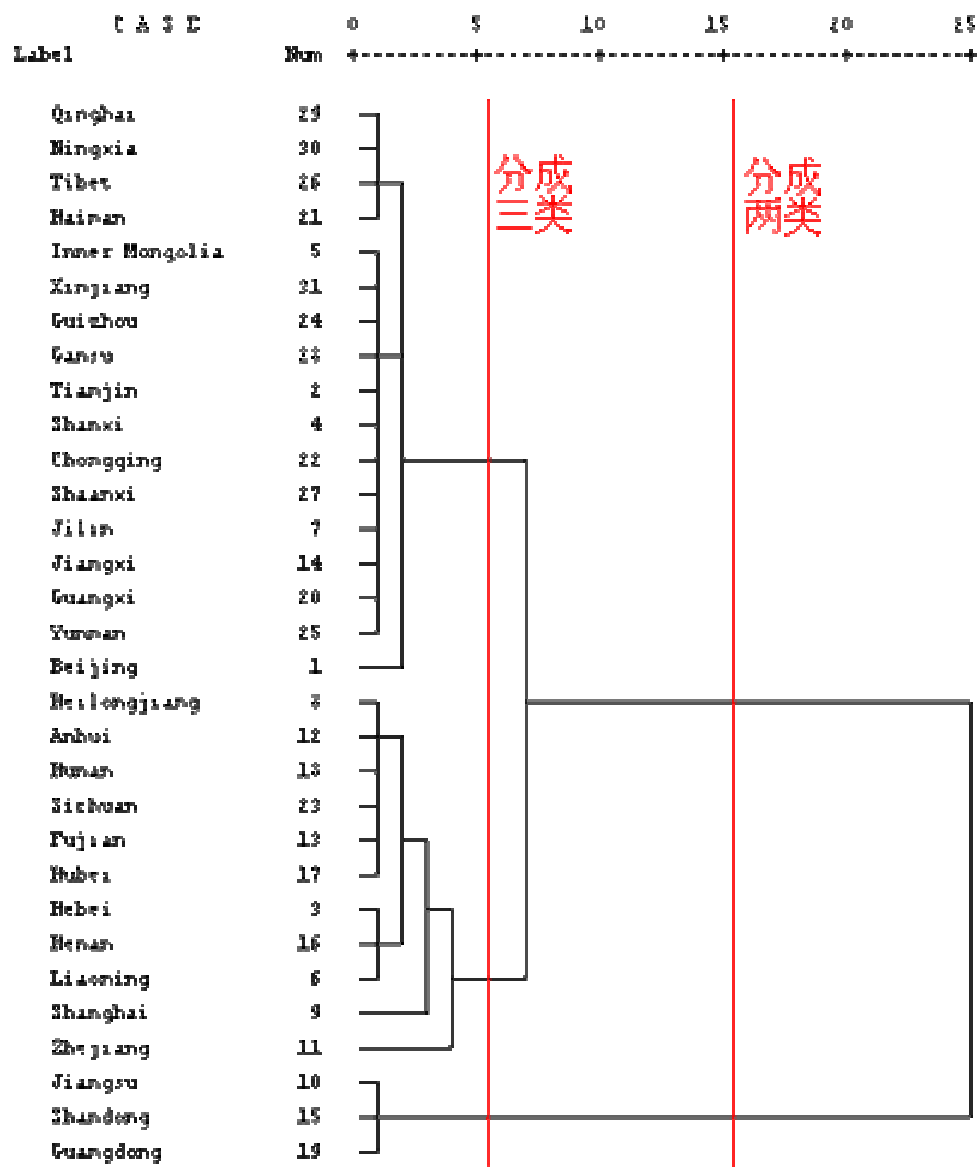
Vertical Icicle (纵向冰柱图)(略)

Dendrogram (树形图)

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine (转换成 1—25 的类间距离)



从冰柱图和树形图看，把各省市分为两类或三类。

第五节 案例：汽车市场需求情况定量研究

以往在我国各地区汽车需求量的研究中，主要是根据国家政策、国民经济发展情况、各地区公路状况等，结合不同时期汽车保有量，对汽车市场进行定性分析和决策。这样往往带有主观因素，下面用聚类分析法对表 15-73 中各地区与汽车市场相关的原始数据进行分类，对我国汽车市场发展

参见张秀芳：《聚类分析法在汽车市场需求中的应用》，哈尔滨理工大学学报，1998.2。

进行了分类预测，用定量分析的方法对汽车市场进行研究，从而对汽车市场需求情况有更深入的分析与评价。

首先对表 15-73 中的原始数据进行标准化变换处理，经过运算使数据标准化得到表 15-74。使它的每列数据的平均值为 0，方差为 1。这样表 15-73 中 5 列具有不同量纲、不同数量级的数据就能放在一起比较；其次用表 15-74 中经过标准化处理后的 30 个不同地区数据求出欧氏距离；最后应用 Wald 离差平方和法，假定将 30 个地区分成 30 类，求得每个地区的离差平方和及总的类内离差平方和，按照使总的类内离差平方和增加最小的原则，使得类的分法逐渐减小。最终的计算结果决定了聚类谱系图中纵轴 30 个不同地区的前后排序，而横轴表示的是样本点之间的距离。用 SPSS 完成以上运算步骤，给出聚类分析的结果的聚类谱系图，如图 15-114 所示。由此图可以看出，可将 30 个样品分成三类，第一类包含 (1, 2, 9, 21, 28, 29, 25)；第二类包含 (4, 5, 26, 14, 20, 8, 13, 7, 30, 23, 27)；第三类包含 (3, 6, 10, 11, 16, 15, 19, 22, 12, 17, 18, 24)。在这个分类中，第一类所反映的为我国经济发展较发达地区与相对欠发达地区。1, 2, 9 所代表为北京、天津、上海三个直辖市，在全国具有举足轻重的地位。它们的汽车市场发展仍将处在全国领先水平。而 21, 28, 29, 25 分别为海南、青海、宁夏、西藏，由于历史、地理、人口、气候及交通等原因，汽车市场的发展将作为今后发展的重要因素，带动该地区经济的腾飞。第二类中 11 个元素，分别代表山西、内蒙古、陕西等。此类样本点从经济发展看处于中等水平，将是今后汽车发展的大市场。第三类为河北、辽宁、江苏等，这些地区处于经济发展前沿，地处沿海、或是交通发达的地区，今后它们仍将对汽车需求保持强劲势头，而且某汽车市场的发展必将带动第一、第二类地区，从整体上促进我国汽车工业及相关产业的发展。

表 15-73 各地区相关数据表

序号	地区	国内生产总值 (亿元)	地区人口 总数 (万人)	地区公路 长度 (km)	全社会货 运量 (万吨)	汽车总保 有量 (万辆)
1	北京	1394.89	1251	11811	29087	58.94
2	天津	920.11	942	4243	19260	26.80
3	河北	2849.52	6437	51630	59860	72.64
4	山西	1092.48	3077	33644	39774	33.29
5	内蒙古	832.88	2284	44753	24384	22.12

6	辽宁	2793.37	4092	43434	69976	63.90
7	吉林	1129.20	2592	31321	20478	23.61
8	黑龙江	2014.53	3701	48819	19281	36.36
9	上海	2462.57	1415	3787	24645	30.71
10	江苏	5155.25	7066	25970	49578	51.19
11	浙江	3524.79	4319	34121	45052	35.92
12	安徽	2003.58	6013	35178	30236	24.46
13	福建	2160.52	3237	46574	23720	20.08
14	江西	1205.11	4063	34915	18078	16.90
15	山东	5002.34	8705	54243	63525	76.20
16	河南	3002.74	9100	49707	42692	46.93
17	湖北	2391.42	5772	48728	28649	34.89
18	湖南	2195.70	6392	59125	41272	35.24
19	广东	5381.72	6868	84567	71158	114.73
20	广西	1606.15	4543	40904	20686	24.90
21	海南	364.17	724	14808	6809	9.52
22	四川	3534.00	11325	100724	80001	53.35
23	贵州	630.07	3508	32487	8957	15.89
24	云南	1206.68	3990	68236	36042	33.39
25	西藏	55.98	240	22391	360	2.94
26	陕西	1000.03	3514	39620	25560	24.52
27	甘肃	553.35	2438	35194	17719	15.14
28	青海	165.31	481	17223	2887	5.76
29	宁夏	169.75	513	8554	3465	5.00
30	新疆	834.57	1661	30298	17000	24.70

表 15-74 数据标准化

	Z1	Z2	Z3	Z4	Z5
1	-0.35679	-0.98672	-1.22446	-0.1061	0.99732
2	-0.67872	-1.09728	-1.57081	-0.56892	-0.32325
3	0.62953	0.86881	0.59782	1.34323	1.56023
4	-0.56185	-0.33338	-0.2253	0.39723	-0.05659
5	-0.73787	-0.61711	0.2831	-0.32759	-0.51555

6	0.59145	0.02978	0.22274	1.81966	1.20112
7	-0.53695	-0.50691	-0.33161	-0.51156	-0.45433
8	0.06336	-0.11012	0.46918	-0.56793	0.06955
9	0.36715	-0.92804	-1.59167	-0.3153	-0.1626
10	2.19294	1.09387	-0.57649	0.85897	0.67889
11	1.0874	0.111	-0.20347	0.64581	0.05147
12	0.05593	0.71711	-0.15509	-0.05198	-0.4194
13	0.16234	-0.27614	0.36643	-0.35887	-0.59937
14	-0.48548	0.0194	-0.16713	-0.62459	-0.73003
15	2.08926	1.68029	0.7174	1.51584	1.7065
16	0.73342	1.82162	0.50981	0.53466	0.50385
17	0.31891	0.63088	0.46501	-0.12672	0.00915
18	0.1862	0.85271	0.94082	0.46778	0.02353
19	2.3465	1.02302	2.10515	1.87533	3.28962
20	-0.21355	0.19115	0.10695	-0.50176	-0.40132
21	-1.05568	-1.17528	-1.08731	-1.15533	-1.03326
22	1.09364	2.61772	2.84456	2.29181	0.76764
23	-0.87538	-0.17917	-0.27824	-1.05416	-0.77153
24	-0.48441	-0.00671	1.35778	0.22147	-0.05248
25	-1.26465	-1.34845	-0.74028	-1.45906	-1.30362
26	-0.62453	-0.17703	0.04819	-0.27221	-0.41693
27	-0.92741	-0.56201	-0.15436	-0.6415	-0.80234
28	-1.19052	-1.26222	-0.97679	-1.34004	-1.18775
29	-1.18751	-1.25077	-1.37352	-1.31282	-1.21898
30	-0.73672	-0.84002	-0.37842	-0.67536	-0.40954

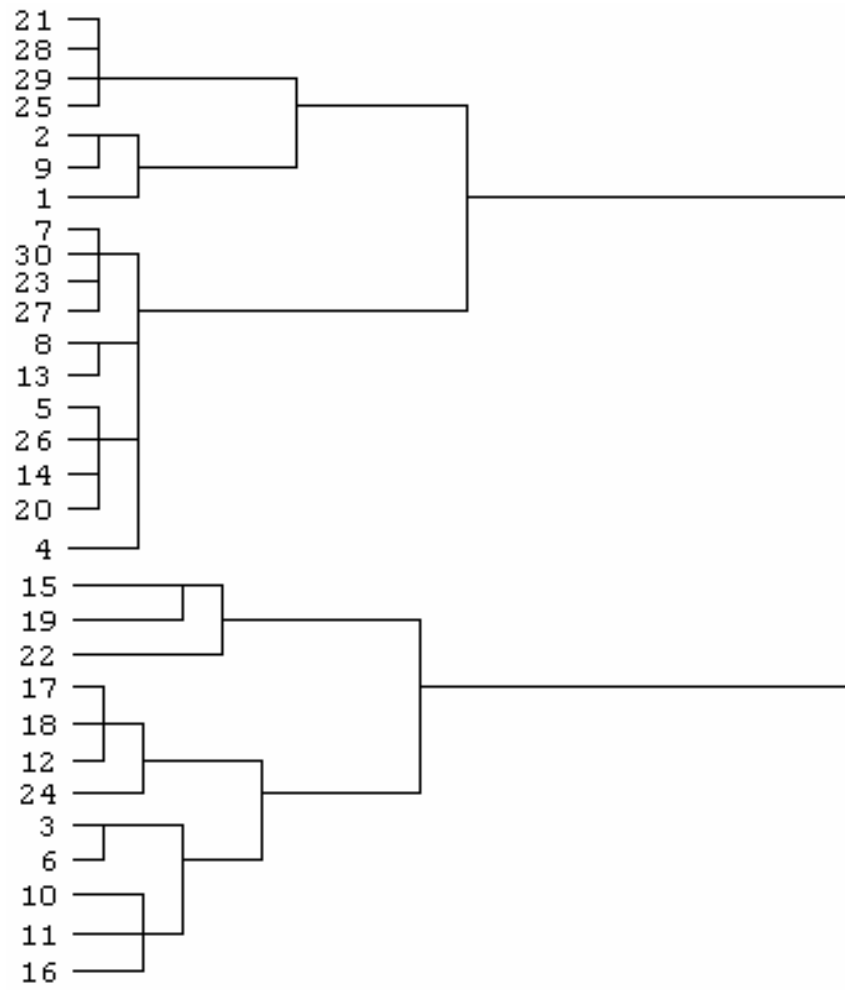


图 15-114 聚类谱系图

第十三章 主成分分析

第一节 主成分分析的基本思想

主成分分析(Principal Components Analysis, PCA)也称为主分量分析,是一种通过降维来简化数据结构的方法:如何把多个变量(指标)化为少数几个综合变量(综合指标),而这几个综合变量可以反映原来多个变量的大部分信息。为了使这些综合变量所含的信息互不重叠,应要求它们之间互不相关。

例如在评价企业的经营业绩时,要考虑许多指标,如利润、产值、产品数量、产品质量、固定资产、流动资产等等。若要全部列出,也许可以有几十个变量。因此用少量的几个综合变量代替原来的许多变量是有实际意义的。由这几个综合变量出发还有可能得到一个总的指标,按此总指标来排序、分类,问题就可能简单多了。

为了方便,下面通过一个例子在二维空间中讨论主成分的几何意义。

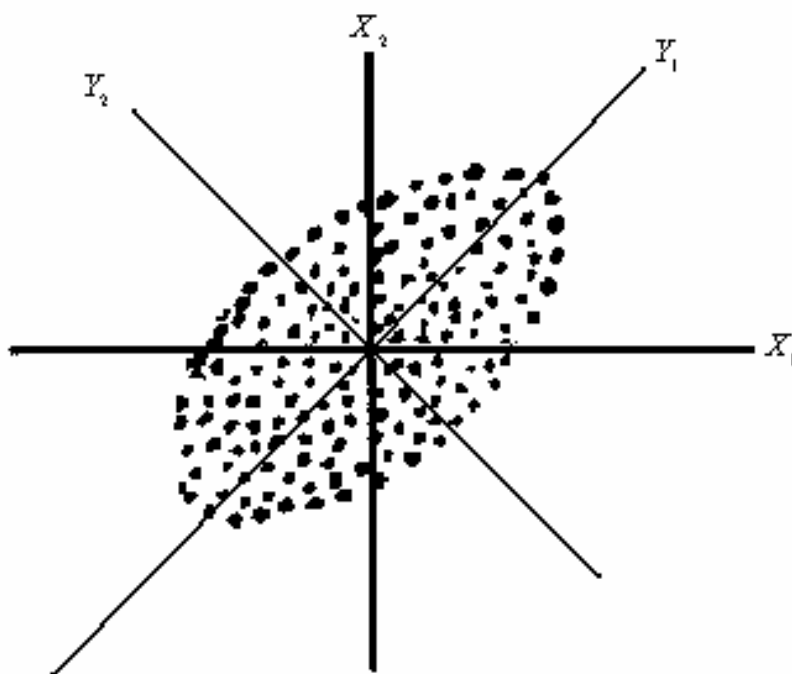


图 17-115 主成分的几何意义

假定某年级学生的语文成绩(X_1)和数学成绩(X_2)的相关系数 $\rho = 0.6$ 。

设 X_1 和 X_2 分别为标准化后的分数，其散点图如所示。那么随机向量 $\mathbf{X}' = (X_1, X_2)$ 的方差—协方差矩阵 为：

$$\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

由式有

$$(\Sigma - \lambda \mathbf{I})\mathbf{e} = 0$$

可求出 Σ 的特征值分别为：

$$\lambda_1 = 1.6 \quad \lambda_2 = 0.4$$

及其对应的特征向量分别为

$$\mathbf{e}_1' = (e_{11}, e_{21}) = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right)$$

$$\mathbf{e}_2' = (e_{21}, e_{22}) = \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right)$$

显然，这两个特征向量是相互正交的单位向量。而且它们与原来的坐标轴 X_1 和 X_2 的夹角都分别等于 45° 。如果将坐标轴 X_1 和 X_2 旋转 45° ，那么点在新坐标系中的坐标 (Y_1, Y_2) 与原坐标 (X_1, X_2) 有如下的关系：

$$Y_1 = \frac{\sqrt{2}}{2} X_1 + \frac{\sqrt{2}}{2} X_2 = \mathbf{e}_1' \mathbf{X}$$

$$Y_2 = \frac{\sqrt{2}}{2} X_1 - \frac{\sqrt{2}}{2} X_2 = \mathbf{e}_2' \mathbf{X}$$

在新坐标系中（如图 17-115 所示），可以发现：虽然散点图的形状没有改变，但新的随机变量 Y_1 和 Y_2 已经不再相关。而且大部分点沿 Y_1 轴散开，在 Y_1 轴方向的变异较大（即 Y_1 的方差较大），相对来说，在 Y_2 轴方向的变异较小（即 Y_2 的方差较小）。事实上，随机变量 Y_1 和 Y_2 的方差分别为：

$$\text{Var}(Y_1) = E(Y_1^2) = \mathbf{e}_1' \Sigma \mathbf{e}_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = 1.6 = \lambda_1$$

在变量标准化的情况下的方差—协方差矩阵与其相关矩阵相等。

$$Var(Y_2) = E(Y_2^2) = \mathbf{e}_2' \boldsymbol{\Sigma} \mathbf{e}_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} = 0.4 = \lambda_2$$

可以看出，最大变动方向是由特征向量所决定的，而特征值则刻画了对应的方差。

在上面的例子中 Y_1 和 Y_2 就是原变量 X_1 和 X_2 的第一主成分和第二主成分。实际上第一主成分 Y_1 就基本上反映了 X_1 和 X_2 的主要信息，因为图 17-115 中的各点在新坐标系中的 Y_1 坐标基本上就代表了这些点的分布情况，因此可以选 Y_1 为一个新的综合变量。当然如果再选 Y_2 也作为综合变量，那么 Y_1 和 Y_2 则反映了 X_1 和 X_2 的全部信息。

第二节 总体主成分

一、主成分的定义

设 $X' = (X_1, X_2, \dots, X_p)$ 是 p 维随机向量，它的主成分为：

$$\begin{aligned} Y_1 &= \mathbf{e}_1' X = e_{11}X_1 + e_{21}X_2 + \dots + e_{p1}X_p \\ Y_2 &= \mathbf{e}_2' X = e_{12}X_1 + e_{22}X_2 + \dots + e_{p2}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p' X = e_{1p}X_1 + e_{2p}X_2 + \dots + e_{pp}X_p \end{aligned}$$

其中： $\mathbf{e}_i' \mathbf{e}_i = 1 (i=1, 2, \dots, p)$ ； Y_1 是一切 $\mathbf{Y}_i = \mathbf{e}_i' X$ 中方差最大者， Y_2 是一切

$\mathbf{Y}_i = \mathbf{e}_i' X$ 中方差次大者，……， Y_p 是一切 $\mathbf{Y}_i = \mathbf{e}_i' X$ 中方差最小者； $Y_1, Y_2, \dots, Y_p$

互不相关。因此， p 个变量的 p 个主成分就是这 p 个变量的 p 个线性组合，其中线性组合的系数向量是单位向量。

二、主成分的性质

设 p 维随机向量 $X' = (X_1, X_2, \dots, X_p)$ 的方差—协方差矩阵为 $\boldsymbol{\Sigma}$ ，即

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p) \end{pmatrix} \text{记} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

Σ 的 p 个特征值为 $\lambda_1, \lambda_2, \dots, \lambda_p$ ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$), 对应的 p 个单位特

征向量为 $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$, 那么:

1. X 的第 i 个主成分 Y_i 的系数向量就是第 i 个特征值 λ_i 所对应的正交化特征向量 $\mathbf{e}_i$, 即

$$Y_i = \mathbf{e}_i' \mathbf{X} = e_{1i} X_1 + e_{2i} X_2 + \cdots + e_{pi} X_p$$

2. 第 i 个主成分 Y_i 的方差为第 i 个特征值 λ_i , 每两个不相同主成分间的协方差为 0。也就是说, 令主成分向量 $\mathbf{Y}' = (Y_1, Y_2, \dots, Y_p)$, 则 $\mathbf{Y}$ 的方差—协方差矩阵是一对角矩阵 Λ , 其对角元素分别是 $\lambda_1, \lambda_2, \dots, \lambda_p$, 即 $\mathbf{Y}$ 的方差—协方差矩阵为

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

3. Σ 和 Λ 的对角元素之和相等, 即两个方差—协方差矩阵的迹相等, 即

$$\text{tr}(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \text{tr}(\Lambda)$$

由此可以看出, 主成分分析把 p 个原始变量 $X_1, X_2, \dots, X_p$ 的总方差 $\text{tr}(\Sigma)$ 分

解为 p 个不相关的变量 $Y_1, Y_2, \dots, Y_p$ 的方差之和 $\text{tr}(\Lambda) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$ 。主成

分分析的目的就是为了减少变量的个数, 一般是不会使用所有 p 个主成分的, 忽略一些带有较小方差的主成分将不会给总方差带来大的影响。由此

可进一步得到

$$\text{第 } i \text{ 个主成分的方差占总方差的比例} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad (i = 1, 2, \cdots, p)$$

称此方差比例为主成分 Y_i 的贡献率。第一主成分的贡献率最大，这表明 Y_1 综合原始变量 $X_1, X_2, \cdots, X_p$ 的能力最强，而 $Y_2, \cdots, Y_p$ 的综合能力依次减弱。

若只取前 m ($m < p$) 个主成分，则称

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_m}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

为主成分 $Y_1, Y_2, \cdots, Y_m$ 的累计贡献率，累计贡献率表明 $Y_1, Y_2, \cdots, Y_m$ 综合 $X_1, X_2, \cdots, X_p$ 的能力。通常取 m ，使得累计贡献率达到一个较高的百分数（如 85% 以上）。

4. 主成分 Y_i 与变量 X_j 的相关系数

$$\rho(Y_i, X_j) = \frac{\text{Cov}(X_j, Y_i)}{\sqrt{\text{Var}(X_j)}\sqrt{\text{Var}(Y_i)}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}} e_{ji} \quad (i, j = 1, 2, \cdots, p)$$

称为因子负荷量（或因子载荷量）。（注：一些书中也称主成分为因子。）因此，第 i 个特征向量 e_i 的第 j 个分量 e_{ji} 描述了第 j 个变量对第 i 个主成分的重要性，它与 Y_i 和 X_j 之间的相关系数成比例。

5. 前面的累计贡献率度量了主成分 $Y_1, Y_2, \cdots, Y_m$ 从原始变量 $X_1, X_2, \cdots, X_p$ 中提取了信息的多少，而 $Y_1, Y_2, \cdots, Y_m$ 包含有 X_j ($j = 1, 2, \cdots, p$) 的多少信息用 X_j 与 $Y_1, Y_2, \cdots, Y_m$ 的复相关系数的平方来度量，即

$$\rho_{j,(1 \cdots m)}^2 = \sum_{i=1}^m \lambda_i e_{ji}^2 / \sigma_{jj}$$

称为 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 对原始变量 X_j 的贡献率。 p 个主成分

$Y_1, Y_2, \cdots, Y_p$ 对 X_j 的贡献率为 1，即

$$\rho_{j,(1\cdots p)}^2 = \sum_{i=1}^p \lambda_i e_{ji}^2 / \sigma_{jj} = 1$$

三、从相关矩阵出发求主成分

前面讨论的主成分是从方差—协方差矩阵 Σ 出发求得的，其结果受原始 p 个变量单位的影响。不同的变量往往有不同的单位，对同一变量使用不同的单位会产生不同的主成分，主成分会过于照顾方差大的变量 X_j ，而对方差小的变量却照顾得不够。为使主成分分析能够均等地对待每一个原始变量，消除由于单位的不同而可能带来的一些不合理的影响，常常将各原始变量作标准化处理。即令

$$X_j^* = \frac{X_j - E(X_j)}{\sqrt{\text{Var}(X_j)}} \quad (j=1,2,\cdots,p)$$

这时 $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)'$ 方差—协方差矩阵 Σ 就是其相关矩阵 ρ 。从 ρ 出

发求主成分的方法与从 Σ 出发是类似的。如果从相关矩阵 ρ 来求主成分，那么前面几个性质可以写成

1. $\mathbf{X}^*$ 的第 i 个主成分

$$Y_i = \mathbf{e}_i' \mathbf{X}^*$$

2. $\mathbf{Y}$ 的方差—协方差矩阵等于 Λ

3. $\text{tr}(\rho) = \text{tr}(\Lambda) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$

4. 主成分 Y_i 与变量 X_j^* 之间的相关系数为

$$\rho(Y_i, X_j^*) = \sqrt{\lambda_i} e_{ji} \quad (i, j=1,2,\cdots,p)$$

5. 主成分 $Y_1, Y_2, \cdots, Y_p$ 对变量 X_j^* 的贡献率为

$$\rho_{j,(1\cdots m)}^2 = \sum_{i=1}^m \lambda_i e_{ji}^2$$

由 Σ 和 ρ 出发求得的主成分会有较大的差别，因此变量的标准化并不是无关紧要的，在实际应用中要注意这个问题。

第三节 样本主成分

在实际问题中，总体的方差—协方差矩阵或相关矩阵是不知道的。我们只是从数据出发，计算原始变量的样本方差—协方差矩阵或相关矩阵，然后再进行主成分分析的。

设样本数据矩阵为

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)' = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

则样本方差—协方差矩阵 $\mathbf{S}$ 和样本相关矩阵 $\mathbf{R}$ 分别为

$$\mathbf{S} = (S_{ij}) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

$$\mathbf{R} = (R_{ij}) = \frac{S_{ij}}{\sqrt{S_{ii}} \sqrt{S_{jj}}}$$

其中 $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ 为样本均值。可以分别用 $\mathbf{S}$ 、 $\mathbf{R}$ 代替 Σ 、 ρ ，然后从

$\mathbf{S}$ 或 $\mathbf{R}$ 出发按上一节的方法求主成分。若从 $\mathbf{S}$ 出发，则主成分分析将使得方差大的那些原始变量与具有大特征值的主成分有较密切的联系，而方差小的另一些变量同具有小特征值的主成分有较强的联系。因此，在实际应用中，一般从 $\mathbf{R}$ 出发来求得主成分，除非原始变量所测量的单位是可比较的，或者这些变量已用某些方法标准化了。

[例 17-30] 美国 38 个城市在人口、经济和财政方面的 11 个百分比变化变量上的概括性信息如表 17-75 所示。为了定义能说明数据中大部分差异的少量的几个主成分指标，同时也希望能了解这些城市的结构形态，首先从方差—协方差矩阵出发进行了主成分分析。

表 17-75 关于 11 个变量的概括性信息

变量	均值	标准差	说明
X_1	-0.03	0.234	1970~1977 的人口百分比变化
X_2	0.52	0.085	1969~1975 间（调整后的）收入的百分比变化
X_3	0.016	0.209	1970~1977 间总就业的百分比变化

参见 Dillon & Goldstein, Multivariate Analysis, John Wiley & Sons, 1984 和柯惠新等：《调查研究中的统计分析法》，1992。

X ₄	0.92	0.308	1970~1976 间利润的百分比变化
X ₅	0.52	0.676	1972~1977 间政府税收的百分比变化
X ₆	-1.29	3.301	1972~1977 同税收——支出不平衡的百分比变化
X ₇	0.40	2.062	1971~1976 间有效的财产税率的百分比变化
X ₈	0.58	0.631	1971~1976 间实际财产税底的近似市场值的百分比变化
X ₉	-0.01	0.455	1972~1977 间长期债务相对于最大容限的百分比变化
X ₁₀	1.27	0.496	1970~1977 间在公共设施上人均花费百分比变化
X ₁₁	1.27	0.583	1970~1977 间往市人均消费的百分比变化

表 17-76 (A) 中给出了特征值和可解释方差的累计百分比。可以看到前两个因子就说明了大约 90% 的方差，而前三个因子则超过了 93%，因此似乎可以认为，用两个或三个因子就足以获取原始数据的方差结构了。再看对主成分的解释。表 17-76 (B) 中给出了主成分与各变量的相关系数（即因子负荷）。从因子负荷中我们看到第一个主成分（占总方差的 64%）很明显是由税收——支出不平衡（X₆）所支配的；而第二个主成分（约占总方差的 25%）主要是由有效财产税率（X₇）决定的。我们可能已经注意到支配前两个主成分的两个变量是具有最大方差的两个变量，它们的方差分别是

$$3.301^2 = 10.9 \text{ 和 } 2.062^2 = 4.25$$

在利用方差—协方差矩阵进行主成分分析时，一种心照不宣的假定是：变量的方差不应相差太大。否则前几个主成分将朝着那几个有较大方差的变量的方向被抽取。因此我们就不会奇怪为什么是变量 X₆ 和 X₇ 分别刻画了前两个主成分。对于这种方差相差很大的情况，为了防止主成分趋向方差大的变量，一般应从相关矩阵出发来进行主成分分析。

从相关矩阵出发求得特征值以及由此计算得到的各个主成分的方差贡献率如所示。取对应特征值大于 1 的那 5 个主成分，可解释的方差超过总方差的 80%。

给出了在这 5 个因子上的负荷值。每个变量的最高绝对负荷值下面都划线标明。根据在每个因子上负荷最高的那些变量来说明主成分的意义。

在第一主成分上负荷高的变量分成两组。一组是人口变化 X1、城市居民就业 X3 和有效财产税率变化 X7。这些都属于“增长刺激”变量。另一组是在公共设施上人均花费的变化 X10 和城市人均消费变化 X11。这些属于“下降指标”变量。两组变量的负荷符号相反，因此第一主成分可以解释为是“增长与下降的对比变化”。类似地可以给出对第二主成分的解释。而第三、四、五主成分分别只用一个变量代表，其意义是比较清楚的。

表 17-76 从方差—协方差矩阵出发进行主成分分析的结果

(A)特征值与方差贡献率			(B)主成分(因子)负荷			
主成分 (因子)	可解释 的方差	累计方差 贡献率	变量	因子负荷		
				1	2	3
1	10.936	64.3%	X ₁	-0.0746	0.5667	-0.3723
2	4.323	89.7%	X ₂	0.1164	-0.1136	-0.0601
3	0.623	93.4%	X ₃	0.0043	0.5121	-0.2714
4	0.464	96.1%	X ₄	0.0925	0.4470	0.2375
5	0.376	98.3%	X ₅	0.1311	0.1402	0.0390
6	0.118	99.0%	X ₆	<u>0.9996</u>	0.0089	-0.0091
7	0.078	99.4%	X ₇	-0.0218	<u>0.9988</u>	0.0235
8	0.061	99.8%	X ₈	-0.2472	0.0905	<u>-0.5753</u>
9	0.024	99.9%	X ₉	0.0385	-0.0517	<u>0.6644</u>
10	0.007	100.0%	X ₁₀	0.0206	-0.1461	<u>0.7637</u>
11	0.003	100.0%	X ₁₁	0.0106	-0.1334	<u>0.8195</u>

表 17-77 从相关矩阵出发进行主成分分析的结果

(A) 特征值与方差贡献率			
因子/主成分	特征值	可解释方差的百分比	累计方差贡献率(%)
1	3.130	28.5	28.5
2	1.989	18.1	46.6
3	1.477	13.4	60.0
4	1.154	10.5	70.5
5	1.089	9.9	80.4
6	0.663	6.0	86.4
7	0.545	5.0	91.4
8	0.404	3.6	85.0
9	0.331	3.0	98.0
10	0.145	1.3	99.3

11	0.073	0.7	100.0
合计	11.000	100	

(B) 主成分(因子)负荷

变量	因子负荷				
	1	2	3	4	5
X ₁	0.90936	0.20587	0.15208	-0.07556	-0.13286
X ₂	-0.14097	-0.33839	0.55362	0.20074	0.49161
X ₃	0.81965	0.20944	0.38487	-0.08755	-0.04807
X ₄	0.08799	0.63249	-0.23092	0.51255	0.29409
X ₅	0.38816	0.41821	0.33187	-0.53723	-0.12257
X ₆	-0.09243	0.25691	0.10220	-0.40261	0.78553
X ₇	0.56708	0.46137	0.07907	0.51907	0.08207
X ₈	0.38991	-0.60176	0.37674	0.31465	-0.06529
X ₉	-0.43686	0.63492	-0.11976	0.00047	-0.17723
X ₁₀	-0.64785	0.21916	0.60014	0.09780	-0.21094
X ₁₁	-0.59990	0.34751	0.55012	0.09748	-0.14182

下面进一步利用主成分分析的结果来研究数据集,例如,研究这 38 个城市(样品)的特性。在这种情况下一般都要求计算对应样品的前 m 个主成分得分(即求每个样品的 Y_i 值, $i=1,2,\dots,m$, $m < p$)

$$Y_i = \mathbf{e}_i' \mathbf{X}^* = \frac{\mathbf{p}_i'}{\sqrt{\lambda_i}} \mathbf{X}^* \quad (17-131)$$

其中, $\mathbf{X}^*$ 表示对应某个样品的标准化数据; $\mathbf{p}_i'$ 表示原始变量 $X_1^*, X_2^*, \dots, X_p^*$

在第 i 个主成分上的负荷量构成的向量; λ_i 是相关矩阵 R 的第 i 个特征值。利用 (17-131) 式和 38 个城市的标准化数据,我们可以求出对应每个城市的前 5 个主成分得分值。例如,计算其中两个城市 Boston 和 Honolulu 在第一主成分上的得分值,标准数据如表 17-78 所示。

表 17-78 Boston 和 Honolulu 的标准化数据及第一主成分的负荷量

变量	第一主成分上的负荷量 ($\lambda_1=3.13$)	两个城市的标准化数据	
		Boston	Honolulu
X ₁	0.90936	-0.02123	5.28339
X ₂	-0.14097	-1.55866	-2.41315

X ₃	0.81965	-0.59068	4.32701
X ₄	0.08799	-0.33825	5.39500
X ₅	0.38816	-0.31398	2.14468
X ₆	-0.09243	-0.11679	-0.35764
X ₇	0.56708	-0.15839	2.87651
X ₈	0.38991	-0.84694	0.63674
X ₉	-0.43683	0.86810	-1.04087
X ₁₀	-0.64785	0.09917	-2.78783
X ₁₁	-0.59990	0.00323	-1.98013

Boston 在第一主成分上的得分为

$$[(0.90936)(-0.02123)+(-0.14097)(-1.55866)+\dots+(-0.59990)(0.00323)]$$

$$\div \sqrt{3.13} = -0.729$$

Honolulu 在第一主成分上的得分为

$$[(0.90936)(5.28339)+(-0.14097)(-2.41315)+\dots+(-0.59990)(-0.98013)]$$

$$\div \sqrt{3.13} = -8.68$$

这两个城市在第一主成分上的得分值是很不相同的。前面已经解释过第一主成分是“增长与下降的对比变化”，因此对 Boston，与人口和就业变化相比，人均消费等的变化更强些；而对于 Honolulu，情况则正好相反。根据各个样品在前几个主成分上的得分，我们还可以进一步做聚类分析等研究。

[例 17-31]美国五十个州每十万人中七种犯罪的比率数据如表 13-79 所示。这七种犯罪是：杀人罪（X₁）、强奸罪（X₂）、抢劫罪（X₃）、斗殴罪（X₄）、夜盗罪（X₅）、偷盗罪（X₆）、汽车犯罪（X₇）。我们很难直接从这些数据出发来评价各个州的治安和犯罪情况，但可以使用主成分分析方法，把这些变量概括为两三个综合变量，这样就可以简单地分析这些数据了。

表 13-79 美国 50 州七种犯罪比率数据

州	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
California	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5

Colorado	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Hawaii	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Illinois	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Kansas	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
Maine	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Maryland	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Massachusetts	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Michigan	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Minnesota	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Missouri	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
Montana	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Nebraska	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Nevada	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
New Hampshire	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
New Jersey	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
New Mexico	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
New York	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
North Carolina	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
Ohio	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
North Dakota	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
Oklahoma	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Oregon	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Pennsylvania	5.6	19.0	130.3	128.0	877.5	1642.1	333.2

Rhode Island	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
South Carolina	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
South Dakota	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
Tennessee	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
Texas	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Utah	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Vermont	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
Virginia	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
Washington	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
West Virginia	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Wisconsin	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
Wyoming	5.4	21.9	39.7	173.9	811.6	2772.2	282.0

SPSS 没有提供主成分分析的专用菜单项,但通过因子分析(详见下一章)很容易就可以实现。下面是对[例 13-2]的变量进行主成分分析的操作步骤:

(1) 新建一数据文件,定义变量:State(州)、 X_1 (杀人罪)、 X_2 (强奸罪)、 X_3 (抢劫罪)、 X_4 (斗殴罪)、 X_5 (夜盗罪)、 X_6 (偷盗罪)、 X_7 (汽车犯罪),这些变量中除 State 为字符串型变量外,其余变量均为数值型变量。

(2) 选择菜单[Analyze]=>[Data Reduction]=>[Factor...],打开如图 13-116 所示的[Factor Analysis(因子分析)]主对话框。选定左边列表中的变量 X_1 、 X_2 、 X_3 、 X_4 、 X_5 、 X_6 、 X_7 ,单击按钮使之进入[Variables]列表框。

(3) 单击主对话框中的[Descriptive...]按钮,打开[Factor Analysis: Descriptives]子对话框(如图 13-117所示),在[Statistics]栏中选择[Univariate descriptives(单变量描述统计量)]项要求输出各变量的均值与标准差,在[Correlation Matrix(相关系数矩阵)]栏内选择[Coefficients(系数)]项要求计算相关系数矩阵,单击[Continue]按钮返回[Factor Analysis]主对话框。

(4) 单击主对话框中的[Extraction...]按钮,打开如图 13-118所示的[Factor Analysis: Extraction]子对话框。在[Method]列表中选择默认因子抽取方法——[Principal components(主成分分析法)],在[Analyze]栏中选择默认的[Correlation matrix]项要求从相关系数矩阵出发求解主成分,在[Exact]栏中选择默认项[Eigenvalues over: 1]。单击[Continue]按钮返回主对话框。

(5) 单击主对话框中的[OK]按钮, 输出结果如表 13-81所示。

(6) 应该注意的是, 表 13-81输出结果中给出的是因子负荷, 并没有给出主成分。我们可以把因子负荷除以相应的相关矩阵特征值平方根, 即:

$$e_i = \frac{\rho_i}{\sqrt{\lambda_i}}$$

其中, ρ_i 表示原始变量 $X_1^*, X_2^*, \dots, X_p^*$ 在第 i 个主成分上的负荷量构成的向量; λ_i 是相关系数矩阵 R 的第 i 个特征值。例如, $0.609/4.115=0.3002$, $-0.7/1.239=-0.6289$ 。结果如表 13-80所示。

前两个主成分的累计贡献率已达 76.49%, 因此前两个主成分就能够很好地概括这组数据。由于第一主成分对所有变量都有近似相等的负荷, 因此可认为是对所有犯罪率的度量。第二主成分在变量 X7 和 X6 上有高的正负荷, 而在变量 X1 和 X4 上有高的负负荷; 在 X5 上存在小的正负荷, 而在 X2 上存在小的负负荷。可以认为该主成分是由于度量暴力犯罪在犯罪性质上占的比重。

表 13-80 因子与主成分

	F1	F2	Y1	Y2
X1	0.609	-0.7	0.3002	-0.6289
X2	0.876	-0.189	0.4318	-0.1698
X3	0.805	0.047	0.3968	0.0422
X4	0.805	-0.382	0.3968	-0.3432
X5	0.893	0.226	0.4402	0.2030
X6	0.725	0.448	0.3574	0.4025
X7	0.599	0.559	0.2953	0.5022
特征值	4.115	1.239		
贡献率	58.79%	17.70%		
累计贡献率	58.79%	76.49%		

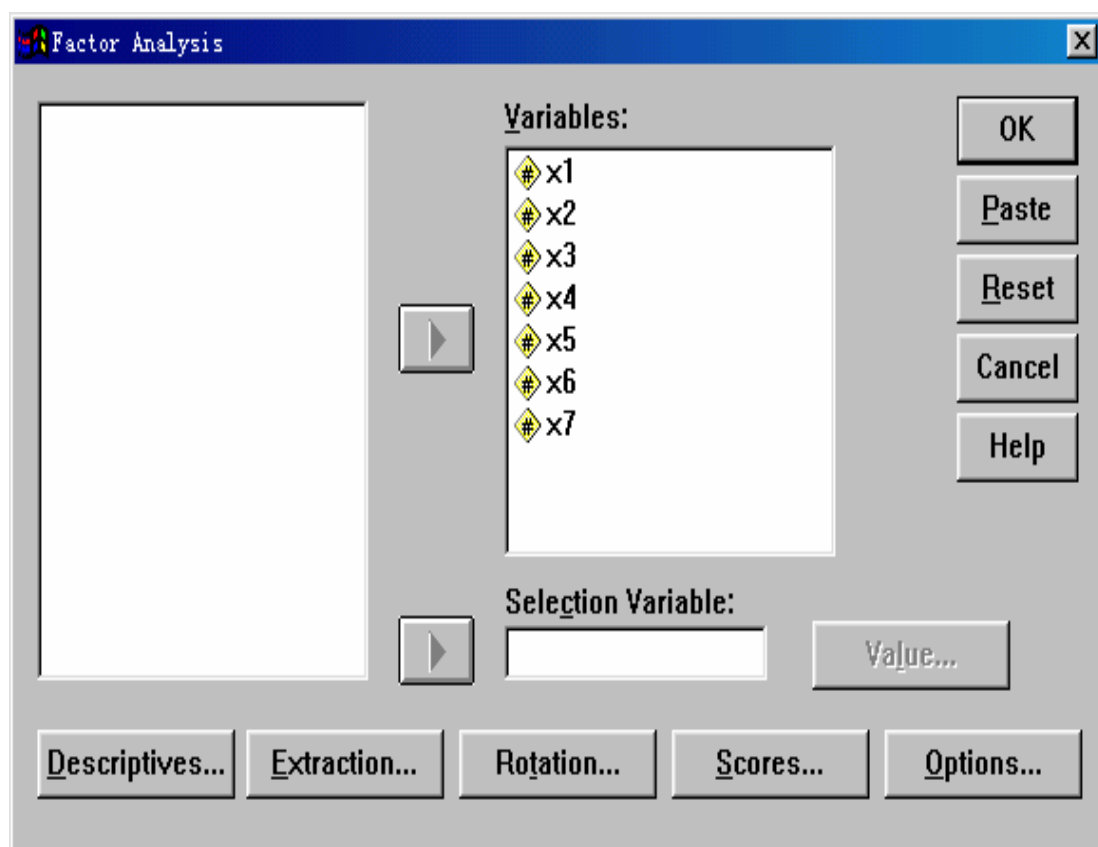


图 13-116 主成分/因子分析主对话框

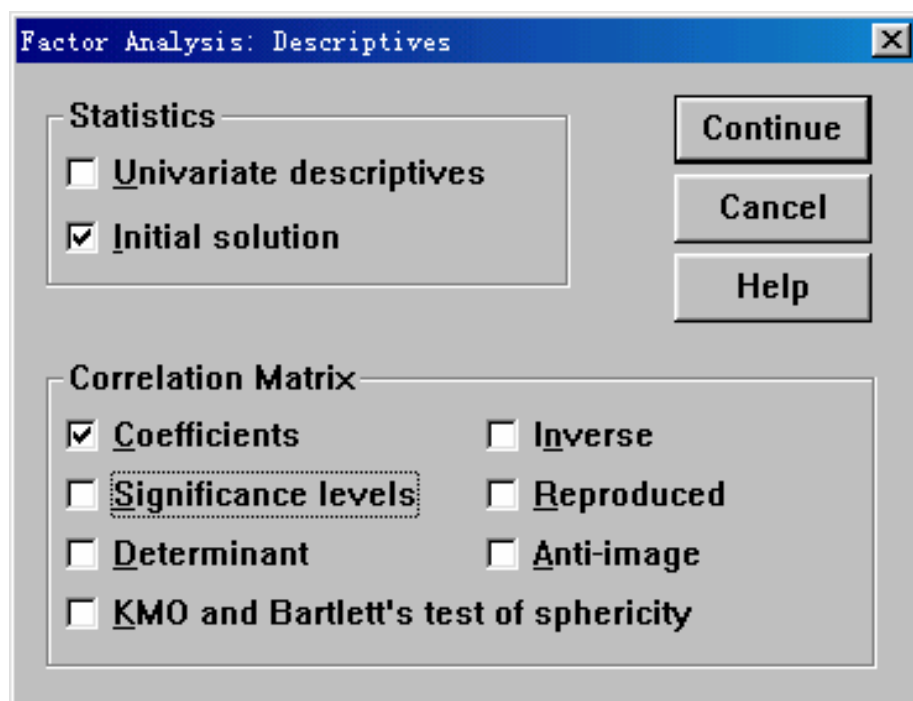


图 13-117 描述统计量子对话框

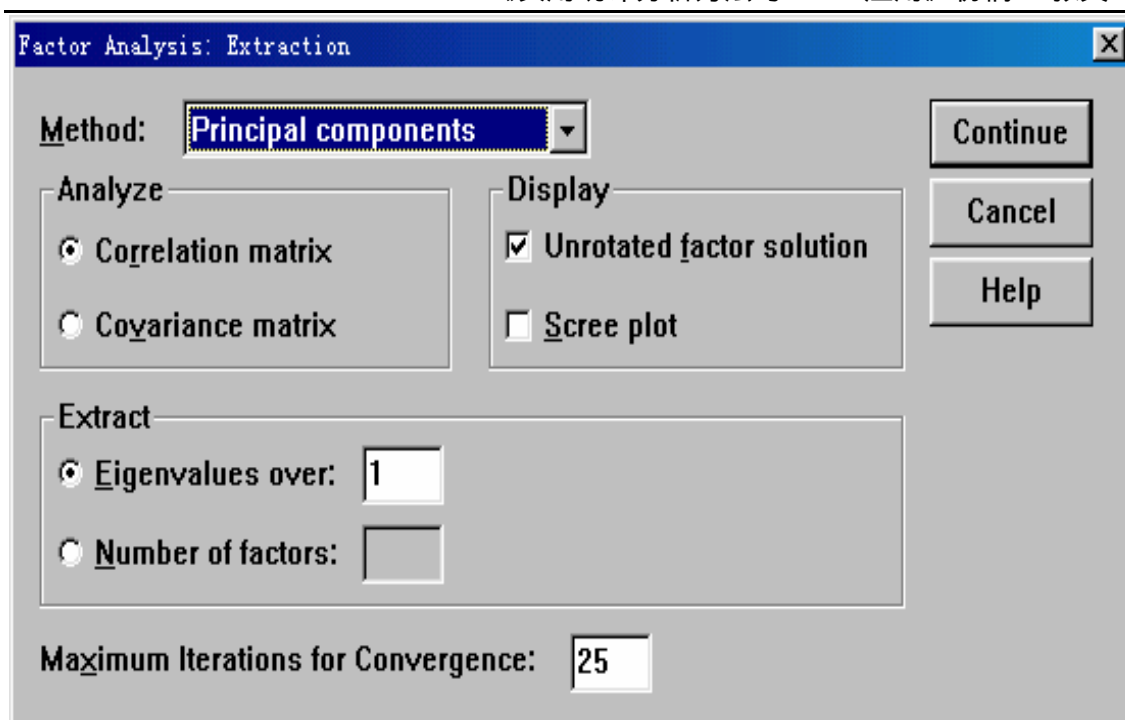


图 13-118 因子提取方法子对话框

表 13-81 主成分分析结果输出

Descriptive Statistics (描述统计量)			
	Mean	Std. Deviation	Analysis N
X1	7.444	3.867	50
X2	25.734	10.760	50
X3	124.092	88.349	50
X4	211.300	100.253	50
X5	1291.904	432.456	50
X6	2671.648	725.383	50
X7	377.526	193.394	50

Correlation Matrix (相关系数矩阵)							
	X1	X2	X3	X4	X5	X6	X7
X1	1.000	.601	.484	.649	.386	.102	.069
X2	.601	1.000	.592	.740	.712	.614	.349
X3	.484	.592	1.000	.557	.637	.447	.591
X4	.649	.740	.557	1.000	.623	.404	.276
X5	.386	.712	.637	.623	1.000	.792	.558
X6	.102	.614	.447	.404	.792	1.000	.444
X7	.069	.349	.591	.276	.558	.444	1.000

Communalities (共同度)

	Initial	Extraction
X1	1.000	.861
X2	1.000	.803
X3	1.000	.650
X4	1.000	.794
X5	1.000	.848
X6	1.000	.726
X7	1.000	.671

Extraction Method: Principal Component Analysis.

Total Variance Explained (可解释的总方差)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.115	58.787	58.787	4.115	58.787	58.787
2	1.239	17.699	76.486	1.239	17.699	76.486
3	.726	10.365	86.852			
4	.317	4.521	91.373			
5	.258	3.685	95.058			
6	.222	3.172	98.230			
7	.124	1.770	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix (主成分/因子矩阵)

	Component	
	1	2
X1	.609	-.700
X2	.876	-.189
X3	.805	.047
X4	.805	-.382
X5	.893	.226
X6	.725	.448
X7	.599	.559

Extraction Method: Principal Component Analysis.

a 2 components extracted.

第四节 案例：新兴股市的多因素模型

一、模型的建立

影响股票价格的因素很多，从长期观点来看，普遍为经济学家承认的宏观经济因素有国民生产总值（GNP）或国内生产总值（GDP）、通货膨胀率、汇率、利率和失业率，许多国家的公司（如日本的大和公司等）运用这 5 个指标（依次记为 X_1 、 X_2 、 X_3 、 X_4 、 X_5 ）的线性模型作为股票收益率的预测模型，其具体形式为：

$$Y = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad (13-132)$$

其中 $\varepsilon \sim N(0, \sigma^2)$ ， Y 为预期收益率，若用 P_t 表示 t 时刻的股票价格，那么 $Y_{t+1} = (P_{t+1} - P_t) / P_t$ 就是 $t+1$ 时刻的收益率。

对股票价格的影响除了上述 5 个宏观因素外，还有一些微观因素，诸如上市公司的数量、发行量和交易量。综合这二类因素，本文考虑如下股票价格模型：

$$P = a_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon \quad (13-133)$$

其中 $\varepsilon \sim N(0, \sigma^2)$ ， P 代表股票价格指数， X_1 、 X_2 、 X_3 分别为上市公司数量、发行量、交易量， X_4 、 X_5 、 X_6 、 X_7 分别为 GNP（或国内生产总值 GDP）、通胀率、对美元的汇率、利率。前 3 个是微观因素，后 4 个是宏观因素。

二、模型应用

下面我们利用模型(13-133)来对一些主要的新兴股市发展情况进行分析。在运用中 X_4 表示 GDP， X_5 、 X_7 分别是累积通货膨胀率、累积利率，

参考何基报、茆诗松：《影响新兴股市的多因素模型及与中国股市的比较》，《统计与

其计算如下：

设 1984 年、1985 年的通货膨胀率(或一年期存款利率)分别为 a_1 、 a_2 ，那么以 1983 年为基期，1984 年、1985 年的 x_5 (或 x_7) 分别为 $1+a_1$ 、

$(1+a_1) \lambda (1+a_2)$ ，对以后的年份以此类推得 x_5 (或 x_7)。我们选了

14 个具有代表性的新兴股市，他们分别是：

拉美：智利 哥伦比亚 墨西哥 委内瑞拉

亚洲：印度 韩国 马来西亚 巴基斯坦 菲律宾 中国台湾 泰国

欧洲：希腊 葡萄牙

非洲：尼日利亚

在估计参数时，以年为单位，诸 x_i 取年平均值，对每一个新兴股市都取

1984 ~ 1993 年共 10 年的数据，按年顺序排号为： $(P_j, x_{1j}, x_{2j}, \dots,$

$x_{7j}) j = 1, \dots, 10$ ，其中 P_j 是 1983 + j 年的股票指数的年平均值。

为消除量纲的影响，对每个新兴股市的数据进行如下变换：

$$P'_j = \frac{P_j}{P_1} X'_{ij} = \frac{X_{ij}}{X_{i1}} \quad i=1, \dots, 7, j=1, \dots, 10 \quad (13-134)$$

对每一新兴股市，用经过 (13-134) 式变换后的数据去拟合模型 (13-133)，其计算结果见表 13-82 第 栏。表 13-82 中第 栏、第 栏分别列出了用经过 (13-134) 式变换后的微观因素数据、宏观因素数据拟合每个新兴股市股价线性模型时的一些计算结果。表 13-83 列出了每个新兴股市的股价指数 P 与 t 个变量的相关系数。从表 13-82 可以看出，模型 (13-133) 基本上适合于这 14 个新兴股市。从股价指数与 7 个变量的复相关系数平方 R^2 来看，除菲律宾为 0.7709，台湾为 0.8426 外，都大于 0.90。从模型检验统计量 F 值来看，除了巴基斯坦、菲律宾、台湾、葡萄牙分别为 5.589、0.961、1.530、2.767 外，其余都大于临界值 $F_{0.1}(7, 2) = 9.35$ 。

下面我们将利用表 13-82、表 13-83 的结果分析新兴股市的发展情况。

1 . 新兴股市既受宏观因素影响 , 又受微观因素影响。从表 13-82 中第 、 第 栏中可以看出有 7 个新兴股市的 $R_1^2 > 0.9$, 而 $R_2^2 > 0.9$ 只有 6 个 , 另外

$R_{1(\min)}^2 = 0.4348 > R_{2(\min)}^2 = 0.2642$, 因此 , 总的说来 , 微观因素对新兴股市

的影响程度比宏观因素似乎要强。 $R_{1(\min)}^2 = 0.9777$, $R_{2(\min)}^2 = 0.4348$,

$R_{2(\max)}^2 = 0.9919$, $R_{2(\min)}^2 = 0.2642$, 这又表明和微观因素相比 , 宏观因素对

各新兴股市的影响程度参差不齐 , 相差悬殊 , 微观因素对各新兴股市的影响程度差别相对较小 , 对新兴股市的影响更具有普遍性

2 . 股市的成熟性比较。股票市场的发展是经济发展的 “ 晴雨表 ” , 因此 , 好的股票市场的运行应和宏观经济运行相一致。一般认为 , 如果股市运行和国民经济的运行有比较稳定的相关关系 , 那么该股市基本上是成熟的。如何反映这种稳定的相关关系 ? 反映宏观经济运行好坏的指标有许多 , 而最能综合反映国民经济运行的指标是 G N P 或 G D P , 如用股价指数 P 作为衡量股市运行的指标 , 用 G N P 或 G D P 作为衡量国民经济运行的指标 ,

那么就可以用 P 与 x_4 (即 G D P) 的相关系数 ρ 度量这种 “ 稳定的相关

关系 ”。取 $\alpha = 0.05$, 相关系数检验临界值 $\rho_0 = 0.6319$, 这样就可把新兴

股市分为二类 , 一类趋于成熟 ($\rho \geq \rho_0$) , 它们为智利、印度、墨西哥、

尼日利亚、泰国、菲律宾和委内瑞拉。另一类为不成熟股市 ($|\rho| < \rho_0$

= , 它们为哥伦比亚、希腊、韩国、马来西亚、巴基斯坦、台湾和葡萄牙。从股价指数 P 与 G D P 的相关系数可以看出 , 新兴股市的运行与国民经济的关系表现出参差不齐。两极分化现象严重。成熟度高的股市 (如智利、印度、墨西哥和尼日利亚) 与 G D P 的相关系数为 $\rho > 0.96$, 股市的总体表现和国民经济的运行比较一致。表现比较差的股市 (如韩国) 与 G D P 的相关系数为 $\rho < -0.60$, 股市的运行和经济的发展不能很好地接轨。

表 13-82 模型(13-133)关于每个新兴股市的拟合情况表

国家	I	II	III
----	---	----	-----

和地区	P 与 T 个变量复 相关 R^2	检验统计 量 F 值 $F_{0.1}(7,2)=$ 9.35	P 与微 观因 素 R_1^2	检验 F 值 $F_{0.05}=4.76$ $F_{0.1}(3,6)=$ 3.29	P 与宏 观因 素 R_2^2	检验 F 值 $F_{0.05}(4,5)=$ 5.19 $F_{0.1}(4,5)=$ 3.52
智利	0.9981	148.1	0.9195	22.84	0.9734	45.72
哥伦比亚	0.9765	11.87	0.4384	1.561	0.9259	15.63
希腊	0.9890	25.68	0.9777	87.78	0.8570	7.49
印度	0.9928	39.46	0.9226	23.85	0.9919	153.1
韩国	0.9884	24.26	0.8373	10.29	0.9304	16.71
马来西亚	0.9736	10.53	0.9597	47.60	0.8722	8.530
墨西哥	0.9958	68.25	0.9089	19.94	0.9898	112.70
尼日利亚	0.9922	36.18	0.9739	74.62	0.9898	121.10
巴基斯坦	0.9514	5.589	0.7888	7.471	0.5701	1.658
菲律宾	0.7709	0.961	0.6708	4.075	0.7469	3.688
泰国	0.9727	10.20	0.9352	28.87	0.8728	8.577
中国台湾	0.8426	1.530	0.6824	4.296	0.7239	3.278
委内瑞拉	0.9743	10.85	0.8541	11.70	0.8396	6.544
葡萄牙	0.9064	2.767	0.7818	7.166	0.2642	0.4489

表 13-83 股价指数 P 与 7 个变量的相关系数表

国 家 和	P 与 上	P 与股票	P 与股票	P 与	P 与通胀	P 与汇	P 与利	P 与股市
-------	-------	-------	-------	-----	-------	------	------	-------

地区	市 公 司 数量 X1	发 行 量 X2	交 易 量 X3	GDP X4	率 X5	率 X6	率 X7	规模 (市价总 值/GDP)
智利	0.7773	0.7960	0.3949	0.9796	0.9702	0.9168	0.9790	0.9821
哥 伦 比 亚	-0.1891	-0.1399	0.2072	0.1866	0.2329	0.1759	-0.0519	-0.5223
希腊	0.9200	0.9871	0.4402	0.9048	0.8334	0.8672	0.7310	0.9273
印度	0.9605	0.0350	0.9691	0.9743	0.9953	0.9775	0.9929	0.8972
韩国	0.7557	0.6947	-0.2051	0.6406	-0.7430	0.7156	0.9478	0.7802
马 来 西 亚	0.9111	0.9366	-0.6031	0.8865	0.1937	0.8942	0.9753	0.9263
墨西哥	0.4507	0.9090	-0.1069	0.9684	0.9151	0.7612	0.9244	0.9720
尼 日 利 亚	0.9851	0.6215	0.6980	0.9908	0.9676	0.9821	0.9942	0.7388
巴 基 斯 坦	-0.2713	-0.1564	-0.1862	0.3002	0.3210	0.4004	0.3207	0.3335
菲律宾	0.8059	0.6748	0.5060	0.8250	0.8318	0.7597	0.8425	0.9917
中 国 台 湾	0.5313	0.6059	0.6387	0.5772	0.4582	-0.7700	0.5515	0.9886
泰国	0.8980	0.9140	0.7019	0.9070	0.8938	-0.1161	0.9029	0.9857
委 内 瑞 拉	0.4627	0.6588	0.9132	0.8471	0.8408	0.8340	0.7698	0.9507
葡萄牙	0.8165	0.6104	0.4032	0.4615	0.4898	-0.1468	0.4977	0.9696

表 13-84 微观因素和宏观因素对各个股市影响情况表

国家 和地 区	常数项	回归方程系数							F 值
		X1	X2	X3	X4	X5	X6	X7	
智利	-10.45	0	0	0	6.966	0	0	0	190.07
哥伦 比亚	回归方程不显著								
希腊	0.8785	0	0	0.5132	0	0	0	0	304.60
印度	-6.434	0	0	0	0	0	7.596	0	838.79

韩国	-1.808	3.406	0	0	0	0	0	0	12.44
马来西亚	-4.574	0	0	0.1216	0	5.283	0	0	83.08
墨西哥	-100.81	0	93.478	0	16.88	0	-15.15	0	310.81
尼日利亚	-3.472	0	0	0	0	0	0	3.914	679.83
巴基斯坦	回归方程不显著								
菲律宾	-14.8	0	0	0	0	0	0	13.82	19.57
泰国	-1.196	0	3.048	-0.288	0	0	0	0	50.24
中国台湾	18.076	0	0	0	0	0	-17.472	0	11.65
委内瑞拉	0.224	0	0	2.899	0	0	01	0	40.14
葡萄牙	-2.864	3.028	-1.676	0	2.9128	0	0	0	11.85

3. 股票价格指数与宏观因素诸指标的关系。在股价指数 P 与通胀率的相关系数中, 除哥伦比亚, 巴基斯坦、台湾、葡萄牙都小于 0.5 外, 其余 10 个股市中有 9 个股市的相关系数 (韩国为 0.6406) 都大于 0.8, 这表明这些股市股价的上涨很大程度上受物价的上涨的影响。股价指数 P 与汇率的相关系数中, 韩国、台湾为最小, 分别为 -0.7430、-0.7700, 表明了这两个亚洲股市受美元的负影响比较大。在其余的 12 个股市中有 7 个股市的相关系数大于 0.75, 剩下 5 个股市的相关系数都小于 0.45, 这说明了新兴股市受汇率的影响大小不一, 呈现出两极分化趋势。由于 X_7 是累积利率, 呈递增趋势, 而股价指数在股市发展正常时一般也有增长趋势, 因此二者呈现正相关。在成熟的股市, 长期说来, 二者的相关性较强, 但从表 13-83 中可以看出, 哥伦比亚、巴基斯坦、台湾、葡萄牙这些股市中, 二者的相关系数不大, 说明了二者的增长幅度和趋势相差很大, 这几个股市的利率波动不大, 因此累积利率 x_7 稳定增长, 这就说明了这几个股市股价暴涨暴跌现象严重, 并没有一种稳定的增长趋势, 而且短期投机者比较多, 使得

股价大起大落。

4 . 股价指数与微观因素诸指标的关系。在股价指数 P 与上市公司数量 X_1 的相关系数中, 有 9 个是大于 0.77 的, 在 P 与股票发行量 X_2 的相关系数中, 也有 9 个是大于 0.65 的, 在 P 与交易量 X_3 的相关系数中也有一半是大于 $\rho_0 = 0.6319$ 的, 这说明了虽然新兴股市的规模在不断扩大, 但对股价的稳定所起的作用不大, 在表 13-83 中的股价指数 P 与股市规模的相关系数中, 有 10 个大于 0.95, 这更进一步地说明了新兴股市的股市规模的扩大并没有缓解股市的供需紧张局面, 投资大众的热情很高, 股市发展规模亟需进一步扩大。

5 . 各个股市受宏观因素和微观因素影响的具体情况不同, 找出影响每个股市的重要因素, 对其以后的发展无疑是有帮助的。在这里, 我们用逐步回归方法找出影响各新兴股市的重要因素, 表 13-84 列出了有关的计算结果 (其中引入变量时作检验的 F 统计量 F_1 和剔除时的统计量 F_2 均为 3.0)。除哥伦比亚、巴基斯坦不能挑出重要因素外, 其他 12 个股市挑出的重要因素中, X_3 有 4 个, X_2 、 X_4 、 X_6 各有 3 个, X_1 和 X_7 各有 2 个, X_5 有 1 个。没有一个因素超过 7 个, 说明了虽然同是新兴股市, 但不同的股市受不同的重要因素的影响不同。在这 7 个因素中, 没有一个共同的影响这些新兴股市的重要因素, 各新兴股市的具体情况相差很大, 表现出明显的参差不齐, 不象发达股市 (如美国、英国) 无一例外地受到重要因素 GNP 、汇率、通胀率、利率的影响。另外, 从表 13-84 中可以看出, 大部分的新兴股市只能挑选出一个重要因素, 说明了新兴股市在发展过程中没有注意到股市的均衡性, 往往只受一个重要因素的左右, 这样的股市容易出现大起大落现象。

三、主成份分析

现在, 我们用主成份分析这 7 个因素, 找出几个综合指标, 从而在分析时可用较少的变量来代替上述 7 个变量, 给分析问题带来很大的

方便，将 7 个指标作为自变量即 $X = (x_1, x_2, \dots, x_7)$ ，对每一新兴股市取 1984~1993 年共 10 年的年平均数据进行 (13-134) 式变换组成 $X(140, 7)$ 的样本矩阵，以相关阵出发来求主成份。由于样本矩阵 $X(140, 7)$ 数据太多，在这里只列出它们的相关阵 $X(7, 7)$ 如下：

	1	2	3	4	5	6	7
1	1.000	0.522	0.801	-0.038	-0.063	-0.092	0.01
2	0.522	1.000	0.692	0.090	0.059	0.024	0.550
3	0.801	0.692	1.000	0.040	-0.015	-0.085	0.282
4	-0.38	0.09	0.040	1.000	0.908	0.650	0.314
5	-0.063	0.059	-0.015	0.908	1.000	0.727	0.245
6	-0.092	0.024	-0.085	0.650	0.727	1.000	0.315
7	0.010	0.550	0.282	0.314	0.245	0.293	1.000

其特征值为：2.750、2.442、0.946、0.387、0.254、0.145、0.077，特征值贡献率为：39.3%、34.9%、13.5%、5.5%、3.6%、2.1%、1.1%，由于前 3 个特征值累积贡献率达 87.7%，故可取 3 个主成份，得到因子载荷阵如下：

因子（主成份）载荷阵 $A(7, 3)$

	1	2	3
1	-0.038	0.954	-0.078
2	0.002	0.644	0.656
3	-0.026	0.916	0.249
4	0.930	0.030	0.112
5	0.962	0.000	0.049
6	0.842	-0.096	0.130
7	0.217	0.045	0.939

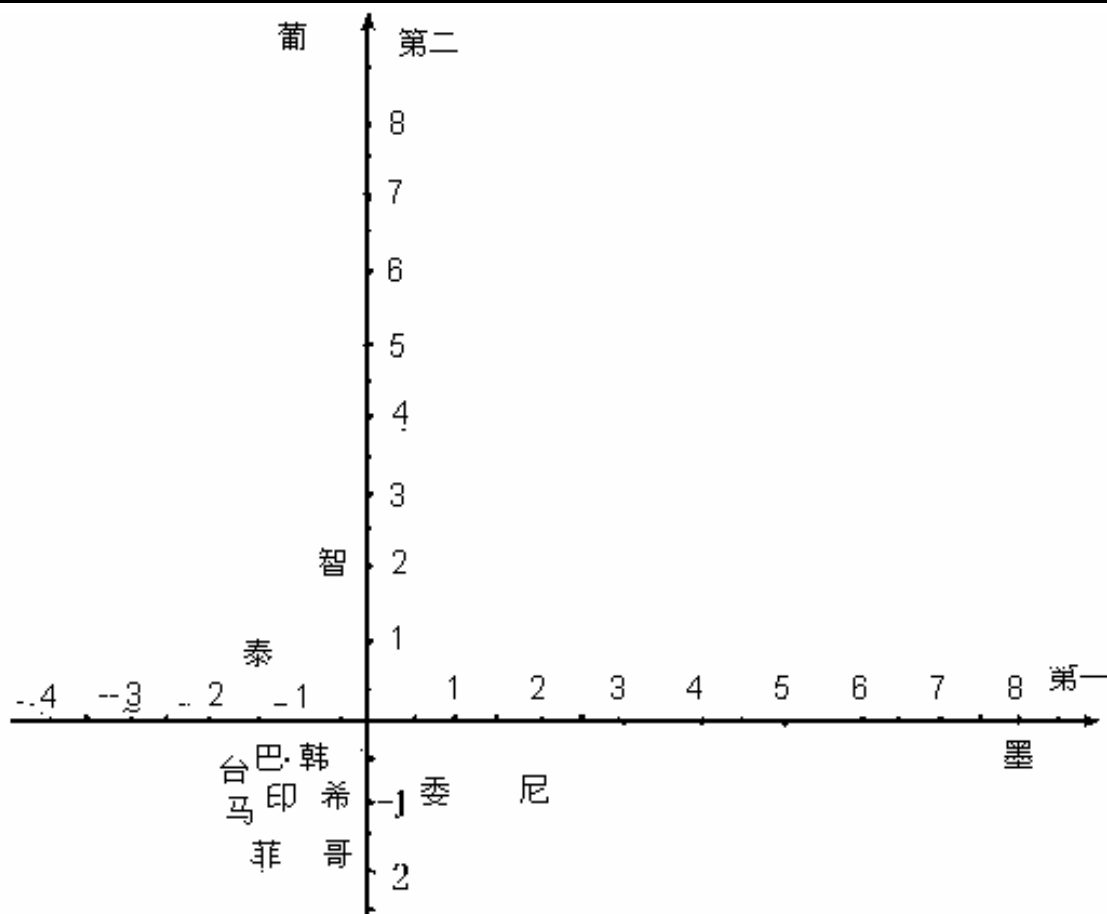
3 个主成份对各变量的方差贡献分别为：91.7%、84.5%、90.2%、87.9%、92.8%、73.4%和 93.1%。以上 3 个主成份提取的信息量占 87.7%，其中，在第一主成份对以上 7 个变量的因子载荷中，比较大的有第 4、第 5、第 6 分量，分别是 0.930、0.962、0.842，因此第一主成份主要反映了 GDP、通胀率、汇率 3 个变量，我们称之为宏观因子。在第二主成份对以上 7 个变量的因子载荷中，比较大的有第 1、第 2、第 3 分量，分别是 0.954、0.644、0.916，因此第二主成份主要反映了上市公司数量、发

行量、交易量三个变量，我们称之为微观因子。在第三主成份对 7 个变量的因子载荷中，比较大的有第 2、第 7 分量，分别是 0.656、0.939，因此第三个主成份主要反映了发行量和利率这两个变量，但更主要是反映了利率变量，我们称之为利率因子。通过以上分析我们知道，在众多的影响因素中，宏观因子、微观因子、利率因子是影响新兴股市的主要的共同的因素。由于每个新兴股市有 10 个样本，共组成 140 个样本，计算经过(3)式变换后的 140 个数据的前二个主成份的值，对每一新兴股市，分别取属于该股市的 10 个数据的前二个主成份的均值，列于表 13-85，并将它作成附图。从表 13-85 可以看到，在第一主成份中，墨西哥的数值为最大，而其第二主成份的值仅为 -0.141，表明了宏观因素对墨西哥股市的影响比微观因素要大，以马来西亚和台湾的第一主成份的值为最小，均比第二主成份的值小，说明了这两个股市中，宏观因素对股市的总的影响不大。在第二主成份中，以葡萄牙为最大，其值为 4.33，而第一主成份仅为 -0.372，表明了葡萄牙股市受微观因素的影响比受宏观因素的影响大得多，而且股市和宏观经济运行联系不大，菲律宾和哥伦比亚的第二主成份最小，说明了相对于宏观因素而言，微观因素对股市影响不大，股市主要受宏观因素影响。亚洲几个股市的第一主成份比较接近，因此，大体上可分成如附图所示的 6 组，从图中可看出两点：1. 以上 7 个因素对新兴股市的影响很不均匀，呈两极分化状态，有的只受微观因素的影响（如葡萄牙），有的只是宏观因素起作用（如哥伦比亚），体现了新兴股市发展的不均衡性，这种现象令人担忧。2. 亚洲的几个股市比较靠近，呈板块结构。

表 13-85 第一、第二主成份对各股市的影响

国家和地区	第一主成份（均值）	第二主成分（均值）
智利	-0.129	0.767
哥伦比亚	-0.171	-0.682
希腊	-0.214	-0.86
印度	-0.524	-0.247
韩国	-0.509	-0.09
马来西亚	-0.739	-0.285
尼日利亚	0.865	-0.573
墨西哥	3.97	-0.141
巴基斯坦	-0.605	-0.143
菲律宾	-0.569	-0.633

中国台湾	-0.656	-0.156
泰国	-0.615	0.8
委内瑞拉	0.266	-0.554
葡萄牙	-0.372	4.33



附图 第一、第二主成份点图

四、我国股市与其他股市比较

我国股市起步晚，数据较少，还不能用模型来说明，只能从几个侧面来比较。

1. 与发达股市比较。在表 13-86 中，加拿大、美国、英国股市的股价指数与 GNP 的相关系数都很大，这些股市与国民经济运行非常一致，我国股市 1992~1995 年的上证指数与 GDP 的相关系数为 $\rho = -0.3467$ ，呈负相关，说明了我国的股市发展还没有和宏观经济接轨。这几个发达股市股价指数与 GNP、物价指数的复相关系数分别为 0.998、1.000、1.000，从而可看出，发达股市与宏观经济是息息相关的。我国股市的上证指数与物价指数、GDP 的复相关系数只有 0.385，这表明我国股市的发展游离于宏

观经济之外。从以上比较可以看出，我国股市与发达股市的差距还很大。

2. 与新兴股市比较。我国股市的股价指数与 GDP 的相关系数为 $\rho =$

-0.3467，与其他新兴股市相比，只比马来西亚的 $\rho = -0.6801$ 大，按 $\rho_0 =$

0.6319 归类，应归入差的一类，说明了我国股市的总体表现和表现好的新

兴股市是有一段差距的。另外，大部分的新兴股市既受宏观因素的影响，又

受微观因素的影响，而我国股市受二者的影响不大。一方面表现在国民经济持经济持续稳定增长，但股票价格这几年却大起大落，股市波动频繁。

另一方面，股票价格的波动与政策、扩容、消息有密切关系，一有风吹草动，马上就是草木皆兵，股市动荡不安，市场人士反映这几年我国股市是政策市、消息市。

表 13-86 我国股市与发达股市的比较

(1985 年 = 100%)

	1986 年	1987 年	1988 年	1989 年	股票指 数与 G N P 相 关系数	股票与 物价指 数相关 系数	复相 关系 数 R
加拿大 多伦多 股票指 数	105.7	108.9	116.9	136.9	0.9298	0.9525	0.998
伦敦金 融时报 股票指 数	116.1	121.4	128.6	169.4	0.8932	0.9536	1.000
美国 道·琼 斯工业 平均指 数	122.6	125.4	140.2	178.0	0.9226	0.9453	1.000
上证指 数	1992 年	1993 年	1994 年	1995 年	与 G D P 相关 系数		
1990 年	668.52	1013.4	674.1	660.8	-0.3467	-0.384	0.385

= 100							
-------	--	--	--	--	--	--	--

资料来源：《证券市场导报》，《国际金融统计》。1995 年上证指数为当年 1-10 月指数

通过以上分析可知，和发达股市及其他新兴股市相比，我国股市与国民经济的运行还没有很好地接轨。虽然我国股市的规模不断扩大，但由于发展较晚，使得股市规模相对于国民生产总值来说，显得太小（1994 年流通股票市值只占 G N P 的不足 5 % ）。这样，股市在国民经济运行中所起的作用也小。因此，我国股市在以后的发展中要进一步扩大股市规模，同时要规范化建设，使市场朝有效化方向发展。唯有这样，才能吸引更多的投资者，更好地保证股市规模进一步扩大，从而使股市与宏观经济运行更好地接轨。

第十四章 因子分析

第一节 因子分析模型

一、因子分析的基本思想

因子分析可以看成是主成分分析的一种推广。它的基本目的是，用少数几个因子 $F_1, F_2, \dots, F_m$ 去描述许多变量之间的关系。被描述的变量 $X_1, X_2, \dots, X_p$ 是可以观测的随机变量，即显在变量。而这些因子是不可观测的潜在变量。在社会科学、经济科学、管理科学、心理学、行为科学、教育学等领域中，许多基本特征例如“态度”、“认识”、“爱好”、“能力”、“智力”等等实际上是不可能直接观测的，我们把它们看成是潜在变量。而对人的测量例如“教育水平”、“收看电视频度”、“是否喜欢某种节目”、“考试成绩”、“平均收入”等等是显在的，可以观测的。对人的测量可以看成是一些潜在变量（不可观测的基本特征）的表现。因子分析正是利用这些潜在变量或本质因子（基本特征）去解释可观测的变量的一种工具。

因子分析的思想是，将观测变量分类，将相关性较高即联系比较紧密的变量分在同一类中，而不同类的变量之间的相关性则较低。那么每一类的变量实际上就代表了一个本质因子，或一个基本结构。因子分析就是寻找这种类型的结构，或者叫做模型。

“因子分析”的名称于 1931 年由 Thurstone 首次提出，但它的概念起源于二十世纪初 Karl Pearson 和 Charles Spearman 等人关于智力测验的统计分析。近年来，随着电子计算机的高速发展，人们将因子分析方法成功地应用于各个领域，使得因子分析的理论和方法更加丰富。

二、正交因子模型的定义

设 $\mathbf{X}' = (X_1, X_2, \dots, X_p)$ 是 $p \times 1$ 的随机向量。 $\mathbf{X}$ 的协方差矩阵

$$\text{Cov}(\mathbf{X}) = \Sigma \quad (18-135)$$

$\mathbf{F}' = (F_1, F_2, \dots, F_p)$ 是 $m \times 1$ 的标准化的正交公共因子向量 ($m < p$)，即假定

$$E(\mathbf{F}) = \mathbf{0}, \text{Cov}(\mathbf{F}) = \mathbf{I} \quad (18-136)$$

$\boldsymbol{\varepsilon}' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ 是 $p \times 1$ 的特殊因子向量 (或误差向量), 并假定其均值为 0, 协方差矩阵为对角矩阵 (说明各个 ε 之间互不相关), 即

$$E(\boldsymbol{\varepsilon}) = 0$$

$$\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Phi} = \text{diag}(\Phi_1, \Phi_2, \dots, \Phi_p) = \begin{pmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_p \end{pmatrix} \quad (18-137)$$

并假设公共因子 $F_1, F_2, \dots, F_p$ 与各个特殊因子 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ 都互不相关 (或 $\mathbf{F}$ 与 $\boldsymbol{\varepsilon}$ 相互独立), 即

$$\text{Cov}(\mathbf{F}, \boldsymbol{\varepsilon}) = \mathbf{0} \quad (18-138)$$

在以上假定下, 正交因子模型可以写成以下的矩阵形式

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \boldsymbol{\varepsilon} \quad (18-139)$$

其中矩阵 $\mathbf{A} = (a_{ij})$ ($p \times m$ 阶) 称为因子负荷矩阵, a_{ij} 表示第 i 个变量 X_i 在第 j 个因子 F_j 上的负荷。

因子分析模型(18-139)可以具体地写成:

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + \varepsilon_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + \dots + a_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= a_{p1}F_1 + a_{p2}F_2 + \dots + a_{pm}F_m + \varepsilon_p \end{aligned} \quad (18-140)$$

该模型中, 第 i 个特殊因子 ε_i 仅与第 i 个变量 X_i 有关系。而第 i 个公共因子 F_i 则与所有 p 个变量都有关系。

三、正交因子模型与回归模型的比较

比较(18-140)的其中一个式子

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + \varepsilon_i \quad (18-141)$$

与第九章的回归模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (18-142)$$

可以发现它们的形式是类似的, 但参数的意义与“自变量”的性质都不相同, 现把它们的比较列于表 18-87 中。

表 18-87 正交因子模型与回归模型的比较

	正交因子模型	回归模型
要估计的主要参数	因子负荷系数 $a_{i1}, a_{i2}, \dots, a_{im}$ 和 ε_i 的方差 $\text{Var}(\varepsilon_i)$	回归系数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 和 ε 的方差 $\text{Var}(\varepsilon)$
“自变量”的性质	$F_1, F_2, \dots, F_m$ 是不可观测的潜在变量	$X_1, X_2, \dots, X_p$ 可观测的 显在变量
“自变量”之间的个数	p 是未知的, 需要估计	p 是已知的
“自变量”之间的关系	是相互独立的	可能是相关的

四、负荷矩阵 A 的意义

负荷矩阵 A 是 $p \times m$ 矩阵 ($m < p$)

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{im} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pj} & \cdots & a_{pm} \end{pmatrix} \quad (18-143)$$

1、A 中任一元素 a_{ij} 是第 i 个变量 X_i 与第 j 个公共因子 F_j 的协方差，
即

$$\text{Cov}(X_i, F_j) = a_{ij} \quad (18-144)$$

如果观测变量 X_i 也是标准化变量，那么 a_{ij} 就是 X_i 与 F_j 的相关系数，它表示 X_i 与 F_j 线性联系的紧密程度。第 i 行的因子负荷量 $a_{i1}, a_{i2}, \dots, a_{im}$ 说明了第 i 个变量 X_i 依赖于各个因子的程度，而第 j 列的因子负荷量 $a_{1j}, a_{2j}, \dots, a_{pj}$ 则说明了第 j 个因子 F_j 与各个变量的联系程度，常常根据该列负荷中绝对值较大的负荷所对应的变量来说明这个因子的意义。

2、A 中任一元素 a_{ij} 的平方和 h^2 等于第 i 个变量 X_i 的方差与特殊因子 ε_i

的方差之差，也就是

$$Var(X_i) = h_i^2 + Var(\varepsilon_i) \quad (18-145)$$

$$\sigma_{ii} = (a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2) + \Phi_i$$

变量方差=公因子方差+特殊因子方差

称

$$h_i^2 = \sum_{j=1}^m a_{ij}^2 \quad (i=1,2,\cdots,p) \quad (18-146)$$

为公因子方差，或叫做 X_i 的共通性（或共同度）(Communalities)，它表示 m 个公共因子对第 i 个变量 X_i 的方差贡献。 h_i^2 越大，表示 X_i 对这 m 个因子的共同依赖程度越大，也就是说，用这 m 个因子描述变量 X_i 就越有效。

如果 X_i 是标准化变量，那么(18-145)式就变成了

$$h_i^2 + \Phi_i = 1 \quad (18-147)$$

因此共同度 h_i^2 就等于公共因子的方差在变量 X_i 的总方差中所占的比例。

3、A 中任一列元素的平方和

$$g_j^2 = \sum_{i=1}^p a_{ij}^2 = a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2 \quad (j=1,2,\cdots,m) \quad (18-148)$$

表示第 j 个公共因子的方差贡献，它与 p 个变量 $X_1, X_2, \cdots, X_p$ 的总方差之比

$$F_j \text{ 的贡献率} = \frac{g_j^2}{g^2} = \frac{a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2}{Var(X_1) + Var(X_2) + \cdots + Var(X_p)} \quad (18-149)$$

叫做第 j 个公共因子 F_j 的方差贡献率。贡献率越大，该因子就相对地越重要。

如果 $X_1, X_2, \cdots, X_p$ 都是标准化变量，那么第 j 个因子的贡献率就等于

$$F_j \text{ 的贡献率} = \frac{g_j^2}{p} = \frac{a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2}{p} \quad (18-150)$$

4、正交因子模型的协方差结构

根据正交因子模型(18-139)以及有关的假定，可以证明

$$\Sigma = \mathbf{A}\mathbf{A}' + \Phi \quad (18-151)$$

也就是说，在正交因子模型的假定下，随机向量 $\mathbf{X}$ 的协方差矩阵 Σ 要分解成两部分。应该注意，这种分解并不是唯一的。例如，设 $\mathbf{T}$ 是一个 $m \times m$ 的正交矩阵，即

$$\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I} \quad (18-152)$$

它在几何上对应 $\mathbf{X}$ 坐标系的一个旋转。旋转后的因子负荷矩阵 $\mathbf{A}^*$ 和公共因子向量 $\mathbf{F}^*$ 分别为

$$\mathbf{A}^* = \mathbf{A}\mathbf{T}, \mathbf{F}^* = \mathbf{T}'\mathbf{F} \quad (18-153)$$

显然 $\mathbf{A}^*$ 和 $\mathbf{F}^*$ 也满足正交因子模型(18-139)

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{A}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \quad (18-154)$$

而且可以验证 $\mathbf{F}^*$ 满足模型要求的假定(18-136)。这样，不同的负荷矩阵 $\mathbf{A}$ 和 $\mathbf{A}^*$ 都产生了相同的协方差矩阵

$$\Sigma = \mathbf{A}\mathbf{A}' + \Phi = \mathbf{A}^*\mathbf{A}^{*'} + \Phi \quad (18-155)$$

不过，尽管正交因子模型具有不确定性，但是如果不考虑旋转（不考虑正交矩阵 $\mathbf{T}$ ），那么因子负荷矩阵还是唯一确定的。而且即使在旋转的情况下，由(18-154)式可知，共通性也是保持不变的。

由协方差结构(18-151)式，我们还可以知道，两个变量 X_r 和 X_s 之间的协方差等于因子负荷阵中第 r 行与第 s 行对应元素乘积之和：

$$\text{Cov}(X_r, X_s) = a_{r1}a_{s1} + a_{r2}a_{s2} + \cdots + a_{rm}a_{sm} \quad (18-156)$$

如果 X_r 和 X_s 都是标准化变量，那么上式就表示变量间的相关系数。

[例 18-32]假定在一个很大的总体中（例如某学校的学生）进行了测量语言能力和数字能力的六项考试。考试成绩都化为标准分数。假定 X_1, X_2, X_3 是语言能力的三项不同考试的标准分数。 X_4, X_5, X_6 是数字能力的三项不同考试的标准分数。 $\mathbf{X}' = (X_1, X_2, X_3, X_4, X_5, X_6)$ 的相关系数矩阵为

$$\mathbf{p} = \begin{pmatrix} 1.00 & & & & & \\ 0.24 & 1.00 & & & & \\ 0.28 & 0.42 & 1.00 & & & \\ 0.20 & 0.30 & 0.35 & 1.00 & & \\ 0.24 & 0.36 & 0.42 & 0.78 & 1.00 & \\ 0.28 & 0.42 & 0.49 & 0.75 & 0.72 & 1.00 \end{pmatrix}$$

按照两个正交因子的模型，求出其因子负荷矩阵为

$$\mathbf{A} = \begin{pmatrix} 0.272 & 0.293 \\ 0.409 & 0.439 \\ 0.477 & 0.513 \\ 0.926 & -0.179 \\ 0.848 & 0.031 \\ 0.843 & 0.172 \end{pmatrix}$$

- 1) 写出正交因子模型；
- 2) 求各个变量的共同度（公共因子方差）以及对应的特殊因子方差；
- 3) 计算每个因子方差贡献率以及两个因子的累计方差贡献率；
- 4) 试说明两个因子的意义。

解：1) 正交因子模型可以写成

$$X_1 = 0.272F_1 + 0.293F_2 + \varepsilon_1$$

$$X_2 = 0.409F_1 + 0.439F_2 + \varepsilon_2$$

$$X_3 = 0.477F_1 + 0.513F_2 + \varepsilon_3$$

$$X_4 = 0.926F_1 - 0.179F_2 + \varepsilon_4$$

$$X_5 = 0.848F_1 + 0.031F_2 + \varepsilon_5$$

$$X_6 = 0.843F_1 + 0.172F_2 + \varepsilon_6$$

2)、3)的计算结果归纳在表 18-88中。

表 18-88 因子分析解

变量	因子负荷		共同度	特殊方差
	F_1	F_2		
X_1	0.072	0.293	0.16	0.84
X_2	0.409	0.439	0.36	0.64
X_3	0.477	0.513	0.49	0.51
X_4	0.926	-0.179	0.89	0.11
X_5	0.848	0.031	0.72	0.28
X_6	0.843	0.172	0.74	0.26
方差贡献率	45.9%	10.1%	56%	44%
累计方差贡献率	45.9%	56%		

4) 由第一因子 F_1 上的负荷看, X_4 、 X_5 、 X_6 的负荷量都很大, 因此 F_1 主要是数字能力因子; 相反地, F_2 则可称为语言能力因子。但从共同度来看, X_1 和 X_2 , 特别是 X_1 对这两个因子的依赖程度较小。它们的方差有相当大的部分仍不能被这两个公共因子所解释, 因此被包含在特殊因子的方差之中。

第二节 因子分析模型估计方法

当给定 p 个变量 $X_1, X_2, \dots, X_p$ 的 n 组观测值时, 如何从样本协方差矩阵 S 或样本相关矩阵 R 出发(将 S 或 R 看成是总体协方差矩阵 Σ 或总体相关矩阵 ρ 的估计), 抽取较少的 m 个因子, 估计因子负荷 a_{ij} 及特殊方差 Φ_i , 从而建立因子模型, 这是因子分析首先要解决的问题。

估计参数 a_{ij} 和 Φ_i 的方法比较多, 计算都比较复杂, 必须借助于电子计算机。此处只简单介绍两种较常用方法的主要步骤。原理及计算过程尽量省略。其次介绍 SPSS 软件提供的几种方法及相应的子命令。

一、主成分分解

设 Σ 的特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, 对应的标准化正交特征向量为 $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ 。设由列向量 $\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2, \dots, \sqrt{\lambda_p} \mathbf{e}_p$ 构成的矩阵用 $\mathbf{B}$ 表示, 即

$$\mathbf{B} = (\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2, \dots, \sqrt{\lambda_p} \mathbf{e}_p) \quad (18-157)$$

那么可以证明, Σ 的分解式为

$$\Sigma = \mathbf{B}\mathbf{B}' + \mathbf{0} = \mathbf{B}\mathbf{B}' \quad (18-158)$$

与因子模型的协方差结构(18-151)相比较, 我们看到(18-158)式表示了一个精确的可行的因子分解式, 实际上这就是主成分分析法的根据。但是它仍用 p 个因子来解释协方差结构, 而我们希望只用 m 个因子 ($m < p$)。这就启发我们想到能否去掉 $\mathbf{B}$ 的最后几列 $\sqrt{\lambda_{m+1}} \mathbf{e}_{m+1}, \sqrt{\lambda_{m+2}} \mathbf{e}_{m+2}, \dots, \sqrt{\lambda_p} \mathbf{e}_p$, 当

最后 $p-m$ 个特征值 $\lambda_{m+1}, \lambda_{m+2}, \dots, \lambda_p$ 比较小时, 我们可以得到如下的近似式

$$\Sigma = \mathbf{B}\mathbf{B}' = (\mathbf{A} \quad \mathbf{C}) \begin{pmatrix} \mathbf{A}' \\ \mathbf{C}' \end{pmatrix} = \mathbf{A}\mathbf{A}' + \mathbf{C}\mathbf{C}' \approx \mathbf{A}\mathbf{A}' + \Phi \quad (18-159)$$

其中, $\mathbf{A} = (\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2, \dots, \sqrt{\lambda_p} \mathbf{e}_p)$, $\mathbf{C} = (\sqrt{\lambda_{m+1}} \mathbf{e}_{m+1}, \sqrt{\lambda_{m+2}} \mathbf{e}_{m+2}, \dots, \sqrt{\lambda_p} \mathbf{e}_p)$,

$\Phi = \text{diag}(\Phi_1, \Phi_2, \dots, \Phi_p)$, $\Phi_i = \sigma_{ii} - \sum_{j=1}^m a_{ij}^2$ 。事实上就是忽略了 CC' 中的非

对角元素。

由此我们得到对 A 和 Φ 的一个解，叫做因子模型的主成分分解：

设样本协方差矩阵 S （或样本相关矩阵 R ）的特征值仍用 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 表示，对应的标准化正交特征向量为 $e_1, e_2, \dots, e_p$ ，则 S 的主成分因子分析的负荷矩阵可用

$$A = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \dots, \sqrt{\lambda_m} e_m) \quad (18-160)$$

（ $m < p$ ）来估计，特殊方差用 $S - AA'$ 的对角元素估计，即

$$\Phi = \text{diag}(\Phi_1, \Phi_2, \dots, \Phi_p) \quad (18-161)$$

其中 $\Phi_i = S_{ii} - \sum_{j=1}^m a_{ij}^2, (i=1, 2, \dots, p)$ 。

主成分分解中的因子负荷矩阵的各列实际上就分别是前 m 个主成分的系数的 $\sqrt{\lambda_i}$ 倍。

在实际应用中有一个如何确定公共因子个数 m 的问题。根据以上的解，我们知道 S 和 $AA' + \Phi$ 的对角元素是相等的，但非对角元素却不相等。如果取 m 个因子后使得残差矩阵

$$S - (AA' + \Phi) \quad (18-162)$$

的元素绝对值都很小（对角元素为 0），那么就可以认为含 m 个因子的模型是合适的。可以证明，残差矩阵的元素的平方和不会超过后面 $p-m$ 个特征值的平方和。实际上常常以这种平方和来估计因子模型的误差。

还有一种实用方法是根据因子的累计方差贡献率

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_m}{S_{11} + S_{22} + \dots + S_{pp}} \quad (18-163)$$

来确定因子数。如果达不到所要求的比例，则可以逐步增加因子个数，直到达到合适的比例为止。

也有时选择 $\lambda > 1$ 的个数为因子个数 m 。不过不管采用什么方法，原则上是贡献率要适当，因子个数也要尽可能少。

[例 18-33]假定[例 18-32]中 6 项考试的成绩 $X_1 \sim X_6$ 是对某个学生样本的数据，并假定样本相关矩阵 R 就等于

$$\mathbf{R} = \begin{pmatrix} 1.00 & & & & & \\ 0.24 & 1.00 & & & & \\ 0.28 & 0.42 & 1.00 & & & \\ 0.20 & 0.30 & 0.35 & 1.00 & & \\ 0.24 & 0.36 & 0.42 & 0.78 & 1.00 & \\ 0.28 & 0.42 & 0.49 & 0.75 & 0.72 & 1.00 \end{pmatrix}$$

它的前两个特征值和对应的标准化正交向量为

$$\lambda_1 = 2.756, \mathbf{e}_1' = (0.164, 0.246, 0.287, 0.558, 0.511, 0.508)$$

$$\lambda_2 = 0.604, \mathbf{e}_2' = (0.377, 0.565, 0.660, -0.230, 0.040, 0.221)$$

求因子分析的主成分解。

利用(18-160), 求出 $m=2$ 个公共因子的因子负荷矩阵为

$$\mathbf{A} = (\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2) = \begin{pmatrix} 0.272 & 0.293 \\ 0.409 & 0.439 \\ 0.477 & 0.513 \\ 0.926 & -0.179 \\ 0.848 & 0.031 \\ 0.843 & 0.172 \end{pmatrix}$$

$$\Phi = \text{diag}(\Phi_1, \Phi_2, \dots, \Phi_6) = \text{diag}(0.84, 0.64, 0.51, 0.11, 0.28, 0.26)$$

因子分析模型为

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \boldsymbol{\varepsilon}$$

其协方差结构为

$$\Sigma = \mathbf{A}\mathbf{A}' + \Phi$$

其残差矩阵为

$$\mathbf{R} - (\mathbf{A}\mathbf{A}' + \Phi) = \begin{pmatrix} 0 & & & & & \\ 0.00013 & 0 & & & & \\ -0.00005 & -0.00030 & 0 & & & \\ 0.00058 & -0.00015 & 0.00013 & 0 & & \\ 0.00026 & -0.00044 & -0.00040 & 0.00030 & 0 & \\ 0.00276 & -0.00030 & 0.00395 & 0.00017 & -0.00020 & 0 \end{pmatrix}$$

可以看到残差矩阵中的元素都很小, 说明由估计的负荷矩阵 $\mathbf{A}$ 和特殊方差矩阵 Φ 所再生的相关矩阵 $\mathbf{A}\mathbf{A}' + \Phi$ 与实际的相关矩阵 $\mathbf{R}$ 是十分接近的。

一般将因子分析的解, 包括负荷、共同度、累计方差贡献等, 都总结在一张表中, 如[例 18-32]的表 18-88 所示。

[例 18-34]表 18-89给出的数据是在洛杉矶十二个标准大都市居民统计地区中进行人口调查获得的。它有五个社会经济变量，分别是人口总数 (X_1)、居民的教育程度或中等教育的年数 (X_2)、佣人总数 (X_3)、各种服务行业的人数 (X_4) 和中等的房价 (X_5)。

表 18-89 五个社会经济变量

编号	X_1	X_2	X_3	X_4	X_5
1	5700	12.8	2500	270	25000
2	1000	10.9	600	10	10000
3	3400	8.8	100	10	9000
4	3800	13.6	1700	140	25000
5	4000	12.8	1600	140	25000
6	8200	8.3	2600	60	12000
7	1200	11.4	400	10	16000
8	9100	11.5	3300	60	14000
9	9900	12.5	3400	180	18000
10	9600	13.7	3600	390	25000
11	9600	9.6	3300	80	12000
12	9400	11.4	4000	100	13000

利用 SPSS 进行因子分析的步骤如下：

(1) 选择[Analyze]=>[Data Reduction]=>[Factor]，显示的[Factor Analysis]主对话框如图 18-119所示。

(2) 把 X_1 、 X_2 、 X_3 、 X_4 、 X_5 选入[Variables]列表框。

(3) 单击[Descriptives]显示如图 18-120所示的对话框，在[Correlation Matrix]选择中选择[Coefficients]以计算变量之间的相关系数。

(4) 其余选项采用默认设置。

结果输出如下：

数据转引自：王学民，《应用多元分析》，上海财经大学出版社，1999 年。

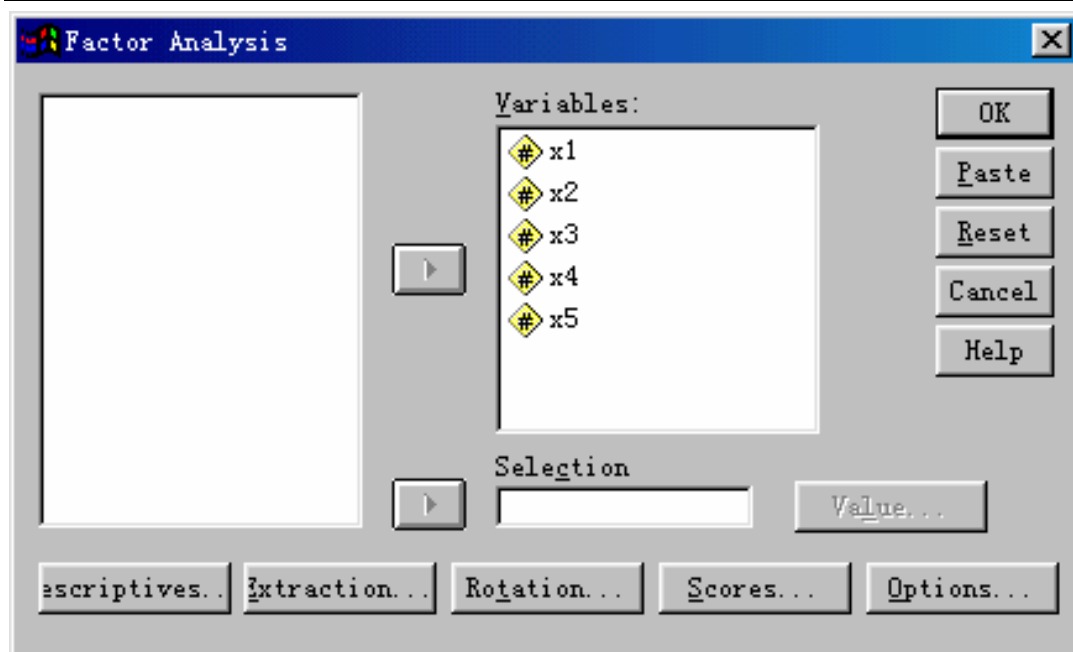


图 18-119 因子分析主对话框

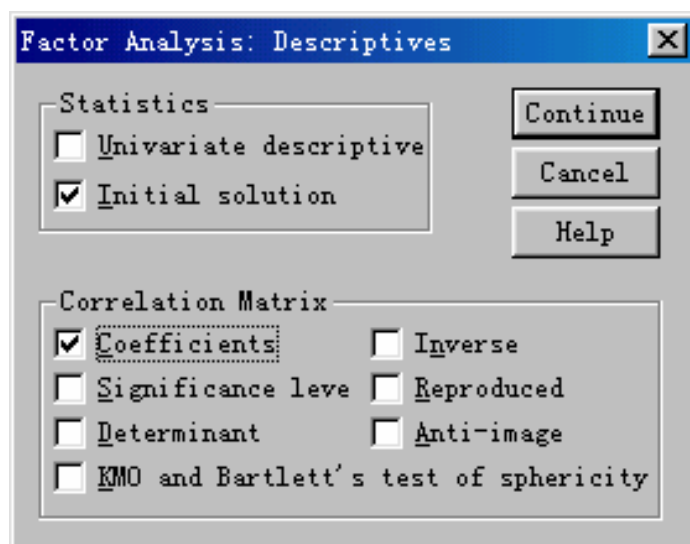


图18-120 因子分析描述统计量对话框Correlation Matrix

Correlation Matrix (相关系数矩阵)

		X1	X2	X3	X4	X5
Correlation	X1	1.000	.010	.972	.439	.022
	X2	.010	1.000	.154	.691	.863
	X3	.972	.154	1.000	.515	.122
	X4	.439	.691	.515	1.000	.778
	X5	.022	.863	.122	.778	1.000

Communalities (共同度)

	Initial	Extraction
X1	1.000	.988
X2	1.000	.885
X3	1.000	.979
X4	1.000	.880
X5	1.000	.938

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.873	57.466	57.466	2.873	57.466	57.466
2	1.797	35.933	93.399	1.797	35.933	93.399
3	.215	4.297	97.696			
4	9.993E-02	1.999	99.695			
5	1.526E-02	.305	100.000			

Extraction Method: Principal Component Analysis. (主成分法)

Component Matrix (主成分)

	Component	
	1	2
X1	.581	.806
X2	.767	-.545
X3	.672	.726
X4	.932	-.104
X5	.791	-.558

Extraction Method: Principal Component Analysis. (主成分法)

a 2 components extracted. (抽取2个主成分)

从结果可以看出(1)五个变量在第一个因子上都具有大的正负荷,尤其是X₄的负荷特别大。在第二个因子上变量X₁和X₃都有较大的正负荷,X₂和X₅都有较大的负负荷,X₁、X₃和X₂、X₅形成了鲜明的对照,而在X₄上的负荷非常小。(2)两个因子对所有变量的共同度都很大,在0.880到0.988之间。

二、主因子解

在主成分解中，我们从 $\mathbf{R} = \mathbf{B}\mathbf{B}'$ 出发求得 $\mathbf{B}$ ，然后去掉其最后几列来估计 $\mathbf{A}$ ，再由 $\mathbf{R}$ 和 $\mathbf{A}$ 去估计 Φ 。从共同度的观点考虑，这实际上就是把共同度的初始估计当作 $h_i^2 = 1$ （因为 $\mathbf{R}$ 的对角元素为 1），特殊方差的初始估计为 $\Phi_i = 1 - h_i^2$ 来进行的。

主因子解是先给出特殊方差 Φ_i 的一个初始估计 Φ_i^* ，得到共同度的初始估计 $h_i^{*2} = 1 - \Phi_i^*$ ，将 $\mathbf{R} = (r_{ij})$ 的对角线上的元素 1 都换成 h_i^{*2} ，其它元素不变，从而得到所谓约化相关矩阵 $\mathbf{R}^*$ ，即

$$\mathbf{R}^* = \begin{pmatrix} h_1^{*2} & & & & \\ r_{21} & h_2^{*2} & & & \\ \vdots & \vdots & \ddots & & \\ r_{p-1,1} & r_{p-1,2} & \cdots & h_{p-1}^{*2} & \\ r_{p1} & r_{p2} & \cdots & r_{p,p-1} & h_p^{*2} \end{pmatrix}$$

然后求出 $\mathbf{R}^*$ 的前 m 个特征值 $\lambda_1^*, \lambda_2^*, \dots, \lambda_p^*$ 对应的标准化正交特征向量 $\mathbf{e}_1^*$ 、

$\mathbf{e}_2^*, \dots, \mathbf{e}_m^*$ ，将 $(\sqrt{\lambda_1^*} \mathbf{e}_1^*, \sqrt{\lambda_2^*} \mathbf{e}_2^*, \dots, \sqrt{\lambda_p^*} \mathbf{e}_p^*)$ 作为负荷矩阵 $\mathbf{A}$ 的估计，再计算特殊方差矩阵的估计为 $\Phi^* = \mathbf{R} - \mathbf{A}^* \mathbf{A}'$ 。

显然，如果特殊方差 Φ_i 的初始估计为 0，那么共同度的初始估计为 1，这样，约化相关矩阵 $\mathbf{R}^*$ 与原相关矩阵 $\mathbf{R}$ 是相同的。因此这时主因子解与生成份解是一致的。

一般情况下 Φ_i 的初始估计可以取

$$\Phi_i^* = \frac{1}{r^{ii}} \quad (18-164)$$

其中 r^{ii} 是 $\mathbf{R}^{-1} = (r^{ii})$ 的对角线上的第 i 个元素。因此，共同度的估计为

$$h_i^{*2} = 1 - \Phi_i^* = 1 - \frac{1}{r^{ii}} \quad (18-165)$$

值得注意的是,用小于 1 的 h_i^{*2} 去代替 $\mathbf{R}$ 的对角元素后,得到的 $\mathbf{R}^*$ 的某些特征值可能会是负数。这时候,正的特征值之和将超过总共同度(即 $\mathbf{R}^*$ 的积),因为全部特征值之和就是总共同度。一种常用用的处理办法就是,逐步抽取因子并计算对应的特征值之和,直到特征值之和接近总共同度为止。

如果因子分析是从样本协方差矩阵 $\mathbf{S}$ 出发进行的,那么特殊方差的初始估计可以取作 $1/S^{ii}$, 其中的 S^{ii} 是 $\mathbf{S}^{-1} = (S^{ij})$ 的对角线上的第 i 个元素。

还有一种直接估计共同度的 h_i^{*2} 的初始值的方法,就是利用 X_i 及对其它 $p-1$ 个变量的回归的判定系数即复相关系数的平方来估计。我们在学习多元线性回归分析时已经知道, R_i^2 表示 X_i 的总离差平方和中可以用其它 $p-1$ 个变量的回归来解释的比例。可以证明, R_i^2 是对应共同度的下限,因此用它估计 h_i^{*2} 是偏于保守的。由于 R_i^2 比较容易计算,因此 R_i^2 比(18-165)式更常用。事实上可以证明它们是等价的,即 $R_i^2 = 1 - 1/r^{ii}$ 。

[例 18-35]在[例 18-34]中采用主因子法。SPSS 的输出结果如下:

Communalities (共同度)

	Initial	Extraction
X1	.969	.978
X2	.822	.818
X3	.969	.972
X4	.786	.798
X5	.847	.885

Extraction Method: Principal Axis Factoring.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of	Cumulativ	Total	% of	Cumulativ

		Variance	e %		Variance	e %
1	2.873	57.466	57.466	2.734	54.686	54.686
2	1.797	35.933	93.399	1.716	34.321	89.007
3	.215	4.297	97.696			
4	9.993E-02	1.999	99.695			
5	1.526E-02	.305	100.000			

Extraction Method: Principal Axis Factoring.

Factor Matrix

	Factor	
	1	2
X1	.625	.766
X2	.714	-.555
X3	.714	.679
X4	.879	-.158
X5	.742	-.578

Extraction Method: Principal Axis Factoring.

该结果与[例 18-34]的结果是类似的,在第一个公共因子上, X_4 有最大的正负荷,而 X_1 有最小的正负荷。在第二个公共因子上, X_1 和 X_3 有大的正负荷,而 X_2 和 X_5 有大的负负荷, X_4 有小的负负荷。所有的共同度估计都很接近于初始的共同度。

三、其它解法

除了以上两种因子分析法外,常用的还有最大似然法、 α 因子分析法、未加权的最小二乘法、一般的加权最小二乘法、映象因子分析法、最小残差法、典型最大似然法等等。根据一些统计学家的研究或蒙特卡罗模拟考察的结果,认为如果样本含量很大,变量数也很大(大于 30),并且所有变量都没有低共同度(不低于 0.40)的情况下,那么所有的不同的抽取因子方法最终都将得到大致相同的结果(对因子的解释相类似)。如果样本含量相当大(超过 1500),那么最大似然法给出的因子负荷估计是更精确的。

一般情况下不同方法给出的结果是不相同的。对一些“成问题”的数据(例如样本含量小,或变量数小的情况),最近一些学者的研究认为采用 α 因子分析法和映象因子分析法比主成分分析法更好。

下面将 SPSS 提供的七种抽取因子的方法及相应的选项名称总结在表

18-90中。

表 18-90 SPSS 提供的抽取因子方法

方法	选项名称	备注
主成分法	Principal components	默认项
未加权最小二乘法	Unweighted least squares	
广义最小二乘法	Generalized least squares	
最大似然法	Maximum Likelihood	
主因子法	Principal Axis factoring	
α 因子提取法	Alpha factoring	
映象因子提取法	Image factoring	

第三节 因子旋转

在实际应用中，我们希望对因子分析解作出合理的解释。也就是要知道各个公共因子的意义，到底代表了什么。为此就要考察各个变量 $X_1, X_2, \dots, X_p$ 在某个因子上的负荷，负荷绝对值大的变量显然与该因子的联系就更密切。可是如果因子负荷的大小相差不大，对因子的解释可能就有困难。为此我们想到通过旋转坐标轴，使因子负荷在新的坐标系中能按列向 0 或 1 两极分化，以便得到一个更简单的易于解释的结构。其原理有些象调整显微镜的焦距，以便更清楚地观察物体。

由线性代数知识我们知道乘以一个正交矩阵 T 就相当于作了一个正交变换或因子旋转。在第一节中我们已经看到旋转前和旋转后的负荷矩阵及因子轴分别为

A 和 $A^* = AT$

F 和 $F^* = T'F$

都产生相同的协方差矩阵

$$\Sigma = AA' + \Phi = A^*A^{*'} + \Phi$$

因此我们的任务就是找到适当的变换矩阵 T ，使旋转后的因子负荷阵尽可能具有简单结构，即

- (1) 每一列上的负荷大部分应是很小的尽可能接近 0 的值；
- (2) 每一行中只有少量的最好是只有一个较大的负荷值；
- (3) 每两列中大负荷和小负荷的排列模式应该不同。

先看看 $m=2$ 的情形。这时可用直观的图解法。

[例 18-36]重新考虑[例 18-32]的主成分解。因子 F_1 与表示数字能力的变量 X_4 、 X_5 、 X_6 关系比较密切，似乎这是一个数学能力因子，不过 X_2 和 X_3 在 F_1 上也有较大的负荷，因此对因子 F_1 的解释还是有些问题的；同理 F_2 也可以勉强地解释为是语言能力因子。

图为因子负荷对的散点图，每个点代表一个变量，它在 (F_1, F_2) 坐标系中的坐标就是该点在这两个因子上的负荷。从图中可以看到，如果将坐标轴按顺时针方向旋转一个角度 θ ，让所有的点都落在第一象限内。那么这些变量就比较明显地分成了两类，一类在 F_1^* 上有较大的负荷（变量 X_4 ，

X_5 ， X_6 ），在 F_2^* 上负荷较小或中等；而另一类在 F_2^* 上有较大的负荷（变

量 X_1 ， X_2 ， X_3 ），在 F_1^* 上的负荷较小或中等。我们可以进一步求出旋转

后的负荷阵 A^* 。用量角器量得 $\theta \approx 23^\circ 5'$ ，根据线性代数的知识，变换矩阵

$$T = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} = \begin{pmatrix} 0.920 & 0.392 \\ -0.392 & 0.920 \end{pmatrix}$$

因此，旋转后的负荷矩阵

$$A^* = AT = \begin{pmatrix} 0.272 & 0.293 \\ 0.409 & 0.439 \\ 0.477 & 0.513 \\ 0.926 & -0.179 \\ 0.848 & 0.031 \\ 0.843 & 0.172 \end{pmatrix} \begin{pmatrix} 0.920 & 0.392 \\ -0.392 & 0.920 \end{pmatrix} = \begin{pmatrix} 0.135 & 0.376 \\ 0.204 & 0.564 \\ 0.238 & 0.659 \\ 0.922 & 0.198 \\ 0.768 & 0.361 \\ 0.708 & 0.489 \end{pmatrix}$$

由于旋转后每个变量的共同度保持不变，因此特殊因子方差也是不变的。将表 18-88 的因子分析解和本例得到的旋转因子解列于表 18-91 中对比。

表 18-91 因子负荷矩阵

变量		因子负荷		因子负荷		共同度	特殊方差
		F_1	F_1	F_2	F_2		
语言考试	X_1	0.072	0.293	0.135	<u>0.376</u>	0.16	0.84
	X_2	0.409	0.439	0.204	<u>0.564</u>	0.36	0.64
	X_3	0.477	0.513	0.238	<u>0.659</u>	0.49	0.51
数字考试	X_4	0.926	-0.179	<u>0.922</u>	0.198	0.89	0.11
	X_5	0.848	0.031	<u>0.768</u>	0.361	0.72	0.28
	X_6	0.843	0.172	<u>0.708</u>	0.489	0.74	0.26

方差贡献率	45.9%	10.1%	34.3%	21.7%	56%	44%
累计贡献率	45.9%	56%	34.3%	56.0%		

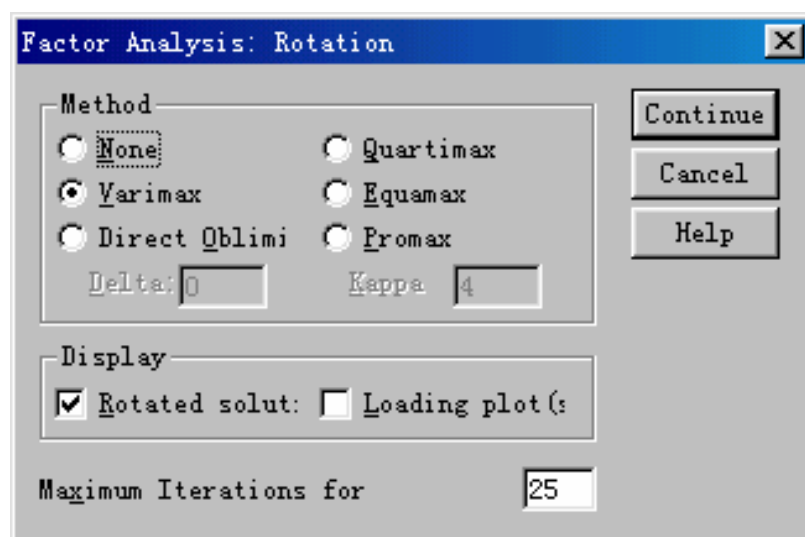
对于 $m=2$ 的情形，从图形上先对变量分类，然后就可以直观地确定新的公共因子。可是对于 $m>2$ 的情形，图解法就不再实用，需要采用解析的方法。

最常用的因子旋转方法是所谓“方差最大正交旋转”，它的思想是选择适当的正交变换矩阵 T ，使得各个因子负荷（除以共同度之后）平方的方差的总和达到最大。其效果是使各个因子对应的负荷量向 0 和 1 的两极分化，以达到负荷平方的方差最大。将负荷平方是为了消除符号的影响，除以共同度是为了消除各变量对公共因子依赖程度不同的影响。具体计算公式比较复杂，这里不再给出。在 SPSS 中会按要求进行因子旋转计算的。

方差最大的准则可以应用于从各种解法（主成分法、主因子法等）得到的初始因子负荷矩阵，不过结果一般是不相同的。但是经过正交旋转后，得到的因子一般都比初始因子容易解释。

[例 18-37]对[例 18-34]的主成分法($m=2$)用最大方差旋转法求解负荷矩阵的 SPSS 操作步骤及结果如下：

选择完估计方法后，单击[Factor Analysis]主对话框中的[Rotation]，在出现的旋转方法对话框中选择[Varimax]后单击[Continue]返回主对话框并单击主对话框中的[OK]按钮。



输出结果如下：

Component Matrix (主成分矩阵)

	Component (主成分)	
	1	2
X1	.581	.806
X2	.767	-.545
X3	.672	.726
X4	.932	-.104
X5	.791	-.558

Extraction Method: Principal Component Analysis.

Rotated Component Matrix (旋转后的主成分矩阵)

	Component (主成分)	
	1	2
X1	.016	.994
X2	.941	-.009
X3	.137	.980
X4	.825	.447
X5	.968	-.006

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 3 iterations.

从以上结果可以看出，在第一个公共因子上，X2、X4 和 X5 有大的正负荷，而 X1 和 X3 的负荷很小，这个因子可解释为福利条件因子。在第二个公共因子上，X1 和 X3 有大的正负荷，X4 有较小的正负荷，而 X2 和 X5 只有很小的负荷，这个因子可解释为人口因子。

上例中采用的是方差最大正交旋转。这是一种最常用的旋转方法。SPSS 还提供了其它一些正交和斜交方法，请见第五节。

第四节 因子得分

前面讨论的是因子分析的第一个也是主要的问题，即将 p 个观测变量 $X_1, \dots, X_p$ 表示成 m 个潜在变量（因子） $F_1, \dots, F_m$ 的线性组合，从而可以得到正交因子解（如果允许因子间有相关关系，得到的就是斜交因子解，本书对此没有讨论）。因子分析的第二个问题是将潜在变量（因子）表示为观

测变量的线性组合，也就是对公共因子的取值进行估计，计算各个样品的公共因子得分的问题。由此可以在公共因子的空间中，按照各个样品的因子得分值标出其对应的位置。因子得分可以通过多元回归分析的方法估计，也可以用 Bartlett、Anderson-Rubin 或 Thompson 等方法估计。这里仅简单介绍回归因子得分。

为了书写简便，我们将第 i 个样品对第 j 个因子得分的估计值

$$\hat{F}_{ij} = b_1 X_{i1} + b_2 X_{i2} + \cdots + b_p X_{ip} \quad (i=1,2,\cdots,n, \quad j=1,2,\cdots,m) \quad (18-166)$$

(其中 X_{ik} 表示第 i 个样品在第 k 个变量上的观测值， $b_1, b_2, \cdots, b_p$ 相当于回归系数，对应于第 j 个因子) 以矩阵的形式表示为

$$\begin{array}{ccc} \hat{\mathbf{F}} & = & \mathbf{XB} \\ (n \times m) & & (n \times p)(p \times m) \end{array} \quad (18-167)$$

其中 $\hat{\mathbf{F}}$ 是一个因子得分矩阵，表示 n 个样品分别在 m 个公共因子上得分的回归估计； $\mathbf{X}$ 是原始数据矩阵，每行代表一个样点在 p 个量上的观测值， $\mathbf{B}$ 则表示回归系数，每列对应于一个因子的 p 个回归系数。如果原始数据是标准化的数据。用矩阵 $\mathbf{Z}$ 表示，则有

$$\begin{array}{ccc} \hat{\mathbf{F}} & = & \mathbf{X}\boldsymbol{\beta} \\ (n \times m) & & (n \times p)(p \times m) \end{array} \quad (18-168)$$

其中 $\boldsymbol{\beta}$ 表示标准化的回归系数矩阵。可以证明它的最小二乘估计为

$$\begin{array}{ccc} \boldsymbol{\beta} & = & \mathbf{R}^{-1} \mathbf{A} \\ (p \times m) & & (p \times p)(p \times m) \end{array} \quad (18-169)$$

其中 $\mathbf{R}^{-1}$ 是变量 $X_1, X_2, \cdots, X_p$ 的样本相关矩阵， $\mathbf{A}$ 是因子负荷矩阵。因此因子得分的估计为

$$\begin{array}{ccc} \hat{\mathbf{F}} & = & \mathbf{X} \mathbf{R}^{-1} \mathbf{A} \\ (n \times m) & & (n \times p)(p \times p)(p \times m) \end{array} \quad (18-170)$$

因子得分实际上给出的是各个样品在公共因子上的投影值或坐标值。因此，以公共因子为坐标轴，在公共因子空间中，就可以按各样品的得分值标出其空间的相对位置。这样就可以进一步得到关于原始数据的结构方面的信息。下面的例子说明了因子得分在这方面的应用。

[例 18-38]在[例 18-37]基础上,计算洛杉矶标准大都市居民统计地区的因子得分。

在[例 18-37]操作基础上,在主对话框中单击[Scores]。显示如下所示的对话框。选择[Save as variables]和[Regression]。SPSS 把用回归分析方法计算得到的因子得分分别在数据文件中存为一新变量(默认为 fac1_1 和 fac2_2)。其得分如表 18-92所示。

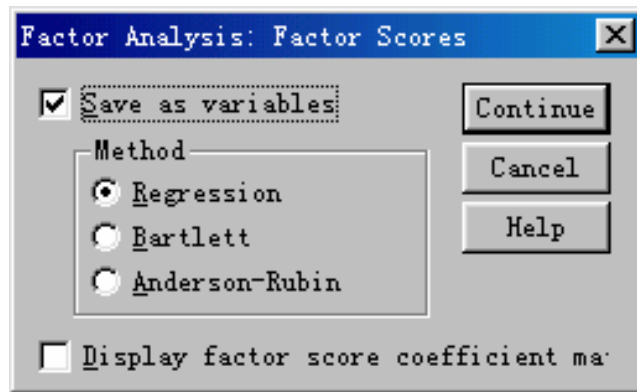


表 18-92 因子得分

编号	$\hat{F}_1$	$\hat{F}_2$
1	1.20297	-.03080
2	-.65918	-1.38351
3	-1.25937	-.76740
4	1.11491	-.79697
5	.93710	-.76320
6	-1.22514	.54857
7	-.16819	-1.54932
8	-.44127	.73444
9	.32032	.91306
10	1.57628	1.02554
11	-.94628	.96184
12	-.45215	1.10776

第五节 案例：研究生院规模的因子分析

我国自 1984 年试办研究生院，已有三十多所试办研究生院的高校正式建立研究生院。这些研究生院通过十多年的实践，在扩大研究生办学规模，保证和提高研究生培养质量和学位授予质量，开展学科建设等方面进行了积极的探索，对促进我国研究生教育事业的发展，起了重要的作用。为总结经验，也为研究生教育的总体布局和新建研究生院提供决策依据，我国于 1995 年首次开展了对全国 33 所研究生院的评估工作。这次评估的数据处理涉及了多种数学方法，如加权和法、百分线性规划法、区域线性规划法、平均值规划法等。国外也经常评估研究生院，各国评估研究生院的数据处理方法各异，有用简单数据处理方法，也有用多元统计分析方法。本文运用因子分析和聚类分析等多元统计方法，对我国 18 所设立研究生院的理工科大学表征规模指标的数据进行分析、处理，试图揭示影响理工科大学研究生院办学规模的综合因素，并对评估研究生院办学规模的数据处理方法进行了初步探讨。

一、研究方法

本案例以全国 18 所理工科大学的研究生院为研究对象，选取象征研究生院办学规模的 1994 年部分统计数据（如表 18-93 所示），进行因子分析与聚类分析。

表 18-93 研究生院办学规模统计数据

单位	变量名							
	学位授权点		在校研究生		已授学位		重点学科 X ₇ 个	博士后流动站 X ₈ 个
	博士 X ₁ 个	硕士 X ₂ 个	博士 X ₃ 个	硕士 X ₄ 个	博士 X ₅ 个	硕士 X ₆ 个		
清华大学	64	107	1202	2492	941	7233	29	14
北航空航天大学	20	60	284	1215	231	3456	5	4
北京理工大学	20	57	278	1140	156	2721	4	3

参见徐惠娟、王惠文：《多元分析在大学研究生院规模研究中的应用》，《北京航空航天大学学报》，1998 年第 2 期。

北京科技大学	15	38	268	1005	235	2831	6	3
天津大学	24	75	399	1784	230	4336	6	4
大连理工大学	21	65	276	1415	292	3657	4	5
东北大学	23	50	330	1185	139	3266	2	3
哈尔滨工业大学	29	67	501	1586	342	4952	7	5
同济大学	20	62	271	1177	167	2632	4	4
上海交通大学	34	75	267	1755	278	4086	8	5
东南大学	21	62	302	1181	206	3004	4	1
浙江大学	35	84	658	1882	367	4598	9	10
中国科技大学	24	46	246	778	225	2433	4	5
华中理工大学	29	76	495	1737	299	4087	4	4
中国地质大学	13	20	215	472	201	1942	5	2
国防科技大学	14	40	212	776	85	2182	2	1
西安交通大学	33	65	690	1694	403	4812	11	8
西北工业大学	19	54	277	1102	207	3078	3	3

1、因子分析

因子分析的基本目的是用少数几个变量去描述多个变量间的协方差关系。其思路是将观测变量分类，将相关性较高即联系比较紧密的变量分在同一类中，每一类的变量实际上就代表了 1 个本质因子，从而可将原观测变量表示为新因子的线性组合。本案例运用因子分析的主要目的是简化观测系统，将已知的 8 个变量减少为几个新因子，以再现它们之间的内在联系。因子分析的结果不仅给出因子模型，还得出变量和因子间的相关系数，这些相关系数构成因子结构。1 个完全的因子解包括因子模型和因子结构两个方面，因子结构反映变量与因子间的相关关系，而因子模型则是以回归方程的形式将变量表为因子的线性组合。

本案例中的因子分析主要经过以下几个步骤：

- (1) 将原始数据进行标准化处理。
- (2) 计算所有变量的相关矩阵 R 。
- (3) 因子提取。这里采用主成分分析法，利用相关系数矩阵 R 进行因子提取。可通过研究公共因子在变量总方差中所占的累计百分数（一般为 85% 以上）确定所需要的公共因子数。
- (4) 因子旋转。因子分析的目的不仅是要找出主因子，更重要的是知道每个主因子的意义，为便于对主因子进行解释，一般须对因子载荷矩阵进行

旋转，以达到结构简化的目的。由于代表研究生院规模的各变量之间不可能是彼此无关的，故这里作了斜交旋转。

(5) 计算每一样本点的因子得分，用于诊断模型，并进行因子排序。在本案例中，因子得分还作为进一步聚类分析的原始数据。

2、聚类分析

聚类分析是将一批样品或变量，按其性质的亲疏程度进行分类。本案例中采用了系统聚类法，定义各样本点之间距离采用的是平方欧氏距离，定义各类之间距离用最短距离法。通过系统聚类，可以得到树状谱系图，谱系图能明确清晰地描述各个样本点在不同层次上聚合分类的情况。

3、绘制散点图

根据因子得分，绘制散点图，在散点图上能清楚地观察不同的类别及其性质。

二、计算结果与分析

从相关矩阵 R 得知，各变量之间相关性很强，因此，有可能存在一些变量共享因子。同时，计算表明，几乎总方差的 95.3% 可由第 1 和第 2 因子解释，由此可选择两个主因子。表 18-94 为经斜交旋转后的因子模型矩阵和因子结构矩阵。

表 18-94 斜主因子模型矩阵和结构矩阵

变量	模型矩阵		结构矩阵	
	F_1	F_2	F_1	F_2
X_7	1.08998	-0.14690	0.97663	0.69407
X_5	1.00662	-0.02617	0.98603	0.75018
X_3	0.85073	0.14781	0.96477	0.80419
X_8	0.82976	0.14305	0.94012	0.78324
X_1	0.70331	0.32548	0.95443	0.86812
X_2	-0.02906	1.00664	0.74761	0.98421
X_4	0.07426	0.93417	0.79502	0.99147
X_6	0.45550	0.57878	0.90205	0.93022

注： F_1 为第 1 斜因子， F_2 为第 2 斜因子。

分析表 18-94 看出，经斜交旋转后，第 1 主因子主要由与博士规模有

关的变量即博士学位授权点数、在校博士生数、已授博士学位数、博士后流动站数及重点学科表征，而重点学科是从博士学位授权点遴选产生，因此，因子 1 可解释为博士规模因子。第 2 主因子主要由与硕士规模有关的变量即硕士学位授权点数、在校硕士生数、已授硕士学位数表征，因此，因子 2 可解释为硕士规模因子。表 3 为斜交因子得分和按因子 1 和因子 2 对 18 所院校的博士规模和硕士规模的排序表。

表 18-95 斜交因子得分与因子排序表

序号	单 位	F_1	F_2	P	M
1	清华大学	3.52138	2.28012	1	1
2	北航空航天大学	-0.32322	-0.17917	10	9
3	北京理工大学	-0.55753	-0.39170	16	13
4	北京科技大学	-0.35486	-1.02746	11	16
5	天津大学	-0.18163	0.80076	7	4
6	大连理工大学	-0.24202	0.14284	8	8
7	东北大学	-0.53641	-0.34008	15	12
8	哈尔滨工业大学	0.29318	0.55366	4	6
9	同济大学	-0.51433	-0.26473	17	10
10	上海交通大学	.04071	0.75449	5	5
11	东南大学	-0.58254	-0.21307	17	10
12	浙江大学	0.81128	1.14650	3	2
13	中国科技大学	-0.27316	-0.97400	9	15
14	华中理工大学	-0.05377	0.80448	6	3
15	中国地质大学	-0.52032	-2.01711	14	18
16	国防科技大学	-0.93288	-1.16212	18	17
17	西安交通大学	0.92023	0.52181	2	7
18	西北工业大学	-0.51411	-0.43522	12	14

注： F_1 为斜因子 1 得分， F_2 为斜因子 2 得分，P 为博士规模大学排序，M 为硕士规模大学排序。

此外，本文通过对样本点进行聚类分析，得到谱系图（如图 18-121 所示），从中可看出，我国理工科大学 18 所研究生院的规模情况大致分为 4 类（见表 18-96）。4 类各项指标的平均值及总平均值见表 18-97。

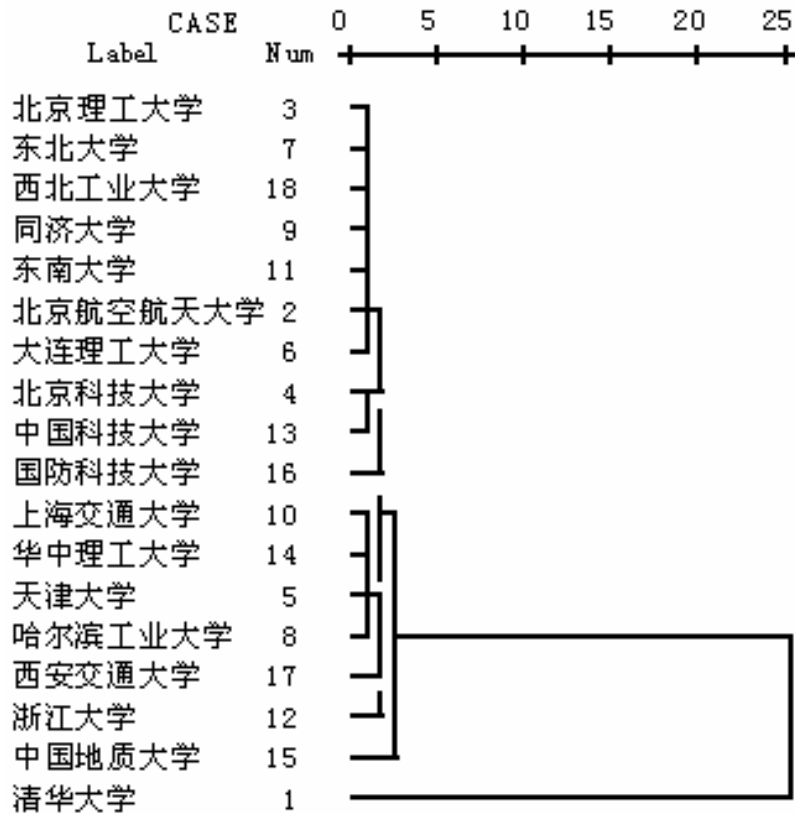


图 18-121 18 所大学研究生院的系统聚类图

图 18-122 为取第 1 斜因子和第 2 斜因子作因子得分散点图 (序号意义见表 3)。

分析表 18-97 及图 18-122 的 4 大类情况，可以看出：第 1 类为清华大学。在图 18-122 中，清华大学的博士规模因子与硕士规模因子的得分值远大于其它院校的平均水平。实际上，该校各项研究生规模指标均大于其它院校的同类指标，其学位授权点数、在校研究生数及已授学位人数等均列全国高校榜首。

第 2 类包含了浙江大学等 6 所大学，此类大学的各项研究生规模指标均在 18 研究生院总平均规模之上。这类大学大都建校历史悠久，学科门类较为齐全。

第 3 类包含北京航空航天大学等 10 所大学，此类大学研究生院各项平均规模指标略低于第 2 类大学，其中大部分大学有较强的行业特色，并有较强的与行业有关的特色学科，属多科性理工科大学。由图 18-122 可见，这一类中，东北大学等 5 所大学规模非常接近，几乎为重叠点。从表 18-95 可见，这 5 所大学博士和硕士规模排序也很接近。

第 4 类为地质大学。由图 18-122 可看出，该校的硕士规模因子得分

值较小，实际上，其各项硕士规模指标均小于第 3 类同类指标的最低值。这与该校学科门类较为单一，特别是与该校工学学位授权点数相对较少有关。然而，该校有较多的重点学科，其已授博士学位的人数高于第 3 类的平均值。

表 18-96 18 所理工科大学研究生院办学规模分类

类别	该类所包含的大学
1 类	清华大学
2 类	浙江大学、天津大学、哈尔滨工业大学、上海交通大学、华中理工大学、西安交通大学
3 类	北京航空航天大学、北京理工大学、北京科技大学、大连理工大学、东北大学、同济大学、东南大学、中国科技大学、国防科技大学、西北工业大学
4 类	中国地质大学

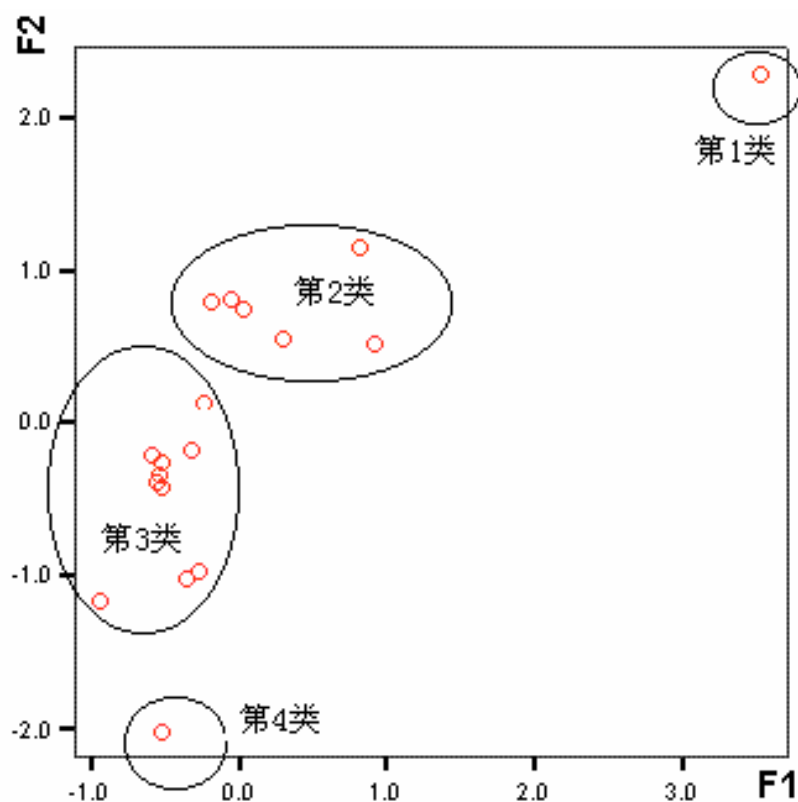
图 18-122 用博士规模因子得分 $\hat{F}_1$ 和硕士规模因子得分 $\hat{F}_2$ 绘制散点图

表 18-97 1994 年各类大学研究生院平均规模

类别	项 目							
	学位授权点		在校研究生		已授学位		重点学科	博士后流动站
	博士	硕士	博士	硕士	博士	硕士		
1	64.0	107.0	1202.0	2492.0	941.0	7233.0	29.0	14.0
2	30.7	73.7	501.7	1739.7	319.8	4478.5	7.5	6.0
3	19.7	53.4	274.4	1097.4	194.3	2926.0	5.0	2.0
4	13.0	20.0	215.0	472.0	201.0	1942.0	5.0	2.0
总平均值	25.4	61.3	398.4	1354.2	278.0	3628.1	6.5	4.7

三、结束语

利用因子分析将众多的变量综合为几个本质因子，并对样本进行聚类分析，这项工作在高等教育研究中很有意义。本节通过对我国 18 所理工科大学研究生院 1994 年的部分统计数据进行因子分析和聚类分析，将表征规模的各统计变量综合为博士规模因子和硕士规模因子两个综合变量；同时，通过对两个综合变量的排序，可大致看出 1994 年 18 所理工科大学研究生院培养博士生规模和培养硕士生规模的实力；通过聚类分析和绘制散点图，清楚地看出哪些大学研究生院的规模是属于同类大学。本案例得到了一些符合高等教育规律的结论，从而为科学评估我国理工科大学研究生院人才培养的规模以及评估研究生院整体条件和水平提供了一种新的途径。

第十五章 事件史分析

事件史分析(Event History Analysis)方法是近年来国际统计学界发展最为迅速的小样本分析技术。本章从事件史分析方法的源流,概念及方法等方面对此作了比较全面的介绍,并举例说明了事件史分析方法的应用。

第一节 事件史分析方法的源流

在社会科学的各个分支领域,有很多极为有趣的事件及其形成这些事件的原因。如:犯罪学家研究犯罪、逮捕、牺牲等事件;经济学家在工作 and 就业的研究中,注意的焦点则是工作变换、提升、解雇和退休等事件;政治学家则关注暴乱、革命、政府的和平更迭等事件;人口学家则关注出生、死亡、结婚、离婚和迁移等事件;医学家则关心病人访问医生的次数等等。

在上述的每个例子中,每个事件包含一些在特定时点上的定性的变化。因此,一般不能用“事件”来普通地描述一些定性变量变化的过程。因为在这些变量的变化之中,往往包含了一些尖锐的突变。

由于事件是按照时间变化来分类的,因此,研究和认识这些事件及其产生的原因的最好的方法是收集描述这些事件的数据。在其最简单的形式中,即当事件集是以个体的集合出现时,它们往往呈现一种纵向的记录。例如,一项人口调查中的结婚的日期记录就是一个典型的事件史数据集。而如果调查的目的是研究事件发生的原因,那么事件史也应当包括描述其变化的一些解释性变量的数据。

事件史是研究事件发生原因的基础,且具有两个特征:截断(censoring)和随时间变化的解释变量(explanatory variable)即:能产生利用诸如多元回归分析等统计分析方法来描述问题的随机变量。在过去的二十年里,一些具有创新性的方法与技术首先在一些特殊的事件史数据的研究得到了应用。因此,事件史方法并非是一种单一的方法,而是一些相关的不断完善的方法与技术的集合体,并在许多领域里由很多深奥和复杂的方法和技术发展起来的。这同时也是引起某些混乱的原因,因为有时事件史中的一些相同的概念经常以不同的方式来表达。因此,有必要对事件史分析方法发展的历史源流作一回顾。

事件史分析方法的源流主要有三个，其一是生物医学统计方法的生存分析 (survival analysis) 学派；其二是可靠性技术的“失效时间分析”(failure time analysis) 学派；其三是数理社会学中的马尔科夫(Markov) 学派。

从人口学中可以得知，生命表方法是最早出现的，也是最著名的、应用最广泛的分析方法。自十八世纪以来在社会学及生物医学等领域被广泛地采用。直到二十世纪五十年代末六十年代初逐渐被更为先进的方法——事件史分析方法所替代。在生物医学中，用来处理其实际问题的方法又可称为生存数据的分析方法，的确，大多数事件史方法的文献是用生存分析和生命分析的名称而写成的。例如，在实验室进行的动物实验中，通过让动物吃不同剂量、毒性不一的药物，实验人员可以观察到在不同的过程中动物是如何存活的。在这里，“事件”就是特指动物的死亡，“截断”的出现则是因为该实验通常是在所有动物死亡以前而结束。生物统计学家已经写了大量的文献说明如何分析这些数据的有效的方法，这些方法已经变成为研究与分析癌症病人存活的数据的主要工具。其中，由该领域中所发展起来的最有影响的回归方法部分似然估计方法(Cox, 1972)成为事件史分析方法中的重要的核心技术之一。

可靠性分析技术是工程上研究与分析机器和电子元件故障数据所发展起来的重要的方法。该领域里所发展起来失效时间分析技术，在过去的十几年里已经被有效的应用于事件史分析之中，并形成了新的流派。

另外，自二十世纪 60、70 年代以来，马尔科夫过程理论在社会科学的应用有了长足的进步和发展，这一方法的转折点是 Tuma(1976)将解释变量引入连续时间的马尔科夫模型，这可以说是沟通生物统计和工程学方法的一座桥梁。其中，用马尔科夫模型研究社会科学现象变化过程中不同状态的个体事件的分布构成了事件史分析中的重要方法。

第二节 事件史分析方法的内容概述

一、事件史分析中的主要概念

■ 重复发生的和不重复发生的事件

事件是事件史分析中最基本的概念，不重复发生的事件则是指只能发生一次的事件；重复发生的事件则是指可以发生两次以上的事件。

如：出生与死亡事件，就属于不重复发生的事件；而工作变换和结婚等在人的一生中可以出现多次的事件，则属于重复性事件。关于重复性事

件的模型往往要比不重复事件的模型更为复杂，因此，也更能引出许多复杂的统计问题。然而，掌握不重复事件的分析方法是理解和掌握重复事件模型方法的基础。

■ 单一种类事件和多种类事件

单一种类事件是指影响某一类事件发生的原因是相同的；多种类事件则是指影响某一类事件发生的原因是不同的。

在单一种类事件分析中，用几乎相同的分析方法可以方便的处理所有的事件。例如，在工作变换的研究中，要区分自愿终止工作和非自愿终止工作是非常困难的，也是不必要的。若不考虑这一区别，则可用单一种类事件史分析方法来分析这类事件的变化。然而，在诸如癌症等疾病治疗有效性的研究中，区分因癌症或其它病因导致的死亡显然是很重要的。为了适应不同类型的事件史的研究，类似于人口学中的多减量生命表一样，该领域中发展了一种被称为“完全风险”(completing risks)的方法。然而，由于多种类型事件的引入会导致更为复杂的统计方法的使用，而在某些实际问题中，要完全区分事件的类型有时是不可能的。因此，在实际问题的分析中，这些方法的应用会受到一些局限。

■ 截断(censored)和解释变量(explanatory variables)

截断是指所观察的个体在观察的时期内并没有发生变化但因其它原因而退出观察，或直到观察结束时仍未发生变化，但因研究停止而被中断的个体状态。一般来说，被截断的个体过多，则分析的结果会有较大的偏差。解释变量是指描述事件发生的原因和变化过程的相互独立的随机变量，这是运用多元统计方法建立事件史分析模型的基础。

■ 风险率(hazard rate)

风险率，也称风险概率，是事件史分析方法的核心概念。它描述了所观察的平均每一个体所发生该事件概率。在离散时间模型中，风险率是指每一个体将在某一特定的时点发生事件的概率，也同时表示了平均每一个体在该时刻所处的风险。在连续时间模型中，设 $P(t, t+s)$ 是每一个体在时间 $(t, t+s)$ 中发生事件的风险概率，显然，当 $s=1$ 时，即为离散时间的风险概率。而当 s 趋近于零时，定义连续时间的风险函数为 $h(t)$ ，则有：

$$h(t) = \lim_{s \rightarrow 0} P(t, t+s) / s \quad (15-1)$$

它描述了个体瞬时发生的风险概率，其可为正值或负值，也可为常数。（这一概念同生命表中的死亡力的概念相类似）

若设 $P(t)$ 为一个体在时刻 t 之后发生的风险概率，则 $P(t)$ 和 $h(t)$ 有：

下列关系（证明过程略）：

$$h(t) = -\frac{P'(t)}{P(t)} \quad (15-1)'$$

在事件史分析中认识风险是极为重要的，而且，风险既控制着个体发生事件的可能性，也控制着个体数随时间的变化。正因为如此，风险率是事件史分析模型中的核心变量。

二、事件史分析的研究方法

关于事件史分析的研究方法可分为两类，一类为离散型的，另一类为连续型的。

在这两类方法中，又分别可分为参数或非参数的方法。非参数方法是通过对事件随时间的分布作出可能的假设，参数方法则是用诸如指数分布、威布尔分布、Gomperz 分布等来假设事件之间的关系和事件随时间变化的关系。联系这两种方法的一个重要的桥梁是可以描述半参数或部分参数的 Cox 的比例风险模型（1972）。限于篇幅，下面将仅对单一类型和不重复事件史分析方法作一介绍，对于多类型和重复事件史分析模型和方法请参考有关文献。

（一）离散型事件史分析模型及计算方法

尽管事件是随时间连续变化的，但实际上，时间总是以离散的单位来测量的，当这些时间单位非常小时，通常可以将其作为连续型变量来处理。另一方面，当时间单位很大时，如以月，年，十年为单位时，则更适合用离散时间模型方法来分析，由于离散时间模型方法特别容易理解和应用，因此它可以作为事件史分析基本理论的导论。

下面要引入单一类型不重复事件的离散时间模型方法，所描述的方法也适用于其它许多的情况，而且也可以推广到不同类型事件的重复事件的分析之中。

首先要讨论风险率是如何依赖于解释变量而发生变化的。不妨设 $P(t)$ 为一个体在时刻 t 之后发生的风险概率。又假定有 n 个相互独立解释变量是 $X_1(t), X_2(t), \dots, X_n(t)$ 。作为粗略的估计，可以认为 $P(t)$ 是 $X_1(t), X_2(t), \dots, X_n(t)$ 的线性方程，即：

$$P(t) = a + b_1 X_1(t) + b_2 X_2(t) + \dots + b_n X_n(t) \quad (15-2)$$

其中， $t=1, 2, \dots, 5$ 。

然而，这里要注意的一个问题是，由于 $P(t)$ 是一个概率，因此，它不应大于 1 和小于零。但上式的右端显然可以是任意实数。因此，这一模型

是不可能用来进行预测的，且带来计算和解释的复杂性。但这一问题可以通过对 $P(t)$ 进行 logit 变换得以避免，即：

$$\log\left(\frac{P(t)}{1-P(t)}\right) = a + b_1 X_1(t) + b_2 X_2(t) + \cdots + b_n X_n(t) \quad (15-3)$$

当 $P(t)$ 在 0—1 之间变化时，方程的左边从负无穷大到正无穷大。尽管还有其它一些形式的变换，但 logit 变换是最方便的和熟悉的形式。回归系数 b_i ($i=1, \dots, n$) 则表示当 X_i ($i=1, \dots, n$) 分别变化一个单位时，logit 所变化的数量。

在上述离散模型中，通过让参数 a 随时间而变化，这样，可以得到描述离散型事件史分析方法的核心模型，即：

$$\log\left(\frac{P(t)}{1-P(t)}\right) = a(t) + b_1 X_1(t) + b_2 X_2(t) + \cdots + b_n X_n(t) \quad (15-4)$$

而关于参数 $a(t)$, b_1 , b_2 , $\dots$, b_n 的估计将通过极大似然估计的方法来进行(从略)。

(二) 连续型事件史分析模型及计算方法

虽然离散时间模型方法应用较广泛，但大多数事件史分析方法却是连续型的模型方法。仿前述离散型变量的分析可以得到描述连续型事件史分析的比例风险模型 (proportional hazards model) 为：

$$\log h(t) = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + u \quad (15-5)$$

这里 $h(t)$ 是连续时间风险率， x_i ($i=1, \dots, n$) 是相互独立的解释变量， $a, b_1, b_2, \dots, b_n, u$ 是一个独立于 x_i ($i=1, \dots, n$) 的随机项。

对于给定 u 以不同的分布，可以得到连续型风险率将适合诸如指数 (exponential) 分布、威布尔 (weibull) 分布、龚帕兹 (Gompertz) 分布、伽马 (gamma) 分布等的情形。

之所以将模型 (15-5) 称为比例风险模型是因为在该模型中在任意时点上的两个个体所处的风险概率之比 $h_i(t)/h_j(t) = c$ 仅依赖于解释变量，而与时间无关。

关于部分似然估计方法 (partial likelihood method) 数学细节见本章附录。这里要谈一些该方法需要注意的问题。部分似然估计方法实际上依

赖于所建立的似然函数中的所引用的数据的情况，这些数据在比例风险模型中可分为两类，第一类因素是决定回归系数 b_i ($i=1, \dots, n$) 的部分，第二类因素是决定回归系数 b_i ($i=1, \dots, n$) 和 $a(t)$ 的部分。特别是，这第一类因素仅依赖于事件发生的顺序，而不是事件发生的具体时间。

关于其结果的估计是渐近无偏的，并服从正态分布。但由于该方法忽视了发生事件的确切时间，所以其结果有时并非是完全有效的。但是这种无效性通常是很小的，以至于没有必要担心。

关于在比例风险模型下利用部分似然估计参数可以通过 SPSS 等进行。

第三节 事件史案例分析

让我们从一个实际例子中考察一下离散型事件史分析方法的应用。这是一个包含 200 个男性生物化学家的样本，他们分别在二十世纪 50 年代末期和 60 年代初期获得博士学位，并且在这期间的某一时刻毕业后在大学里作助理教授。从他们作助理教授开始，他们被观察的最大时间是 5 年。如表 15-1 所示。

表 15-1 200 名生物化学家工作变换的分布

年	工作变换人数	风险个数	风险率估计
1	11	200	0.055
2	25	189	0.132
3	10	164	0.061
4	13	154	0.084
5	12	141	0.085
>5	129		
合计	200	848	

由表 15-1 可以看出，这些事件是以离散时间来记录的，以年为单位。要想区别这些事件是由于他们自愿的还是不自愿的发生是不可能的。因此，我们把这些事件作为单一类(同类)事件来处理。表 15-1 表示了生物化学家们在 5 年中的每一年中变换工作的次数。在总数为 200 个个体的样本中，有 129 个生物化学家在所观察的时期里没有变换工作，因此，可以将这些个体看作为截断事件。

现在，让我们来估计反映一年期工作变换的风险概率和基于选择的 5 个解释变量的回归模型。其中的两个变量描述了初始就业机构的特征，即

为：机构的威信(x_1)和获得的联邦基金(x_2)，根据对它们的计算，可以认为它们都是常量。其它三个变量则按年描述了生物化学家个体的情况：出版文章的累积数($x_3(t)$)；被其他学者所引用的学术著作数($x_4(t)$)；学术级别($x_5(t)$)若是副教授则定为 1，若是助理教授则定为 0。(虽然在观察时期的起始时刻，他们都是助理教授，但有些人很快就得到了提升)。这样，我们就不仅有了常量，而且有了时变的解释变量。

对于这 200 个生物化学家的样本来说，他们在第一年里变换工作的风险就是整个样本在那一年里的风险集。在那一年里，仅有 11 人变换了工作，则这 11 人在第二年里将不在有风险率。(也许他们中的人在第二年仍会变换工作，但由于其概率很小，这里不再考虑这种情形)因此，在每一年结束时，由于事件的总量在减少，因此，风险集合的数量也在减少。

例如，在表 15-1 中，可以看到风险集合的数量从第一年的 200 下降到第 5 年的 141。在本例中，第一次变换工作的概率对那些还没有变换工作的人构成了风险。

若假设风险率随年而变化，但每一年的每一个体的风险率均相同，则可以容易估计出其风险率：例如，在第二年里，在风险集合为 189 个个体中，有 25 个个体变换了工作。则其风险率为 $25/189 = 0.132$ 。其它年度风险率的估计可以类似的得到，如表 15-1 所示。从表 15-1 看到，并没有风险(关于工作变换)明显的增加或减少的变化趋势。但应注意的是，因为风险集合的个数在严格减少，即使是工作变换的个数在减少，其风险率也可能在增加。例如，虽然第一年的工作变换个数要小于第一年，但第三年的风险率要大于第一年的风险率。

下面将根据离散模型(15-6)来估计参数 b_i ($i=1, \dots, 5$)和 $a(t)$ 的 5 个值。

$$\log\left(\frac{P(t)}{1-P(t)}\right) = a(t) + \sum b_i x_i + \sum b_i x_i(t) \quad (15-6)$$

其中 $a(t)$ 代表 5 个不同的常量，这些常量分别代表了每一年被观察到的量，将在模型中被估计出来。

实际上，整个过程可以描述如下：对于处于风险下的每个个体的时间单位，可以产生一个分离的观察记录。在本例中，可以人年作为所观察的时间单位。这样，每个在第一年里改变工作的人就贡献了一人年数。那些在第三年里改变工作的人，每人则贡献了三人年数。而截断的个体那些在第 5 年里仍保持工作场所不变的人，每人贡献 5 人年数。这样，在样本为

200 个生物化学家的案例中，一共有 848 个人年数。从表 15-1 还可以清楚的看到，这一数值是 5 年中每一年风险数的累积。

对每一人年来说，若在该年度改变工作则记为 1，否则记为 0。其它的解释变量被分配给在每个人年应承担的数值。最后的步骤是将 848 个人年数引入这一样本，利用部分似然估计的方法对 logit 模型进行估计。利用 SPSS 对模型(15-6)中的回归系数的估计和统计检验如表 15-2 所示。

表 15-2 用 Logit 模型预测工作变换概率的估计，848 人年数

解释变量数	模型参数	
	b	t
学术权威	.056	.26
基金	-.078	-2.47*
出版物	-.023	-.79
引用著作	.0069	2.33*
任职级别	-1.6	-3.12**
第一年	-.96	-2.11*
第二年	-.025	-.06
第三年	-.74	-1.60
第四年	-.18	-.42
常量(>5)	2.35	
对数似然函数	-226.25	

*通过置信度为 95%的双侧检验

**通过置信度为 99%的双侧检验

尽管事件史分析方法在处理小样本问题上取得了很大的成功。但该方法也存在着两个重要的限制。其一是风险关于时间的变化在模型的估计时被忽略掉了；其二是比例风险模型及其相关的部分似然估计方法并不包括代表未观察到的异质性的随机扰动项。事实上，已经证明(Tuma, 1982)部分似然方法并不包含这样的扰动项的分析和估计。但若仅研究不重复事件时，这一限制的影响是很小的。

总而言之，事件史分析方法是一门正在发展的应用学科，其存在的问题将在未来的实践中得到不断的修正与完善。

练习题：

事件史分析方法的三个主要源流是什么？简述它们对事件史分析方法发展的影响。

简述事件史分析技术的特点和主要内容，并说明它与常规的数理统计方

法的不同之处。

什么是部分似然函数方法？它与极大似然函数方法有什么不同之处？

什么是风险率和风险函数？举例说明其应用。

什么是离散型事件史分析模型？简述其特点。

什么是连续型事件史分析模型？简述其特点。

试举一例说明事件史分析技术的应用。

附录：部分似然估计方法介绍

部分似然估计方法的第一步象极大似然估计方法一样，建立依赖于未知参数和观察值的似然函数；第二步是寻找使该方程极大化的参数值。然而，普通的似然函数是样本中所有个体的似然函数，而部分似然估计是在观察中所发生的事件所构成的似然函数。即：

$$PL = \prod L_k \quad (A1)$$

这里， PL 是部分似然函数， k 是样本中所发生事件的总数。

为了搞清楚 L_k 的构成，让我们仍然考虑表 A1 中的例子，这是一个有 10 个个体的样本，但只有 5 个事件被观察到，其它 5 个事件是截断的，三个是在 12 时截断的，因为研究是在那时终止。第 4 个和第 6 个事件分别是在 5 时和 9 时截断的。这两个截断事件发生的原因也许是由于个体的死亡，或从研究的样本中离开，或是不适于作进一步考察的个体。

表 A1 部分似然估计计算举例

i	t_i	k	L_k
1	2	1	$e^{bx_1} / (e^{bx_1} + e^{bx_2} + \cdots + e^{bx_{10}})$
2	4	2	$e^{bx_2} / (e^{bx_1} + e^{bx_2} + \cdots + e^{bx_{10}})$
3	5	3	$e^{bx_3} / (e^{bx_1} + e^{bx_2} + \cdots + e^{bx_{10}})$
4	5*		
5	6	4	$e^{bx_5} / (e^{bx_1} + e^{bx_2} + \cdots + e^{bx_{10}})$
6	9		

7	11	5	$e^{bx_7} / (e^{bx_1} + e^{bx_2} + \dots + e^{bx_{10}})$
8	12*		
9	12*		
10	12*		

注：i 表示个体， t_i 第 i 个个体事件发生或截断的时刻，k 表示事件数
*表示截断情况

为方便起见，这些被观察的个体按照事件先后出现的或被截断的时刻来排列的。如表 A 1 所示。第一个发生的事件是个体 1 在时刻 2 发生的。而事实上，样本中的十个个体在时刻 2 都处于发生事件的风险之中。则个体 1 在时刻 2 发生事件的概率为：

$$L_1 = \frac{h_1(2)}{h_1(2) + \dots + h_{10}(2)} \quad (A2)$$

假设风险函数服从指数分布，则有：

$$h_i = \exp(a(t) + bx_i) = \exp(a(t)) \exp(bx_i) \quad (A3)$$

代入(A2)并化简得：

$$L_1 = \frac{\exp(bx_1)}{\exp(bx_1) + \dots + \exp(bx_{10})} \quad (A4)$$

同理可得：

$$L_2 = \frac{\exp(bx_2)}{\exp(bx_2) + \dots + \exp(bx_{10})} \quad (A5)$$

其它诸如 L_3, L_4, L_5 的计算结果如表 A1 所示。

需要注意的是，每个 L_i 的值并不依赖于第 i 个事件的出现的确切时刻。其意义只是表示在第 i-1 个事件之后和第 i+1 个事件之前出现，并表示在任何时刻出现第 i 次事件的概率都是相同的。只有事件出现的顺序将影响部分似然函数的变化。

这样，将式(Ai) (i=2,3,4,5,6)代入式 A(1)之中，在取对数，就得到 LogPL 的表达式。一旦对数部分似然函数建立以后，就可以同极大似然方法一样对未知参数进行估计了。

附录一 Excel 在统计分析中的应用

(附录一由林飞编写)

第一节 中文 Excel 概述

一、中文 Excel 简介

Microsoft Excel 是美国微软公司开发的 Windows 环境下的电子表格系统,它是目前应用最为广泛的办公室表格处理软件之一。Excel 自诞生以来历经了 Excel 5.0、Excel 95、Excel 97 和 Excel 2000 等不同版本。随着版本的不断提高,Excel 软件的强大的数据处理功能和操作的简易性逐渐走入了一个新的境界,整个系统的智能化程度也不断提高,它甚至可以在某些方面判断用户的下一步操作,使用户操作大为简化。这些特性,已使 Excel 成为现代办公软件重要的组成部分。

Excel 具有强有力的数据库管理功能、丰富的宏命令和函数、强有力的决策支持工具,它具有以下主要特点。

(一) 分析能力

Excel 除了可以做一些一般的计算工作外,还有 400 多个函数,用来做统计、财务、数学、字符串等操作以及各种工程上的分析与计算。Excel 还专门提供了一组现成的数据分析工具,称为“分析工具库”,这些分析工具为建立复杂的统计或计量分析工作带来极大的方便。

(二) 操作简便

当需要将工作表上某个范围内的数据移到工作表上的另一个位置时,

只需按鼠标键,选取要移动的资料,将该范围资料拖动至所需的位置,

松开鼠标即可。如果要将公式或数据复制到临近的单元格内,可以拖

动“填充柄”,公式或数据就会被复制到目标单元格中。

此外,在使用 Excel 时,可以单击鼠标右键,屏幕上将出现相应的“快捷菜单”它将帮助用户尽快地寻找到所需要的常用命令。

Excel 内有许多工具按钮,每一个按钮代表一个命令。例如,要建立一个新的工作簿文件,就可直接按下工具栏第一个按钮,而不必先选择“文件”菜单,然后再选择其中的“打开”命令。在 Excel 97 中,系统共有 14 组常用的工具栏,用户可以自由选择加入或隐藏这些工具栏。

(三) 图表能力

在 Excel 中,系统大约有 100 多种不同格式的图表可供选用,用户只要做几个简单的按键动作,就可以制作精美的图表。通过图表指南一步步的引导,可使用不同的选项,得到所需的结果,满意的话就继续,不满意则后退一步,重新修改选项,直到最后出现完美的图表。

(四) 数据库管理能力

对于一个公司,每天都会产生许多新的业务数据,例如,销售数据、库存的变化、人事变动的数据资料等。这些数据必须加以处理,才能知道每段时间的销售金额、某个时候的存货量、要发多少薪水给每个员工等。要对这些数据进行有效的处理,就离不开数据库系统。所谓数据库系统,就是一组有组织的信息。例如,将每位职员简历写在一张卡片上,卡片放在盒子内,盒子内的数据通常组成列和行。再如,每种产品的产地、规格、单位、单价、数量组成一行,每一行都包含同一属性的数据,每列都包含同一品种的数据,即每列都有产地、规格、单位、单价、数量。

管理数据库可用专门的数据库管理系统,如 Access、FoxPro、SQL Server、Oracle、Sybase 等。在 Excel 中提供了类似的数据库管理功能,保存在工作表内的数据都是按照相应的行和列存储的,这种数据结构再加上 Excel 提供的有关处理数据库的命令和函数,使得 Excel 具备了组织和管理大量数据的能力。

(五) 宏语言功能

利用 Excel 中的宏语言功能,用户可以将经常要执行的操作的全过程记录下来,并将此过程用一简单的组合按键或工具按钮保存起来。这样,在下一次操作中,只需按下所定义的宏功能的相应按键或工具按钮即可,而不必重复整个过程。例如,可以定义一个打开最后编辑文件且可以自动执行的宏,以后当用户打开 Excel 后,将自动打开上一次编辑的工作簿。

在 Excel 中,用户可使用 Visual Basic For Application (VBA) 语言,进行

宏命令的开发。利用宏命令，用户可以将 Excel 的下拉菜单和对话框更改或将图形按钮的说明更换，使它们更适合于用户的工作习惯和特殊要求。

（六）样式功能

在 Excel 中，用户可以利用各种文字格式化的工具和制图工具，制作出美观的报表。Excel 工作表里的资料，在打印以前可将其放大或缩小进行观察，用户可以对要打印的文件作微调。

用户可将要打印出的格式制作好，并存储成样本，以后可以读取此样本文件，就可依据样本文件的格式打印出美观的报表。Excel 的专业文书处理程序具有样式工具。所谓样式，就是将一些格式化的组合用一个名称来表示，以后要使用这些格式化的组合时，只要使用此名称即可，因此可大幅度地节省报表格式化的时间。

（七）对象连接和嵌入功能

利用对象连接和嵌入功能，用户可将其他软件（例如，画图）制作的图形插入到 Excel 的工作表中。当需要更改图案时，只要在图案上双击鼠标键，制作该图案的程序就会自动打开，图案将出现在该图形编辑软件内，修改、编辑后的图形也会在 Excel 内显示出来。也可以将一个声音文件或动画文件嵌入到 Excel 工作表中，使工作表变成一幅声形并茂的报表。

（八）连接和合并功能

通常，每个工作在一张工作表上执行即可，早期的工作表软件都只能在一张工作表上执行。但有时需要同时用到多张工作表，例如，公司内每个分公司每月都会有会计报表，要将各分公司区的资料汇总起来，就需要用到连接和合并功能。Excel 很容易将工作表连接起来，并进行汇总工作。Excel 内一个工作簿可以存放许多工作表、图形等，每个工作簿文件最多可以由 255 张工作表组成。

二、Excel 的安装和启动

以下以 Excel 97 为例，说明 Excel 的安装过程。如果没作特别说明，本章介绍其它内容都是针对 Excel 97 软件的。

Excel 97 作为 Office 97 系统中的一个应用程序，它的工作平台为 Windows 95 或其以上的版本。对中文 Excel 而言，最好安装在中文 Windows 系统平台之上。

（一）安装 Excel

1. 在 CD — ROM 中放入 Office 97 光盘。

2. 单击 Windows 的“开始”按钮并选择“运行”命令。在弹出的“运行”对话框，“打开”框中寻找光盘上的安装程序“SETUP.EXE”，找到后单击“确定”按钮即可。（如果光盘具有自动执行的功能，则可省略以上步骤，直接点取屏幕上的“开始安装”按钮。）

3. 开始安装后系统将要求用户选择安装模式：“典型”、“自定义”、“快速”。用户可根据需要选择安装模式，一般可以选择“典型”安装方式，那么安装软件将装入最常用的 Office 组件，如果用户想定制所要安装的组件，那么可以选择“自定义”安装方式，而“快速”安装方式是一种最节约磁盘空间的安装方式，它仅装入 Office 运行中不能缺少的组件。

在剩下安装过程中则用户按屏幕提示选择。

（二）Excel 的启动与退出。

启动 Excel 的常用方法是：单击任务栏上的“开始”按钮，此时屏幕上出现一个弹出式菜单，将鼠标指向“程序”项后，屏幕出现另一个弹出菜单，单击“Microsoft Excel”，就可以启动 Excel 系统，此时屏幕上出现如图附-1 所示的 Excel 主工作画面。若安装 Excel 时，生成了“快捷工具栏”，则双击其中的 Excel 按钮也可立即启动 Excel。

退出 Excel 常用的方法是，用鼠标单击标题栏的  按钮或进入“文件”菜单栏，单击“退出”选项。

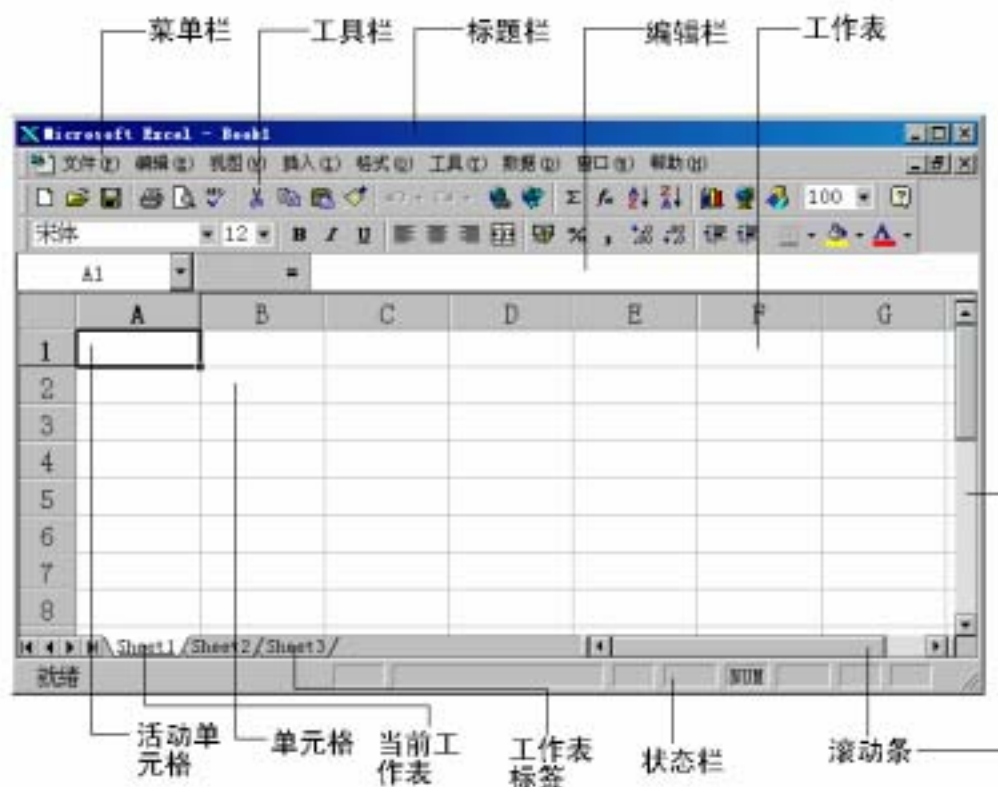


图 附-1 Excel 的工作界面

三、Excel 工作界面简介

按图附-1 从上到下的顺序, Excel 工作界面包含如下几项内容:“标题”栏、“菜单”栏、“工具”栏、“编辑”栏、工作表、工作表标签、滚动条、和“状态”栏。下面分别介绍它们的作用。

(一)“标题”栏

“标题”栏告诉用户正在运行的程序名称和正在打开的文件名称。如图附-1 所示,标题栏显示“Microsoft Excel-Book1”表示此窗口的应用程序为 Microsoft Excel,在 Excel 中打开的文件的文件名为 Book1.xls。

(二)“菜单”栏

“菜单”栏按功能把 Excel 命令分成不同的菜单组,它们分别是“文件”、“编辑”、“视图”、“插入”、“格式”、“工具”、“表格”、“帮助”。当菜单项被选中时,引出一个下拉式菜单,可以从中选取相应的子菜单。

另外,在屏幕的不同地方单击鼠标右键时,“快捷菜单”将出现在鼠标

指针处。选取“快捷菜单”中的命令同从菜单栏的菜单上选取相应命令的效果是一样的，但选取速度明显增快。

（三）“工具”栏

Excel 可显示几种工具栏，这些工具可控制简化用户的操作。“工具”栏中的按钮都是菜单中常用命令的副本，当鼠标指向某一按钮后，稍等片刻在按钮右下方会显示该按钮命令的含意。用户可以配置“工具”栏的内容，通过“视图”菜单中的“工具”栏子菜单来选择显示不同类型的“工具”或全部显示出来。下面介绍出现在 Excel 开始屏幕中的两种“工具”栏。

1. “常用”工具栏

“常用”工具栏中为用户准备了访问 Excel 最常用命令的快捷按钮，如“新建文件”按钮，“打开文件”按钮，“保存文件”按钮等。

2. “格式”工具栏

“格式”工具栏专门放那些和文本外观有关的命令，如字体、字号、对齐方式及其他选项。

（四）“编辑”栏

“编辑”栏给用户提供活动单元格的信息。在“编辑”栏中用户可以输入和编辑公式，“编辑”栏位于图附-1 中第 5 行。

“编辑”栏由“名字”栏和“公式”栏组成。位于“编辑”栏左侧的“名字”栏中显示的是活动单元格的坐标，也可在“名字”栏中直接输入一个或一块单元格的地址进行单元格的快速选定；位于“编辑”栏右侧的“公式”栏可用于编辑活动单元格的内容，它包含三个按钮和一个编辑区。当向活动单元格输入数据时，公式栏中便出现三个按钮，三个按钮从左至右分别是：“ ”（取消）按钮、“√”（确认）按钮和“=”（公式指南）按钮。

通常 Excel 在工作区中显示“编辑”栏。在“视图”菜单中的“编辑栏”命令是一个开关命令，它可以用于隐藏或显示“编辑”栏。

（五）工作表

工作簿窗口包含了 16 张独立的工作表（sheet）。开始时，窗口中显示第一张工作表“Sheet1”，该表为当前工作表。当前工作表只有一张，用户可通过点击工作表下方的标签激活其他工作表为当前工作表。

工作表是一个由行和列组成的表格。行号和列号分别用字母和数字区别。行由上自下范围 1 ~ 65536，列号则由左到右采用字母编号 A ~ IV。因此每张表为 256 列 X 65536 行，若从 Excel 导入的数据超过以上范围，则会被 Excel 自动截去。每一个行、列坐标所指定的位置称之为单元格。在单元格中用户可以键入符号、数值、公式以及其他内容。

（六）工作表标签

工作表标签通常用“Sheet1”、“Sheet2”等名称来表示，用户也可以通过用鼠标右击标签名，选择弹出菜单中“重命名”命令来修改标签名。Excel 一般同时显示工作表队列中的前 3 个标签。利用标签队列左边的一组标签滚动按钮可显示队列中的后续工作表的标签。工作簿窗口中的工作表称之为当前工作表，当前工作表的标签为白色，其他为灰色。

（七）“滚动”栏

当工作表很大时，如何在窗口中查看表中的全部内容呢？可以使用工作簿窗口右边及下边的滚动栏，使窗口在整张表上移动查看，也可以通过修改常用“工具”栏中“显示比例框”的参数来扩大整个工作表的显示范围。

（八）“状态”栏

“状态”栏位于 Excel 窗口底部，它的左端是信息区，右端是键盘状态区。

在信息区中，显示的是 Excel 的当前工作状态。例如，当工作表准备接受命令或数据时，信息区显示“就绪”；当在“编辑”栏中键入新的内容时，信息区显示“输入”；当选取菜单命令或工具按钮时，信息区显示此命令或工具按钮用途的简要提示。

在键盘状态区中，显示的是若干按键的开关状态。例如，当按[Caps Lock]键时，状态栏中便显示“CAPS”。

与“编辑”栏相同，在“视窗”菜单中的“状态”命令是一个开关命令，它可以用于隐藏或显示“状态”栏。

第二节 Excel 基本操作

一、Excel 操作方法概述

要完成任一项 Excel 操作一般都可以找到三种操作方法：鼠标操作、

菜单操作和键盘命令操作。例如,想要将 A1 单元格的数据复制到 A2 单元格去,有如下几种操作方法:

(一) 鼠标操作法:先用鼠标选中 A1 单元格,然后缓慢移动鼠标到 A1 单元格的右下角,当鼠标的形状变为黑色实心“十”字形之后,拖动鼠标到 A2 单元格,然后放开鼠标,则 A1 的数据就复制到 A2 单元格了。

(二) 菜单操作法:先用鼠标选中 A1 单元格,选择菜单[编辑]=>[复制],然后用鼠标选中 A2 单元格,再选择[编辑]=>[粘贴]命令,数据就复制到 A2 单元格了。

(三) 键盘命令操作法:直接用鼠标选中 A2 单元格,从键盘输入“=A1”命令,则复制即告完成。

以上是 Excel 中很典型的三种操作方法。在实际使用过程中,应根据实际情况,尽量选择三种方法中最简洁的操作方法,以提高操作速度。

二、文件基本操作

(一) 新建文件:选择[文件]=>[新建]即可创建一个新的 Excel 文件。

(二) 打开文件:选择[文件]=>[打开],可在 Excel 中打开一个已经存在的数据文件。它可以是 Excel 的数据文件,也可能是 Excel 兼容的其它软件的数据文件。可在不同窗口中同时打开多个数据文件,通过[窗口]菜单下方的不同选项,进行不同窗口的切换。

(三) 保存文件:选择[文件]=>[保存],可保存当前数据文件。如果选择[另存为],可将当前工作簿存为一个新的文件。保存文件的格式可以是 Excel 的数据文件,也可能是 Excel 兼容的其它软件的数据文件。

(四) 文件打印:选择[文件]=>[打印],可打印当前的工作簿文件。打印之前,可以选择[文件]=>[页面设置]和选择[文件]=>[打印预览],进行打印前的页面设置操作和打印效果的预先浏览。

三、数据的输入输出操作

(一) 数据的手动输入

建立一个新的 Excel 文件之后,便可进行数据的输入操作。Excel 中以单元格为单位进行数据的输入操作。一般用上下左右光标键,Tab 键或用鼠标选中某一单元格,然后输入数据。

Excel 中的数据按类型不同通常可分为四类:数值型,字符型,日期

型, 和逻辑型。Excel 根据输入数据的格式自动判断数据属于什么类型。如日期型的数据输入格式为“月/日/年”, “月-日-年”或“时:分:秒”。要输入逻辑型的数据, 输入“true”(真)或“false”(假)即可。若数据由数字与小数点构成, Excel 自动将其识别为数字型, Excel 允许在数值型数据前加入货币符号, Excel 将其视为货币数值型, Excel 也允许数值型数据用科学记数法表示, 如 2×10^9 在 Excel 中可表示为 2E+9。除了以上三种格式以外的数据, Excel 将其视为字符型处理。

(二) 公式生成数据

Excel 的数据中也可由公式直接生成。例如: 在当前工作表中 A1 和 B1 单元格中已输入了数值数据, 欲将 A1 与 B1 单元格的数据相加的结果放入 C1 单元格中, 可按如下步骤操作: 用鼠标选定 C1 单元格, 然后输入公式“=A1+B1”或输入“=SUM(a1:b1)”, 回车之后即可完成操作。C1 单元格此时存放实际上是一个数学公式“A1+B1”, 因此 C1 单元格的数值将随着 A1、B1 单元格的数值的改变而变化。Excel 提供了完整的算术运算符, 如 + (加)、- (减)、* (乘)、/ (除)、% (百分比)、^ (指数) 和丰富的函数, 如 SUM (求和)、CORREL (求相关系数)、STDEV (求标准差) 等, 供用户对数据执行各种形式的计算操作, 在 Excel 帮助文件中可以查到各类算术运算符和函数的完整使用说明。

(三) 复制生成数据

Excel 中的数据也可由复制生成。实际上, 在生成的数据具有相同的规律性的时候, 大部分的数据可以由复制生成。可以在不同单元格之间复制数据, 也可以在不同工作表或不同工作簿之间复制数据, 可以一次复制一个数据, 也可同时复制一批数据, 为数据输入带来了极大的方便。普通单元格的复制结果与公式单元格的复制结果相差较大, 下面分别予以说明。

1. 普通单元格指的是非公式的单元格。普通单元格的复制, 一般可以按如下步骤进行:

- (1) 拖动鼠标选定待复制的区域, 选定之后该区域变为黑色。Excel 可以进行整行、整列或整个表格的选定操作, 例如, 如果要选定表格的第一列, 可直接用鼠标单击列标“A”, 如果要选定表格的第一行, 可直接用鼠标单击行标“1”, 如果要选定整个表格, 可直接点击全选按钮, 如图附-2 所示:

	A	B	C
1	3	4	7
2	32	33	7
3	432	2323	
4			
5			
6			
7			

图 附-2

- (2) 选定完区域之后,用鼠标右击该区域,选择“复制”,将区域内容复制到粘贴版之中。可以发现该区域已被虚线包围。
- (3) 用鼠标右击目标区域,选择“粘贴”,则单元格区域的复制即告完成。

2. 公式单元格的复制,一般可分为两种,一种是值复制,一种是公式复制。值复制指的是只复制公式的计算结果到目标区域,公式复制指的是仅复制公式本身到目标区域。下面对它们的操作步骤分别予以说明。

- (1) 值复制: 拖动鼠标选定待复制区域。

用鼠标右击选定区域,选择“复制”选项。

用鼠标右击目标区域,再单击“选择性粘贴”子菜单。出现复制选项,选定“数值”选项,然后用鼠标单击“确定”按钮,则公式的值复制即告完成。

- (2) 公式复制:

公式复制是 Excel 数据成批计算的重要操作方法,要熟练公式复制的操作首先要区分好两个概念:单元格的相对引用与绝对引用。Excel 中的公式中一般都会引用到别的单元格的数值,如果你希望当公式复制到别的区域之时,公式引用单元格不会随之相对变动,那么你必须要在公式中使用单元格的绝对引用。如果你希望当公式复制到别的区域之时,公式引用单元格也会随之相对变动,那么你必须要在公式中使用单元格的相对引用。在公式中如果直接输入单元格的地址,那么默认的是相对引用单元格,如果在单元格的地址之前加入“\$”符号那么意味着绝对引用单元格。例如,在当前工作表中 A1 和 B1 单元格中已输入了数值数据,用鼠标选定 C1 单元格,然后输入公式“=A1+B1”,此公式引用的便是两个相对的单元格 A1、B1,也就是说,如果将该公式复制到 C2 的单元格,公式所引用的单元格的地址将随着发生变化,公式将变为“=A2+B2”,如果将该公式复制到 F100 的单元格,那么公式将变为“=D100+E100”,这就是相对引用的结果,公式的内容随着公式的位置变化而相对变化。如果在 C1 单元格输入的是

“=A\$1+B\$1”那么此公式引用的便是绝对的单元格，不论将公式复制到何处，公式的内容都不会发生变化。当然，绝对引用和相对引用亦可在同一公式之中混合交叉使用，例如，如果在 C1 单元中输入的是公式“=A\$1+B\$1”，那么意味着，公式的内容不会随着公式的垂直移动而变动，而是随着公式的水平移动而变动，如果将该公式复制到 F100 单元格，那么公式将变为，“=D\$1+E\$1”。可以作这样的归纳：公式中“\$”符号后面的单元格坐标不会随着公式的移动而变动，而不带“\$”符号后面的单元格坐标会随着公式的移动而变动。

在实际的使用中，如果能把单元格的相对引用与绝对引用灵活应用到 Excel 的公式之中，能为数据成批准确运算带来极大的方便。

四、数据的移动操作

数据的移动操作可按如下步骤进行：

- （一）拖动鼠标选定待移动区域。
- （二）用鼠标右击选定区域，选择“剪切”选项。
- （三）用鼠标右击目标区域，选择“粘贴”，则单元格区域的移动即告完成。

与数据的复制操作不同，公式单元格的移动操作不存在值移动或公式移动的区别，也不存在绝对引用已相对引用的区别，移动操作将把公式单元格的公式内容原原本本移动到目标区域，不作任何改动。

五、数据的删除操作

数据的删除操作可按如下步骤进行：

- （一）拖动鼠标选定待删除区域。
- （二）用鼠标右击选定区域，选择“删除”，即可删除单元格区域的内容。

如果不小心删除了不该删除的区域，可以通过“编辑”菜单的“撤消”命令来恢复被删除的内容。“撤消”操作是 Excel 中较常用到的操作，如果不小心实施了错误的操作，那么可以通过“撤消”操作使工作表恢复原样。

六、与其它软件交换数据的方法

在 Excel 中可以打开其它类型的数据文件，如 FOXPRO 系列的 DBF 数据库文件，文本文件，lotus1-2-3 的数据文件等。具体操作方法如下：

(一) 选择[文件]=>[打开]。

(二) 在[打开文件]对话框中选择所要打开的文件类型及其所在的目录。

(三) 用鼠标双击该文件名，并按 Excel 提示步骤操作即可打开该文件。

Excel 文件同样也可存为其它类型的数据文件，具体操作方法如下：

(一) 编辑好文件后，选择[文件]=>[另存为]。

(二) 在[另存为]对话框中选择所要打开文件的类型及其所在的目录。

(三) 输入文件名之后，用鼠标单击[保存]按钮即可。

第二节介绍了 Excel 的一些比较主要的基本操作方法，在 Excel 中还有许多其它的基本操作，如表格显示格式控制，打印格式控制，Excel 帮助的使用等等，在应用 Excel 进行统计数据分析之前熟练这些操作是非常有必要的。第三节开始介绍 Excel 在统计分析中的应用。

第三节 Excel 在描述统计中的应用

在使用 Excel 进行数据分析时，要经常使用到 Excel 中一些函数和数据分析工具。其中，函数是 Excel 预定义的内置公式。它可以接受被称为参数的特定数值，按函数的内置语法结构进行特定计算，最后返回一定的函数运算结果。例如，SUM 函数对单元格或单元格区域执行相加运算，PMT 函数在给定的利率、贷款期限和本金数额基础上计算偿还额。函数的语法以函数名称开始，后面是左圆括号、以逗号隔开的参数和右圆括号。参数可以是数字、文本、形如 TRUE 或 FALSE 的逻辑值、数组、形如 #N/A 的错误值，或单元格引用。给定的参数必须能产生有效的值。参数也可以是常量、公式或其它函数。

Excel 还提供了一组数据分析工具，称为“分析工具库”，在建立复杂的统计分析时，使用现成的数据分析工具，可以节省很多时间。只需为每一个分析工具提供必要的数据和参数，该工具就会使用适宜的统计或数学函数，在输出表格中显示相应的结果。其中的一些工具在生成输出表格时还能同时产生图表。如果要浏览已有的分析工具，可以单击“工具”菜单中的“数据分析”命令。如果“数据分析”命令没有出现在“工具”菜单上，则必须运行“安装”程序来加载“分析工具库”。安装完毕之后，必须通过“工具”菜单中的“加载宏”命令，在“加载宏”对话框中选择并启

动它。

一、描述统计工具

(一) 简介：此分析工具用于生成对输入区域中数据的单变量分析，提供数据趋中性和易变性等有关信息。

(二) 操作步骤：

1. 用鼠标点击工作表中待分析数据的任一单元格。
2. 选择“工具”菜单的“数据分析”子菜单。
3. 用鼠标双击数据分析工具中的“描述统计”选项。
4. 出现“描述统计”对话框，对话框内各选项的含义如下：

输入区域：在此输入待分析数据区域的单元格范围。一般情况下 Excel 会自动根据当前单元格确定待分析数据区域。

分组方式：如果需要指出输入区域中的数据是按行还是按列排列，则单击“行”或“列”。

标志位于第一行/列：如果输入区域的第一行中包含标志项(变量名)，则选中“标志位于第一行”复选框；如果输入区域的第一列中包含标志项，则选中“标志位于第一列”复选框；如果输入区域没有标志项，则不选任何复选框，Excel 将在输出表中生成适宜的数据标志。

均值置信度：若需要输出由样本均值推断总体均值的置信区间，则选中此复选框，然后在右侧的编辑框中，输入所要使用的置信度。例如，置信度 95%可计算出的总体样本均值置信区间为 10，则表示：在 5%的显著水平下总体均值的置信区间为($\bar{X} - 10, \bar{X} + 10$)。

第 K 个最大/小值：如果需要在输出表的某一行中包含每个区域的数据的第 k 个最大/小值，则选中此复选框。然后在右侧的编辑框中，输入 k 的数值。

输出区域：在此框中可填写输出结果表左上角单元格地址,用于控制输出结果的存放位置。整个输出结果分为两列,左边一列包含统计标志项，右边一列包含统计值。根据所选择的“分组方式”选项的不同，Excel 将为输入表中的每一行或每一列生成一个两列的统计表。

新工作表：单击此选项，可在当前工作簿中插入新工作表，并由新工作表的 A1 单元格开始存放计算结果。如果需要给新工作表命名，则在右侧编辑框中键入名称。

新工作簿：单击此选项，可创建一新工作簿，并在新工作簿的新工作表中存放计算结果。

汇总统计：指定输出表中生成下列统计结果，则选中此复选框。这些统计结果有：平均值、标准误差、中值、众数、标准偏差、方差、峰值、偏斜度、极差（全距）最小值、最大值、总和、样本个数。

5. 填写完“描述统计”对话框之后，按“确定”按钮即可。

（三）结果说明：描述统计工具可生成以下统计指标，按从上到下的顺序其中包括样本的平均值($\bar{X}$)，标准误差($S/\sqrt{n}$)，组中值(Medium)，众数(Mode)，样本标准差(S)，样本方差(S^2)，峰度值，偏度值，极差(Max-Min)，最小值(Min)，最大值(Max)，样本总和，样本个数(n)和一定显著水平下总体均值的置信区间。

二．直方图工具

（一）简介：直方图工具，用于在给定工作表中数据单元格区域和接收区间的情况下，计算数据的个别和累积频率，可以统计有限集中某个数值元素的出现次数。例如，在一个有 50 名学生的班级里，可以通过直方图确定考试成绩的分布情况，它会给出考分出现在指定成绩区间的学生个数，而用户必须把存放分段区间的单元地址范围填写在直方图工具对话框中的“接收区域”框中。

（二）操作步骤：

1. 用鼠标点击表中待分析数据的任一单元格。
2. 选择“工具”菜单的“数据分析”子菜单。
3. 用鼠标双击数据分析工具中的“直方图”选项。
4. 出现“直方图”对话框，对话框内主要选项的含义如下：

输入区域：在此输入待分析数据区域的单元格范围。

接收区域(可选)：在此输入接收区域的单元格范围，该区域应包含一组可选的用来计算频数的边界值。这些值应当按升序排列。只要存在的话，Excel 将统计在各个相邻边界直之间的数据出现的次数。如果省略此处的接收区域，Excel 将在数据组的最小值和最大值之间创建一组平滑分布的接收区间。

标志：如果输入区域的第一行或第一列中包含标志项，则选中此复选框；如果输入区域没有标志项，则清除此该复选框，Excel 将在输出表中生成适宜的数据标志。

输出区域：在此输入结果输出表的左上角单元格的地址。如果输出表将覆盖已有的数据，Excel 会自动确定输出区域的大小并显示信息。

柏拉图：选中此复选框，可以在输出表中同时显示按降序排列频率数据。如果此复选框被清除，Excel 将只按升序来排列数据。

累积百分比：选中此复选框，可以在输出结果中添加一列累积百分比数值，并同时在直方图表中添加累积百分比折线。如果清除此选项，则会省略以上结果。

图表输出：选中此复选框，可以在输出表中同时生成一个嵌入式直方图表。

5. 按需要填写完“直方图”对话框之后，按“确定”按钮即可。

（三）结果说明：完整的结果通常包括三列和一个频率分布图，第一列是数值的区间范围，第二列是数值分布的频数，第三列是频数分布的累积百分比。

三、利用 Excel 绘制散点图

（一）简介：散点图是观察两个变量之间关系程度最为直观的工具之一，利用 Excel 的图表向导，可以非常方便的创建并且改进一个散点图，也可以在一个图表中同时显示两个以上变量之间的散点图。

（二）操作步骤：数据如图附-3 所示，

	A	B	C. 编辑
1	x	y	z
2	68	68	312
3	71	69	323
4	72	70	345
5	70	81	366
6	76	85	378
7	77	86	390
8	76	100	411
9	78	108	434
10	79	114	449
11	81	120	469
12	88	133	480

图 附-3

可按如下步骤建立变量 x-y，x-z 的散点图：

1. 拖动鼠标选定数值区域 A2:C12，不包括数据上面的标志项。
2. 选择“插入”菜单的“图表”子菜单，进入图表向导。
3. 选择“图表类型”为“散点图”，然后单击“下一步”。
4. 确定用于制作图表的数据区。Excel 将自动把你前面所选定的数据区的地址放入图表数据区的内。
5. 在此例之中，需要建立两个系列的散点图，一个是 x-y 系列的散点

图，一个是 x - z 系列的散点图，因此，必须单击“系列”标签，确认系列 1 的“X 值”方框与“数值方框”分别输入了 x, y 数值的范围，在系列 2 的“X 值”方框与“数值方框”分别输入了 x, z 数值的范围。在此例中，这些都是 Excel 已经默认的范围，所以，可忽略第 5 步，直接单击“下一步”即可。

6. 填写图表标题为“X-Y 与 X-Z 散点图”，X 轴坐标名称为“X”与 Y 轴坐标名称“Y/Z”，单击“下一步”。

7. 选择图表输出的位置，然后单击“完成”按钮即生成图附-4 的图表。

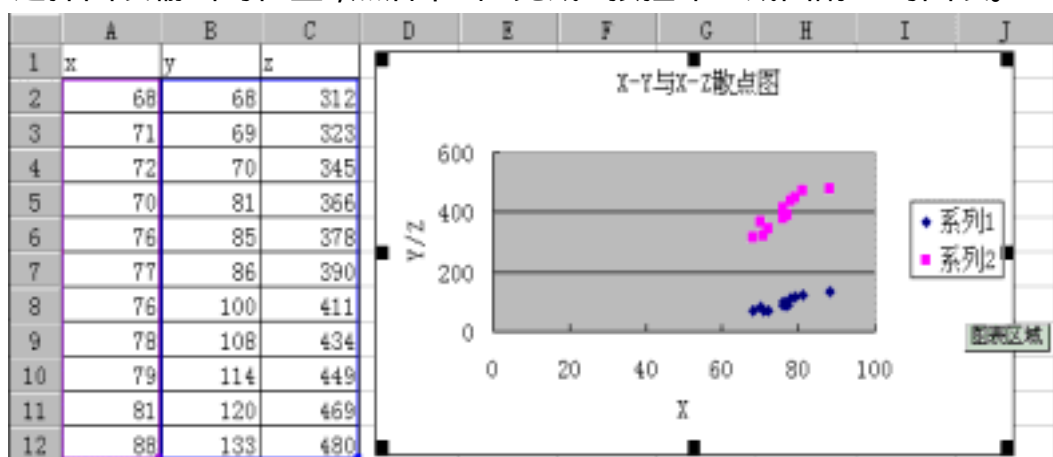


图 附-4

(三) 结果说明：如图附-4 所示，Excel 中可同时生成两个序列的散点图，并分为两种颜色显示。通过散点图可观察出两个变量的关系，为变量之间的建立模型作准备。

四、数据透视表工具

(一) 简介：数据透视表是 Excel 中强有力的数据列表分析工具。它不仅可以用来作单变量数据的次数分布或总和分析，还可以用来作双变量数据的交叉频数分析、总和分析和其它统计量的分析。

(二) 操作步骤：如图附-5 所示，表中列出学生两门功课评定结果，

	A	B	C	D	E	F
1	学号	语文	数学			
2	1001	优	差			
3	1002	良	中			
4	1003	中	中			
5	1004	差	中			
6	1005	差	差			
7	1006	中	良			
8	1007	中	优			
9	1008	差	良			
10	1009	良	中			

图 附-5

可按如下步骤建立交叉频数表：

1. 选中图附-5 中表格中有数据的任一单元格，然后选择“数据”菜单的“数据透视表”子菜单，进入数据透视表向导。
2. 选择“Microsoft Excel 数据清单或数据库”为数据源。单击“下一步”。
3. 选择待分析的数据的区域，一般情况下 Excel 会自动根据当前单元格确定待分析数据区域，因此你只要直接单击“下一步”按钮即可。
4. 确定数据透视表的结构，在此例中，要建立的是一个交叉频数表，分别按语文和数学的成绩对学生的人数进行交叉频数分析，因此可按图附-6 将三个按钮“学号”、“语文”、“数学”分别拖放到表格的指定部位，并且双击“求和项：学号”，将其改为记数项，结果如图附-6 所示，然后单击“下一步”按钮。



图 附-6

5. 选择数据透视表的显示位置之后，单击“完成按钮”，可出现如图附-7 所示的数据透视表。

	A	B	C	D	E	F
1	计数项:学号	语文				
2	数学	差	良	优	中	总计
3	差	2		2	2	6
4	良	1	2		2	5
5	优			1	2	3
6	中	2	2		1	5
7	总计	5	4	3	7	19

图 附-7

(三) 结果说明：如图附-7 的结果所示，数据透视表可以作为一个交叉频数分析工具。完成数据透视表之后，可按需要修改数据表的显示格式。例如，如果想要把表格中的频数替换成为百分比数。可以用鼠标右击频数的任一单元格，选择“字段”子菜单，单击“选项”按钮，将“数据显示方式”替换成为“占总和的百分比”，然后单击“确定”按钮即可。按同样方式，可将数据透视表修改成为其它不同样式。

五、排位与百分比工具

(一) 简介：此分析工具可以产生一个数据列表，在其中罗列给定数据集中各个数值的大小次序排位和相应的百分比排位。用来分析数据集中各数值间的相互位置关系。

(二) 操作步骤：

1. 用鼠标点击表中待分析数据的任一单元格。
2. 选择“工具”菜单的“数据分析”子菜单。
3. 用鼠标双击数据分析工具中的“排位与百分比”选项。
4. 填写完“排位与百分比”对话框，单击“确定”按钮即可。

(三) 结果说明：输出的结果可分为四列，第一列“点”是数值原来的存放位置，第二列是相应的数值，第三列是数值的排序号，第四列是数值的百分比排位，它的计算方法是：小于该数值的数值个数/(数值总个数-1)。

第四节 Excel 在推断统计中的应用

一、二项分布工具

(一) 简介：在 Excel 中想要计算二项分布的概率分布、累积概率，

需要利用 Excel 的工作表函数 BINOMDIST。函数 BINOMDIST 适用于固定次数的独立实验，实验的结果只包含成功或失败二种情况，且每次实验成功的概率固定不变。例如，已知次品概率的情况下，函数 BINOMDIST 可以计算 10 个产品中发现 2 个次品的概率。以下例子说明如何在 Excel 中计算二项分布的概率，以及如何建立二项分布图表。

(二) 操作步骤：例子如下所示，一个推销员打了六个电话，推销成功的概率是 0.3，那么可以按以下步骤建立推销成功次数的概率分布图表。

1. 图附-8 所示，先在 Excel 之下建立好概率分布表格的框架。

	A	B	C	D	E	F
1	二项分布概率分布表					
2	试验总次数	6				
3	每次成功概率	0.3				
4						
5		概率				
6	成功次数(k)	P(Y=k)	P(Y≤k)	P(Y<k)	P(Y>k)	P(Y≥k)
7	0					
8	1					
9	2					
10	3					
11	4					
12	5					
13	6					

图 附-8

2. 图附-9 所示,先在 B7 至 F7 单元格分别输入概率计算公式。

	A	B	C	D	E	F
1	二项分布概率分布表					
2	试验总次数	6				
3	每次成功概率	0.3				
4						
5		概率				
6	成功次数(k)	P(Y=k)	P(Y≤k)	P(Y<k)	P(Y>k)	P(Y≥k)
7	0	=BINOMDIST(A7,6,0.3,0)	=BINOMDIST(A7,6,0.3,1)	=C7-B7	=1-C7	=1-D7
8	1					
9	2					
10	3					
11	4					
12	5					
13	6					

图 附-9

3. 式的拷贝。选取 B7 至 F7 单元格，然后移动鼠标至 F7 单元格的右下角，使其成为黑色实心十字星状，一般称之为“填充柄”，拖

动“填充柄”至 F13 单元格即可完成公式的拷贝操作。结果图附-10 所示。

	A	B	C	D	E	F
1	二项分布概率分布表					
2	试验总次数	6				
3	每次成功概率	0.3				
4						
5		概率				
6	成功次数(k)	P(Y=k)	P(Y≤k)	P(Y<k)	P(Y>k)	P(Y≥k)
7	0	0.117649	0.117649	0	0.88235	1
8	1	0.302526	0.420175	0.11765	0.57983	0.88235
9	2	0.324135	0.74431	0.42018	0.25569	0.57983
10	3	0.18522	0.92953	0.74431	0.07047	0.25569
11	4	0.059535	0.989065	0.92953	0.01094	0.07047
12	5	0.010206	0.999271	0.98906	0.00073	0.01094
13	6	0.000729	1	0.99927	0	0.00073

图 附-10

4. 下面开始创建二项分布图表。选取 B7 至 B13 单元格, 选取“插入”菜单的“图表”子菜单。
5. 选择“柱状图”, 然后单击“下一步”。
6. 单击“系列”标签, 单击“分类(X)轴标志”框, 并用鼠标选取 A7 至 A13 单元格为图表 X 轴的轴标, 然后单击“下一步”。
7. 分别键入图表名称“二项分布图”, X 轴名称“成功次数”, Y 轴名称“成功概率”, 单击“完成”按钮即可生成二项分布图表。

(三) 结果说明: 如图附-10 所示, 利用 Excel 的 BINOMDIST 的函数可以计算出二项分布的概率以及累积概率。BINOMDIST 函数可以带四个参数, 各参数的含义分别是: 实验成功的次数, 实验的总次数, 每次实验中成功的概率, 是否计算累积概率。第四个参数是一个逻辑值, 如果为 TRUE, 函数 BINOMDIST 返回累积分布函数, 如果为 FALSE, 返回概率密度函数。

二、其它分布的函数

(一) 函数 CRITBINOM:

1. 说明: 函数 CRITBINOM 可称为 BINOMDIST 的逆向函数, 它返回使累积二项式分布概率 $P(X \leq x)$ 大于等于临界概率值的最小值。

2. 语法: CRITBINOM(trials, probability_s, alpha)

Trials: 贝努利实验次数。

Probability_s: 一次试验中成功的概率。

Alpha : 临界概率。

- 3 . 举例: CRITBINOM(6,0.5,0.75) 等于 4 , 表明如果每次试验成功的概率为 0.5,那么 6 次试验中成功的次数小于等于 4 的概率恰好超过或等于 0.75 。

(二) 函数 HYPGEOMDIST :

1 . 说明 : 函数 HYPGEOMDIST 返回超几何分布。给定样本容量、总体容量和样本总体中成功的次数 , 函数 HYPGEOMDIST 返回样本取得给定成功次数的概率。使用函数 HYPGEOMDIST 可以解决有限总体的问题 , 其中每个观察值或者为成功或者为失败 , 且给定样本区间的所有子集有相等的发生概率。

2 . 语法 : HYPGEOMDIST(sample_s,number_sample,population_s,number_population)

Sample_s : 样本中成功的次数。

Number_sample : 样本容量。

Population_s : 样本总体中成功的次数。

Number_population : 样本总体的容量。

3 . 举例 : 容器里有 20 块巧克力 , 8 块是焦糖的 , 其余 12 块是果仁的。如果从中随机选出 4 块 , 下面函数计算式计算出只有一块是焦糖巧克力的概率 : HYPGEOMDIST(1,4,8,20)= 0.363261。

(三) 函数 NEGBINOMDIST :

1 . 说明 : 函数 NEGBINOMDIST 返回负二项式分布。当每次试验成功概率固时 , 函数 NEGBINOMDIST 返回在到达指定次数成功之前 , 出现 n 次失败的概率。此函数与二项式分布相似 , 只是它的成功次数固定 , 试验总数为变量。与二项分布类似的是 , 试验次数被假设为自变量。

2 . 语法: NEGBINOMDIST(number_f,number_s,probability_s)

Number_f: 失败次数。

Number_s: 成功的临界次数。

Probability_s: 成功的概率。

3 . 举例 : 例如 , 如果要找出 5 个反应敏捷的人 , 且已知具有这种特征的候选人的概率为 0.3。以下公式将计算出在找到 5 个合格候选人之前 , 需要面试 10 个候选人的概率 :

NEGBINOMDIST(10,5,0.3)= 0.06871

(四) 函数 POISSON:

1. 说明：函数 POISSON 返回泊松分布。泊松分布通常用于预测一段时间内事件发生指定次数的概率，比如一分钟内通过收费站的轿车的数量为 n 的概率。
2. 语法：POISSON($x, mean, cumulative$)
X：事件数。
Mean：期望值。
Cumulative：为一逻辑值，确定所返回的概率分布形式。如果 $cumulative$ 为 TRUE，函数 POISSON 返回累积分布函数，即，随机事件发生的次数在 0 和 x 之间(包含 0 和 1)；如果为 FALSE，则返回概率密度函数，即，随机事件发生的次数恰好为 x 。
3. 举例：POISSON(2,5,FALSE)=0.084224 表明，若某一收费站每分钟通过的轿车平均数量为 5 辆，那么某一分钟通过 2 辆的概率为 0.084224。

(五) 正态分布函数 NORMDIST：

1. 说明：正态分布在模拟现实世界过程和描述随机样本平均值的不确定度时有广泛的用途。函数 NORMDIST 返回给定平均值和标准偏差的正态分布的累积函数。同样可以用类似“七”中的方法，利用 NORMDIST 函数建立正态分布密度函数图，这里不再赘述。
2. 语法：NORMDIST($x, mean, standard_dev, cumulative$)
X：为需要计算其分布的数值。
Mean：分布的算术平均值。
Standard_dev：分布的标准偏差。
Cumulative：为一逻辑值，指明函数的形式。如果 $cumulative$ 为 TRUE，函数 NORMDIST 返回累积分布函数；如果为 FALSE，返回概率密度函数。
2. 举例：例如，公式 NORMDIST(6,5,2,0) 返回平均值为 5、标准差为 2 的正态函数当 $X=6$ 时概率密度函数的数值，公式 NORMDIST(60, 50, 4, 1) 返回平均值为 50、标准差为 4 的正态分布函数当 $X=60$ 时累积分布函数的数值。

(六) 函数 NORMSDIST：

1. 说明：函数 NORMSDIST 返回标准正态分布的累积函数。
2. 语法：NORMSDIST(z)
 Z 为需要计算其分布的数值。
3. 举例：NORMSDIST(0)=0.5

(七) 函数 NORMSINV :

1. 说明：函数 NORMSINV 返回标准正态分布累积函数的逆函数。
2. 语法：NORMSINV(probability)
Probability：正态分布的概率值。
3. 举例：NORMSINV(0.5)=0

(八) t 分布函数 TDIST:

1. 说明：函数 TDIST 返回 student 的 t 分布数值。T 分布用于小样本数据集合的假设检验。使用此函数可以代替 t 分布的临界值表。
2. 语法：TDIST(x,degrees_freedom,tails)
X：为需要计算分布的数字。
Degrees_freedom：为表示自由度的整数。
Tails：指明返回的分布函数是单尾分布还是双尾分布。如果 tails = 1 , 函数 TDIST 返回单尾分布。如果 tails = 2 , 函数 TDIST 返回双尾分布。
3. 举例：TDIST(1.96,60,2)=0.054645

三、随机抽样的工具

- (一) 简介：Excel 中的 Rand()函数可以返回大于等于 0 小于 1 的均匀分布随机数，Rand()不带任何参数运行，每次计算时时都将返回一个新的数值。RAND()函数可以被用来作为不重复抽样调查的工具。
- (二) 操作步骤：如图附-11 所示有 10 个象征性的样本数据，欲从中随机抽取 5 个数据可按如下步骤操作：

	A
1	No
2	1
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9
11	10

图 附-11

1. 选择 B2 单元格，输入公式 “=RAND()” 并回车。
2. 拖动 B2 单元格右下角的填充柄至 B11 单元格，并在 B1 单元格

输入标题“RANDOM”。

3. 选取单元格 B2 至 B11，右击选中的区域选择“复制”，再次右击选中的区域，选择“选择性粘贴”，单击选项“数值”后，点击“确定”按钮。
4. 选取单元格 A2 至 B11 单元格，选择“数据”菜单项下的排序子菜单。
5. 选取“RANDOM”为主要关键字，然后点击“确定”按钮。排序结果如图附-12 所示，A2 至 A6 单元格的样本即为随机抽取的 5 个样本。

	A	B
1	No	random
2	10	0.216688
3	3	0.234034
4	1	0.302342
5	8	0.437267
6	9	0.610631
7	7	0.64232
8	4	0.656722
9	2	0.68924
10	5	0.882674
11	6	0.953918

图 附-12

（三）结果说明：

1. 以上进行的是不重复随机抽样，可以用类似的方法，利用 Excel 的 RANDBETWEEN(TOP,BOTTOM)函数实现总体的重复随机抽样。RANDBETWEEN(TOP,BOTTOM)函数可随机返回介于 TOP 与 BOTTOM 之间的整数，抽取此整数对编号的样本可作为总体的重复随机抽样的结果。
2. RAND()函数返回的是 0 与 1 之间均匀的随机数，利用 Excel 数据分析工具中的随机数发生器，可以生成用户指定类型分布的随机数。例如 0-1 正态分布的随机数，指定 γ 参数的迫松分布的随机数等。
3. 利用 Excel 易于产生各类型随机数的特性，可以用类似的方法方便的进行进行随机数字模拟试验与随机游走模拟试验。

四、样本推断总体

(一) 简介：利用 Excel 的几个函数，如求平均函数 AVERAGE、标准差函数 STDEV、T 分布函数 TINV 等的组合使用可以构造出一个专门用于实现样本推断总体的 Excel 工作表。以下例子先计算样本的平均数和标准差，然后在一定置信水平上估计总体均值的区间范围。

(二) 操作步骤：

1. 构造工作表。如图附-13 所示，首先在各个单元格输入以下内容，其中左边是变量名，右边是相应的计算公式。

2. 为表格右边的公式计算结果定义左边的变量名。选定 A4:B6,A8:B8 和 A10:B15 单元格(先选择第一部分，再按住 CTRL 键选取另外两个部分)，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“最左列”选项,然后点击“确定”按钮即可。

	A	B
1		以样本均值推断总体均值的置信区间
2		
3	样本统计量	
4	样本个数	=COUNT(样本数据)
5	样本均值	=AVERAGE(样本数据)
6	样本标准差	=STDEV(样本数据)
7	用户输入	
8	置信水平	0.95
9	计算结果	
10	抽样标准误	= '样本标准差' / SQRT('样本个数')
11	自由度	= '样本个数' - 1
12	t值	=TINV(1-'置信水平', '自由度')
13	置信区间半径	= 't值' * '抽样标准误'
14	置信区间上界	= '样本均值' - '置信区间半径'
15	置信区间下界	= '样本均值' + '置信区间半径'

图 附-13

3. 输入样本数据，和用户指定的置信水平 0.95,如图附-13 所示。
4. 为样本数据命名。选定 D1:D11 单元格，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“首行”选项,然后点击“确定”按钮，得到图附-14 所示的计算结果。

	A	B	C	D
1	以样本均值推断总体均值的置信区间			样本数据
2				28.5
3	样本统计量			26.4
4	样本个数	10		33.5
5	样本均值	31.4		34.3
6	样本标准差	2.814249456		35.9
7	用户输入			29.6
8	置信水平	0.95		31.3
9	计算结果			31.1
10	抽样标准误	0.889943818		30.9
11	自由度	9		32.5
12	t值	2.262158887		
13	置信区间半径	2.013194318		
14	置信区间上界	29.38680568		
15	置信区间下界	33.41319432		

图 附-14

(三) 结果说明：以上例子说明如何交叉组合使用 Excel 的公式和函数，以构造出一个能实现样本推断总体有关计算的 Excel 工作表。实际上，在用 Excel 进行数据统计处理之时，许多统计功能可以使用和上例类似的方法，通过组合使用 Excel 的各类统计函数和公式加以实现的。

五、假设检验

(一) 简介：假设检验是统计推断中的重要内容。以下例子利用 Excel 的正态分布函数 NORMSDIST、判断函数 IF 等，构造一张能够实现在总体方差已知情况下进行总体均值假设检验的 Excel 工作表。

(二) 操作步骤：

1. 构造工作表。如图附-15 所示，首先在各个单元格输入以下的内容，其中左边是变量名，右边是相应的计算公式。

2. 为表格右边的公式计算结果定义左边的变量名。选定 A3:B4, A6:B8, A10:A11, A13:A15 和 A17:B19 单元格，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“最左列”选项，然后点击“确定”按钮即可。

	A	B
1	总体均值的假设检验	
2	样本统计量	
3	样本个数	=COUNT(样本数据)
4	样本均值	=AVERAGE(样本数据)
5	用户输入	
6	总体标准差	
7	总体均值假设值	
8	置信水平	
9	计算结果	
10	抽样标准误	= '总体标准差' / SQRT('样本个数')
11	计算Z值	= ('样本均值' - '总体均值假设值') / '抽样标准误'
12	单侧检验	
13	单侧Z值	= NORMSINV(1 - '置信水平')
14	检验结果	= IF(ABS('计算Z值') > ABS('单侧Z值'), "拒绝H ₀ ", "接收H ₀ ")
15	单侧显著水平	= 1 - NORMSDIST(ABS('计算Z值'))
16	双侧检验	
17	双侧Z值	= NORMSINV((1 - '置信水平') / 2)
18	检验结果	= IF(ABS('计算Z值') > ABS('双侧Z值'), "拒绝H ₀ ", "接收H ₀ ")
19	双侧显著水平	= IF('计算Z值' > 0, 2 * (1 - NORMSDIST('计算Z值')), 2 * NORMSDIST('计算Z值'))

图 附-15

2. 输入样本数据，以及总体标准差、总体均值假设、置信水平数据。如图附-16 所示。
3. 为样本数据命名。选定 C1:C11 单元格，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“首行”选项，然后点击“确定”按钮，得到如图附-16 中所示的计算结果。

	A	B	C
1	总体均值的假设检验		样本数据
2	样本统计量		28.5
3	样本个数	10	26.4
4	样本均值	31.4	33.5
5	用户输入		34.3
6	总体标准差	5.56	35.9
7	总体均值假设值	35	29.6
8	置信水平	0.95	31.3
9	计算结果		31.1
10	抽样标准误	1.758226379	30.9
11	计算Z值	-2.047517909	32.5
12	单侧检验		
13	单侧Z值	-1.644853	
14	检验结果	拒绝Ho	
15	单侧显著水平	0.020303562	
16	双侧检验		
17	双侧Z值	-1.959961082	
18	检验结果	拒绝Ho	
19	双侧显著水平	0.040607125	

图 附-16

(三) 结果说明：如图附-16 所示，该例子的检验结果不论是单侧还是双侧均为拒绝 H_0 假设。所以，根据样本的计算结果，在 5% 的显著水平之下，拒绝总体均值为 35 的假设。同时由单侧显著水平的计算结果还可以看出，在总体均值是 35 的假设之下，样本均值小于等于 31.4 的概率仅为 0.020303562。

六、双样本等均值假设检验

(一) 简介：双样本等均值检验是在一定置信水平之下，在两个总体方差相等的假设之下，检验两个总体均值的差值等于指定平均差的假设是否成立的检验。我们可以直接使用在 Excel 数据分析中提供双样本等均值假设检验工具进行假设检验。以下通过一例说明双样本等均值假设检验的操作步骤。例子如下，某工厂为了比较两种装配方法的效率，分别组织了两组员工，每组 9 人，一组采用新的装配方法，另外一组采用旧的装配方法。18 个员工的设备装配时间图附-17 中表格所示。根据以下数据，是否有理由认为新的装配方法更节约时间？

	A	B	C	D
1	组别	旧方法装配时间	组别	新方法装配时间
2	1	32	2	35
3	1	37	2	31
4	1	35	2	29
5	1	38	2	25
6	1	41	2	34
7	1	44	2	40
8	1	35	2	27
9	1	31	2	32
10	1	34	2	31

图 附-17

(二) 操作步骤：以上例子可按如下步骤进行假设检验。

1. 选择“工具”菜单的“数据分析”子菜单，双击“t-检验：双样本等方差假设”选项，则弹出图附-18 所示对话框。

t-检验：双样本等方差假设

输入

变量 1 的区域(1): \$B\$1:\$B\$10

变量 2 的区域(2): \$D\$1:\$D\$10

假设平均差(E): 0

☒ 标志(L)

α (A): 0.05

输出选项

☐ 输出区域(O):

☒ 新工作表组(P):

☐ 新工作簿(W)

确定 取消 帮助(H)

图 附-18

2. 分别填写变量 1 的区域 \$B\$1:\$B\$10 ,变量 2 的区域 \$D\$1:\$D\$10 , 由于我们进行的是等均值的检验，填写假设平均差为 0，由于数据的首行包括标志项选择标志选项，所以选择“标志”选项，再填写显著水平 α 为 0.05,然后点击“确定”按钮。则可以得到图附-19 所示的结果。

(三) 结果分析：如图附-19 中所示，表中分别给出了两组装配时间的平均值、方差和样本个数。其中，合并方差是样本方差加权之后的平均值,Df 是假设检验的自由度它等于样本总个数减 2，t 统计量是两个样

	A	B	C
1	t-检验：双样本等方差假设		
2			
3		旧方法装配时间	新方法装配时间
4	平均	35.22222222	31.55555556
5	方差	24.44444444	20.02777778
6	观测值	9	9
7	合并方差	22.23611111	
8	假设平均差	0	
9	df	16	
10	t Stat	1.649484617	
11	P(T<=t) 单尾	0.059269899	
12	t 单尾临界	1.745884219	
13	P(T<=t) 双尾	0.118539799	
14	t 双尾临界	2.119904821	

本差值减去

图 附-19

假设平均差之后再除以标准误差的结果，“ $P(T \leq t)$ 单尾”是单尾检验的显著水平，“t 单尾临界”是单尾检验 t 的临界值，“ $P(T \leq t)$ 双尾”是双尾检验的显著水平，“t 双尾临界”是双尾检验 t 的临界值。由下表的结果可以看出 t 统计量均小于两个临界值，所以，在 5% 显著水平下，不能拒绝两个总体均值相等的假设，即两种装配方法所耗时间没有显著的不同。

Excel 中还提供了以下类似的假设检验的数据分析工具，它们的名称和作用分别罗列如下：

1. “t-检验：双样本异方差假设”：此分析工具可以进行双样本 student t-检验，与双样本等方差假设检验不同，该检验是在两个数据集的方差不等的前提假设之下进行两总体均值差额的检验，故也称作异方差 t-检验。可以使用 t-检验来确定两个样本均值实际上是否相等。当进行分析的样本个数不同时，可使用此检验。如果某一样本组在某次处理前后都进行了检验，则应使用“成对检验”。
2. “t-检验：成对双样本均值分析”：此分析工具可以进行成对双样本学生氏 t-检验，用来确定样本均值是否不等。此 t-检验并不假设两个总体的方差是相等的。当样本中出现自然配对的观察值时，可以使用此成对检验，例如，对一个样本组进行了两次检验，抽取实验前的一次和实验后的一次。
3. “z - 检验：双样本均值分析”：此分析工具可以进行方差已知的双样本均值 z - 检验。此工具用于检验两个总体均值之间存在差异的假设。例如，可以使用此检验来确定两种汽车模型性能之间

的差异情况。

七、正态性的 X^2 检验

(一) 简介： X^2 检验可以用来判断所观测的样本是否来自某一特定分布的总体，这种检验亦称为一致性检验。以下例子，已知某样本的相关统计量和分组频数分布如图附-20 所示，试图用 X^2 检验判断该样本是否来自一正态总体。

	A	B	C
1	正态性的 X^2 检验		
2			
3	样本个数	样本均值	样本标准差
4	200	164	10
5			
6	分组下界	分组上界	真实频数
7		150	15
8	150	160	54
9	160	170	78
10	170	180	42
11	180		11
12		累积值	200

图 附-20

(二) 操作步骤

1. 创建变量名。选定 A3:C4 单元格，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“首行”选项，然后点击“确定”按钮即可。
2. 计算预期正态概率值。如图附-21 表中所示，在 D6 单元格输入标志项，在 D7:D11 单元格输入公式，分别计算各组的预期正态概率值，在 D12 计算累积概率值。

	D
6	预期正态概率
7	=NORMDIST(B7, 样本均值, 样本标准差, 1)
8	=NORMDIST(B8, 样本均值, 样本标准差, 1)-NORMDIST(A8, 样本均值, 样本标准差, 1)
9	=NORMDIST(B9, 样本均值, 样本标准差, 1)-NORMDIST(A9, 样本均值, 样本标准差, 1)
10	=NORMDIST(B10, 样本均值, 样本标准差, 1)-NORMDIST(A10, 样本均值, 样本标准差, 1)
11	=1-NORMDIST(A11, 样本均值, 样本标准差, 1)
12	=SUM(D7:D11)

图 附-21

3. 计算预期频数值。如图附-22 所示，在 E6 单元格输入标志项，在 E7:E11 单元格输入公式，分别计算各组的预期频数，在 E12 计算

累积频数值。

	E
6	预期频数值
7	=D7*样本个数
8	=D8*样本个数
9	=D9*样本个数
10	=D10*样本个数
11	=D11*样本个数
12	=SUM(E7:E11)

图 附-22

4. 计算 X^2 统计量。如图附-23 所示，在 F6 单元格输入标志项，在 F7:F11 分别输入计算公式，分别计算 X^2 值，在 E12 计算 X^2 平方和，这项就是最后计算出的 X^2 统计量。在 E13 单元格输入标志项“卡方统计量”，为以后的引用作准备。先选中 F12、F13 两个单元格，选择“插入”菜单的“名称”子菜单的“指定”选项，用鼠标点击“尾行”选项，然后点击“确定”按钮即可。

	F
1	
2	
3	
4	
5	
6	卡方值
7	=(C7-E7)^2/E7
8	=(C8-E8)^2/E8
9	=(C9-E9)^2/E9
10	=(C10-E10)^2/E10
11	=(C11-E11)^2/E11
12	=SUM(F7:F11)
13	卡方统计量

图 附-23

5. 如图附-24 所示，分别在 A14 到 B20 单元格输入自由度、 X^2 概率值、置信水平、临界值、 X^2 检验结果几项的标志值及计算公式。其中的自由度 = 区间分段数 - 正太分布参数个数 - 1 = 5 - 2 - 1 = 2。

	A	B
14	自由度	2
15	卡方概率值	=CHIDIST(卡方统计量,'自由度')
16		
17	置信水平	0.01
18	临界值	=CHIINV('置信水平','自由度')
19	卡方检验结果	=IF(卡方统计量>临界值,"拒绝总体为正太分布的假设","接受总体为正太分布的假设")

图 附-24

(三) 结果分析：如图附-25 所示，按照以上操作步骤可以得到表中的计算结果。

	A	B	C	D	F
1		正态性的 χ^2 检验			
2					
3	样本个数	样本均值	样本标准差		
4	200	164	10		
5					
6	分组下界	分组上界	真实频数	预期正态概率	卡方值
7		150	15	0.080756711	0.082073
8	150	160	54	0.263821592	0.0289383
9	160	170	78	0.381168632	0.0409231
10	170	180	42	0.219453775	0.0814512
11	180		11	0.054799289	0.000147
12		累积值	200	1	0.2335326
13					卡方统计量
14	自由度	2			
15	卡方概率值	0.889793097			
16					
17	置信水平	0.01			
18	临界值	9.210351036			
19	卡方检验结果	接受总体为正态分布的假设			

图 附-25

按同样的方法可以作总体泊松分布、总体超几何分布等其它分布的检验。此类统计应用也是由 Excel 各类公式和函数综合使用而实现的,为了以后使用方便,和上面的一些例子一样,一般需要将整个表格的计算框架和标志项罗列好,再保存成文件,以后只要对数据项稍作修改即可很快得到计算结果。如果对 Excel 宏语言较为熟悉,还可以将它编成一个宏语言程序,加入 Excel 的工具栏,这样以后使用起来更为方便。

八、列联表分析

(一) 简介：列联表分析经常用来判断同一个调查的对象两个特性之间是否存在明显相关关系。例如,房地产商常常设计列联表问卷,调查顾客的职业和顾客所选房子的类型是否有明显的相关关系。列联表分析同样也可以

由 Excel 加以实现，下面用一个例子给予说明。如图附-26 所示，表中是某装修公司的调查报告，试用列联表分析方法分析在顾客的所在地区和所选房子的地板类型之间是否存在明显的相关关系。

	A	B	C	D	E	F
1	列联表分析					
2						
3	真实频数	地区1	地区2	地区3	地区4	行总数
4	木质地板	72	8	12	23	115
5	钢砖地板	26	10	16	33	85
6	大理石地板	7	10	14	19	50
7	列总数	105	28	42	75	250

图 附-26

(二) 操作步骤

1. 建立期望频数表。如图附-27 所示，先建立期望频数表的框架，然后在 B10 单元格输入公式 “=B\$7*\$F4/\$F\$7”，再利用“填充柄”将公式复制到表格的其它单元格，最后利用 Excel 的求和函数 sum 计算行和与列和。

	A	B	C	D	E	F
9	期望频数	地区1	地区2	地区3	地区4	总数
10	木质地板	=B\$7*\$F4/\$F\$7	=C\$7*\$F4/\$F\$7	=D\$7*\$F4/\$F\$7	=E\$7*\$F4/\$F\$7	=SUM(B10:E10)
11	钢砖地板	=B\$7*\$F5/\$F\$7	=C\$7*\$F5/\$F\$7	=D\$7*\$F5/\$F\$7	=E\$7*\$F5/\$F\$7	=SUM(B11:E11)
12	大理石地板	=B\$7*\$F6/\$F\$7	=C\$7*\$F6/\$F\$7	=D\$7*\$F6/\$F\$7	=E\$7*\$F6/\$F\$7	=SUM(B12:E12)
13	总数	=SUM(B10:B12)	=SUM(C10:C12)	=SUM(D10:D12)	=SUM(E10:E12)	=SUM(B13:E13)

图 附-27

2. 计算 X^2 概率值。在 A15 单元格输入标志项“卡方概率值”，先点击 B15 单元格，从“插入”菜单中“函数”子菜单，选择“统计函数”中的“CHITEST”函数，单击“确定按钮”，然后在弹出的对话框中分别添入实际频数范围“B4:E6”和预期频数范围“B10:E12”。最后单击“确定”按钮即可得到计算结果 1.3E-07，如图附-28 所示。

	A	B	C	D	E	F
1	列联表分析					
2						
3	真实频数	地区1	地区2	地区3	地区4	行总数
4	木质地板	72	8	12	23	115
5	钢砖地板	26	10	16	33	85
6	大理石地板	7	10	14	19	50
7	列总数	105	28	42	75	250
8						
9	期望频数	地区1	地区2	地区3	地区4	总数
10	木质地板	48.3	12.88	19.32	34.5	115
11	钢砖地板	35.7	9.52	14.28	25.5	85
12	大理石地板	21	5.6	8.4	15	50
13	总数	105	28	42	75	250
14						
15	卡方概率值	1E-07				

图 附-28

3. 建立 X^2 统计表。如图附-29 所示,先建立表格的框架,然后在 B20 单元格输入公式 “ $=(B4-B10)^2/B10$ ”,再利用填充柄将公式复制到表格的其它单元格。最后计算 X^2 卡方统计量,分别在 A24 与 B24 单元输入标志项与计算公式。

	A	B	C	D	E
19		地区1	地区2	地区3	地区4
20	木质地板	$=(B4-B10)^2/B10$	$=(C4-C10)^2/C10$	$=(D4-D10)^2/D10$	$=(E4-E10)^2/E10$
21	钢砖地板	$=(B5-B11)^2/B11$	$=(C5-C11)^2/C11$	$=(D5-D11)^2/D11$	$=(E5-E11)^2/E11$
22	大理石地板	$=(B6-B12)^2/B12$	$=(C6-C12)^2/C12$	$=(D6-D12)^2/D12$	$=(E6-E12)^2/E12$
23					
24	卡方统计量	$=SUM(B20:E22)$			

图 附-29

4. 进行假设检验。如图附-30 所示,分别输入置信水平、临界值和假设检验的结果其中 CHINV 函数的自由度=(第一类属性的分类数-1)*(第二类属性的分类数-1)=(3-1)*(4-1)=6。

	A	B
26	置信水平	0.01
27	临界值	$=CHINV(B26,6)$
28	检验结果	$=IF(卡方统计量>临界值,“拒绝两种属性不相关的假设”,“接受两种属性不相关的假设”)$

图 附-30

(三) 结果分析:以上的操作步骤完成整个列联表的分析。其中,B15 单元格的卡方概率值与 B24 单元格的卡方统计量是表格的两个重要计算结果。其中卡方概率值等于 1.3E-07 表明:如果总体的两

类属性，即所在地区和所选地板类型，是不相关的，那么得到以上观察的样本的概率是 0.00000013。这个概率几乎接近于 0，所以可以认为总体的这两个属性是显著相关的。

九、单因素方差分析

(一) 简介：单因素方差分析可用于检验两个或两个以上的总体均值相等的假设是否成立。此方法是对双均值检验（如 t-检验）的扩充。检验假定总体是服从正太分布的，总体方差是相等的，并且随机样本是独立的。这种工具试用于完全随机化试验的结果分析。例子如图附-31 表中所示，一产品制造商雇佣销售人员向销售商打电话。制造商想比较四种不同电话频率计划的效率，他从销售人员中随机选出 32 名，将他们随机分配到 4 种计划中，在一段时期内记录他们的销售情况已经在表中列出，试问其中是否有一种计划会带来较高的销售水平。

	A	B	C	D
1	单因素方差分析			
2				
3	计划1	计划2	计划3	计划4
4	36	39	44	31
5	40	45	43	43
6	32	54	38	46
7	44	53	40	43
8	35	46	41	36
9	41	42	35	49
10	44	35	37	46
11	42	39	37	48

图 附-31

(二) 操作步骤

1. 选择“工具”菜单的“数据分析”子菜单，双击“方差分析：单因素方差分析”选项，弹出单因素方差分析对话框。
2. 按图附-32 所示方式填写对话框。然后单击“确定”按钮。

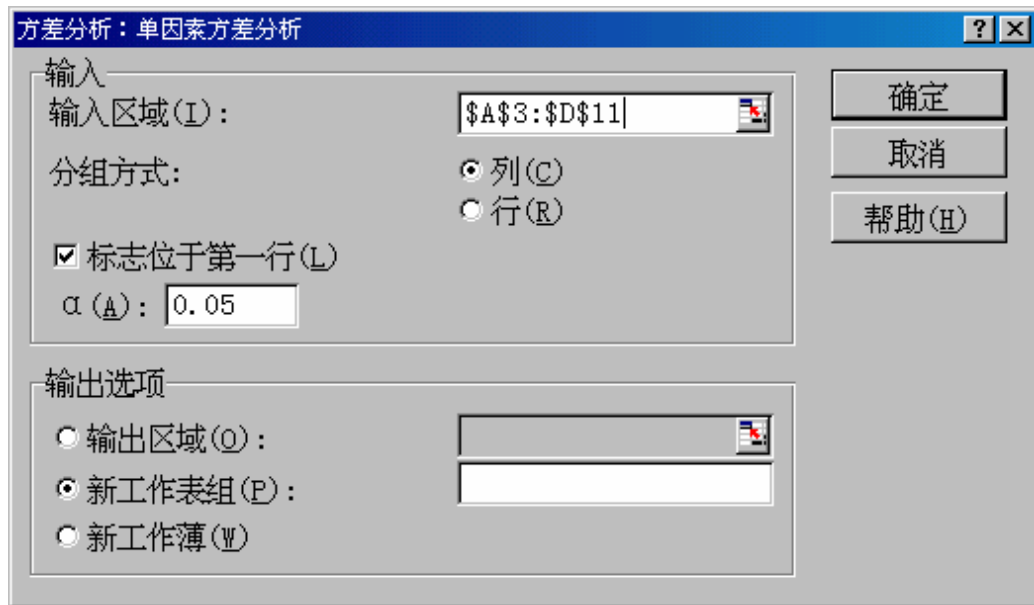


图 附-32

(三) 结果分析：按照如上的操作步骤即可得到图附-33 的计算结果。其中表格的第二部分则是方差分析的结果。SS 列分别给出了四个分组的组间方差、组内方差以及总方差，DF 列分别给出了对应方差的自由度，MS 列是平均值方差，由 SS 除于 DF 得到，它是总体方差的两个估计值。F 列是 F 统计量的计算结果，如果四个总体均值相等的假设成立的化，它应该服从 F 分布，即近似为 1，它是最终的计算结果，通过将

	A	B	C	D	E	F	G
1	方差分析：单因素方差分析						
2							
3	SUMMARY						
4	组	计数	求和	平均	方差		
5	计划1	8	314	39.25	19.6429		
6	计划2	8	353	44.125	45.8393		
7	计划3	8	315	39.375	9.98214		
8	计划4	8	342	42.75	38.7857		
9							
10							
11	方差分析						
12	差异源	SS	df	MS	F	P-value	F crit
13	组间	143.75	3	47.9167	1.67761	0.19442	2.94668
14	组内	799.75	28	28.5625			
15							
16	总计	943.5	31				

图 附-33

它与一定置信水平下的 F 临界值 F_{crit} 比较，可以判断均值相等的假设是否成立，在本例中， $1.67761 < 2.94668$ ，所以不能拒绝四个总

体均值相等的假设。P-value 列,是单尾概率值,表明如果四个总体均值相等的假设成立的化,得到如上样本结果的概率是 19.442%,即得到以上样本并不是小概率事件,同样也得到不能拒绝四个总体均值相等的假设的结论。

按相似方法可进行无重复双因素方差分析,有重复双因素方差分析。

十、线性回归分析

(一) 简介:线性回归分析通过对一组观察值使用“最小二乘法”直线拟合,用来分析单个因变量是如何受一个或几个自变量影响的。例子如图附-34 所示,表中是我国 1987 年至 1997 年的布匹人均产量和人均纱产量,试用线性回归分析的方法分析两组数据之间的关系。

	A	B	C
1	年份	人均布产量(米)	人均纱产量(公斤)
2	1987	15.96	4.03
3	1988	17.06	4.23
4	1989	16.92	4.26
5	1990	16.63	4.07
6	1991	15.79	4
7	1992	16.37	4.31
8	1993	17.23	4.26
9	1994	17.73	4.11
10	1995	21.59	4.5
11	1996	17.17	4.21
12	1997	20.23	4.55

图 附-34

(二) 操作步骤

1. 选择“工具”菜单的“数据分析”子菜单,双击“回归”选项,弹出回归分析对话框。其中主要选项的含义如下:Y 值输入区域,在此输入对因变量数据区域,该区域必须由单列数据组成;X 值输入区域,在此输入对自变量数据区域,Excel 将对此区域中的自变量从左到右按升序排列,自变量的个数最多为 16;置信度,如果需要在汇总输出表中包含附加的置信度信息,则选中此复选框,然后在右侧的编辑框中,输入所要使用的置信度,95%为默认值;常数为零,如果要强制回归线通过原点,则选中此复选框;输出区域,在此输入对输出表左上角单元格的引用。汇总输出表至少需要有七列的宽度,包含的内容有 anova 表、系数、y 估计值的标准误差、 r^2 值、观察值个数,以及系数的标准误差;新工作表,单击

此选项，可在当前工作簿中插入新工作表，并由新工作表的 A1 单元格开始粘贴计算结果，如果需要给新工作表命名，则在右侧的编辑框中键入名称；新工作簿，单击此选项，可创建一新工作簿，并在新工作簿中的新工作表中粘贴计算结果；残差，如果需要以残差输出表的形式查看残差，则选中此复选框；标准残差，如果需要在残差输出表中包含标准残差，则选中此复选框；残差图，如果需要生成一张图表，绘制每个自变量及其残差，则选中此复选框；线形拟合图，如果需要为预测值和观察值生成一个图表，则选中此复选框；正态概率图，如果需要绘制正态概率图，则选中此复选框。

2. 按如下方式填写对话框：X 值输入区域为 \$B\$1:\$B\$12, Y 值输入区域为 \$C\$1:\$C\$12, 并选择“标志”和“线性拟合图”两个复选框，然后单击“确定”按钮即可。

（三）结果分析

按照如上的操作步骤即可得到图附-35 下表的计算结果。结果可以分为四个部分，第一部分是回归统计的结果包括多元相关系数、可决系数 R^2 、调整之后的相关系数、回归标准差以及样本个数。第二部分是方差分析的结果包括可解释的离差、残差、总离差和它们的自由度以及由此计算出的 F 统计量和相应的显著水平。第三部分是回归方程的截距和斜率的估计值以及它们的估计标准误差、t 统计量大小双边拖尾概率值、以及估计值的上下界。根据这部分的结果可知回归方程为 $Y=8.46433 \cdot X-18.288$ 。第四部分是样本散点图，其中蓝色的点是样本的真实散点图，红色的点是根据回归方程进行样本历史模拟的散点。如果觉得散点图不够清晰可以用鼠标拖动图形的边界达到控制图形大小的目的。用相同的方法可以进行多元线性方程的参数估计，还可以在自变量中引入虚拟变量以增加方程的拟合程度。对于非线性方程的参数

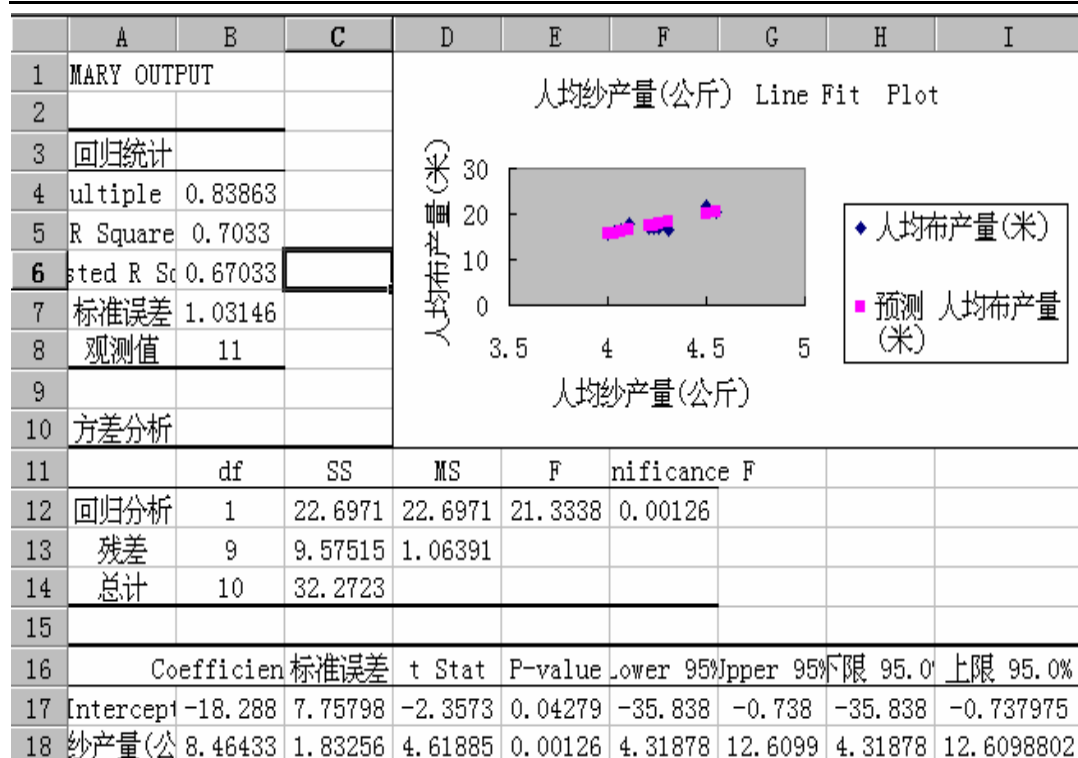


图 附-35

估计，可以在进行样本数据的线性化处理之后，再按以上步骤进行参数估计。

十一、相关系数分析工具

(一) 简介：此分析工具可用于判断两组数据之间的关系。可以使用“相关系数”分析工具来确定两个区域中数据的变化是否相关，即，一个集合的较大数据是否与另一个集合的较大数据相对应(正相关)；或者一个集合的较小数据是否与另一个集合的较小数据相对应(负相关)；还是两个集合中的数据互不相关(相关性为零)。

(二) 操作步骤：采用图附-3 表中的数据，可按如下步骤计算变量 x, y, z 之间的相关系数。

1. 用鼠标点击表中待分析数据的任一单元格。
2. 选择“工具”菜单的“数据分析”子菜单。
3. 用鼠标双击数据分析工具中的“相关系数”选项。
4. 填写完“相关系数”对话框，单击“确定”按钮即可得到各个变量的相关系数矩阵，结果如图附-36 所示。

	A	B	C	D
1		x	y	z
2	x	1		
3	y	0.929167	1	
4	z	0.922982	0.984245	1

图 附-36

(三) 结果说明：以上下三角矩阵计算出三个变量 x, y, z 两两之间的相关系数，如变量 x, y 之间的相关系数为：0.929167, 所以可以判断 x, y 之间存在着较高的正线性相关关系。

十二、协方差分析工具

协方差分析的操作步骤同“六”的相关系数分析较为相似，只须在第 3 步中将“相关系数”选项替换成为“协方差”选项即可。

十三、自回归模型的识别与估计

(一) 简介：时间序列分析已广泛的应用于科研，生产，社会生活的方方面面。它通过对时间序列作统计规律性的分析，构造出拟合时间序列的最佳模型。构造出的时间序列模型一方面而浓缩了时间序列的信息，简化对时间序列的表示，另一方面可以用来预测时间序列未来的可能取值，可以作为人们科学决策的重要依据。例子如图附-37 所示，表中是自 1999 年 4 月 1 日起的 20 个交易日内的上证指数的时间序列，试用自回归模型加以拟合。

(二) 操作步骤

1. 数据的零均值化处理。如图附-37 中所示，在 C1 中输入序列名“Z”，在 C2 中输入公式“=上证指数-AVERAGE(上证指数)”，然后在 C2 单元格中，拖动 Excel“填充柄”将公式复制到 C3 至 C22 单元格，即可生成上证指数的零均值化序列。
2. 计算自相关函数。在 E1 和 F1 单元格分别输入标志项 Lag 和 ac，

	A	B	C
1	日期	上证指数	Z
2	36251	1168.5	=上证指数-AVERAGE(上证指数)
3	36252	1184.4	=上证指数-AVERAGE(上证指数)
4	36255	1194.51	=上证指数-AVERAGE(上证指数)
5	36256	1199.6	=上证指数-AVERAGE(上证指数)
6	36257	1201.63	=上证指数-AVERAGE(上证指数)
7	36258	1202.47	=上证指数-AVERAGE(上证指数)
8	36259	1205.14	=上证指数-AVERAGE(上证指数)
9	36262	1195.34	=上证指数-AVERAGE(上证指数)
10	36263	1178.19	=上证指数-AVERAGE(上证指数)
11	36264	1181.52	=上证指数-AVERAGE(上证指数)
12	36265	1170.62	=上证指数-AVERAGE(上证指数)
13	36266	1166.24	=上证指数-AVERAGE(上证指数)
14	36269	1159.47	=上证指数-AVERAGE(上证指数)
15	36270	1171.6	=上证指数-AVERAGE(上证指数)
16	36271	1144.06	=上证指数-AVERAGE(上证指数)
17	36272	1137.16	=上证指数-AVERAGE(上证指数)
18	36273	1140	=上证指数-AVERAGE(上证指数)
19	36276	1112.79	=上证指数-AVERAGE(上证指数)
20	36277	1091.69	=上证指数-AVERAGE(上证指数)
21	36278	1091.09	=上证指数-AVERAGE(上证指数)

图 附-37

在 E2 到 E9 单元格中分别输入置后期数 1 至 8。在 F2 单元格输入计算自相关函数的公式“=SUMPRODUCT(OFFSET(C\$2,0,0,20-E2), OFFSET(C3,0,0,20-E2))/VAR(\$C\$2:\$C\$21)/19”，然后利用“填充柄”将 F2 单元格公式复制到 F3:F9 单元格，结果如图附-38 所示。

	E	F
1	Lag	ac
2	1	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E2), OFFSET(C3,0,0,20-E2))/VAR(\$C\$2:\$C\$21)/19
3	2	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E3), OFFSET(C4,0,0,20-E3))/VAR(\$C\$2:\$C\$21)/19
4	3	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E4), OFFSET(C5,0,0,20-E4))/VAR(\$C\$2:\$C\$21)/19
5	4	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E5), OFFSET(C6,0,0,20-E5))/VAR(\$C\$2:\$C\$21)/19
6	5	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E6), OFFSET(C7,0,0,20-E6))/VAR(\$C\$2:\$C\$21)/19
7	6	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E7), OFFSET(C8,0,0,20-E7))/VAR(\$C\$2:\$C\$21)/19
8	7	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E8), OFFSET(C9,0,0,20-E8))/VAR(\$C\$2:\$C\$21)/19
9	8	=SUMPRODUCT(OFFSET(C\$2,0,0,20-E9), OFFSET(C10,0,0,20-E9))/VAR(\$C\$2:\$C\$21)/19

图 附-38

2. 计算偏自相关函数。计算偏自相关函数的步骤较为复杂，必须利用 Excel 的逆矩阵等函数求解 Yule-Walker 方程组，由于我们选择了置后期数为 8，为了求解偏自相关函数，我们必须求解 8 个 Yule-Walker 方程组。首先，利用自相关函数的计算结果，填写 H2:O9 范围内

的对称矩阵如图附-39 中 H2:O9 单元格所示。其次，利用 Excel 数组公式分别求解 8 个方程组的结果，结果分别放在 ϕ_{1i} 至 ϕ_{8i} 的 8 列之中，第一个方程组的结果放在 H12 中，第二个方程组的结果放在 I12:I13 中，第三个方程组的结果放在 J12:J14 中，以此类推。

所输入的 8 个数组公式分别为：

“ MMULT(MINVERSE(OFFSET(H2,0,0,1,1)),OFFSET(F2,0,0,1)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,2,2)),OFFSET(F2,0,0,2)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,3,3)),OFFSET(F2,0,0,3)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,4,4)),OFFSET(F2,0,0,4)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,5,5)),OFFSET(F2,0,0,5)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,6,6)),OFFSET(F2,0,0,6)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,7,7)),OFFSET(F2,0,0,7)) ”

,

“ MMULT(MINVERSE(OFFSET(H2,0,0,8,8)),OFFSET(F2,0,0,8)) ”

。

(说明 1.在 Excel 中输入数组公式时，先用鼠标选定所有需放置结果的单元格地址范围然后输入数组公式，例如

“ =MMULT(MINVERSE(OFFSET(H2,0,0,2,2)),
OFFSET(F2,0,0,2)) ”，然后同时按下“ CTRL+SHIFT+回车 ”三个按键，完成数组公式的输入，公式会自动加上一对大括号，它由 Excel 自动添入。2.以上数组公式中包含的各个函数的含义及其用法请参看附表 1。)最后，将每一个方程组的最后一个解，用值复制的方式复制到 pac 这一列，即可得到 8 个偏自相关系数。

如图附-39，表中 H12:O19 单元格的 8 列分别给出了 8 个数组公式计算的结果，F12:F19 单元格的内容即是所求解的 8 个偏自相

关
系
数。

	E	F	G	H	I	J	K	L	M	N	O
1	Lag	ac									
2	1	0.819067942		1	0.819	0.612	0.452	0.327	0.164	0.0388	-0.01
3	2	0.611510228		0.819	1	0.819	0.612	0.452	0.327	0.1642	0.039
4	3	0.451726387		0.612	0.819	1	0.819	0.612	0.452	0.3272	0.164
5	4	0.327219978		0.452	0.612	0.819	1	0.819	0.612	0.4517	0.327
6	5	0.164215603		0.327	0.452	0.612	0.819	1	0.819	0.6115	0.452
7	6	0.038753484		0.164	0.327	0.452	0.612	0.819	1	0.8191	0.612
8	7	-0.009794647		0.039	0.164	0.327	0.452	0.612	0.819	1	0.819
9	8	-0.116257009		-0.01	0.039	0.164	0.327	0.452	0.612	0.8191	1
10											
11		pac		ϕ_{1i}	ϕ_{2i}	ϕ_{3i}	ϕ_{4i}	ϕ_{5i}	ϕ_{6i}	ϕ_{7i}	ϕ_{8i}
12		0.819067942		0.819	0.967	0.971	0.972	0.967	0.972	0.9704	0.997
13		-0.180361799			-0.18	-0.205	-0.21	-0.2	-0.204	-0.184	-0.202
14		0.025905952				0.026	0.048	0.002	0.002	-0.015	-0.085
15		-0.022714997					-0.02	0.19	0.195	0.1951	0.256
16		-0.21914027						-0.22	-0.243	-0.226	-0.23
17		0.024732305							0.025	-0.058	-0.115
18		0.084685469								0.0847	0.386
19		-0.310257383									-0.31

图 附-39

3. 模型的识别与估计。自相关函数序列呈现明显拖尾性，而偏自相关

函数序列在 $k > 1$ 之后，都在区间 $(-\frac{1.96}{\sqrt{20}}, \frac{1.96}{\sqrt{20}})$ ，即 $(-0.438, 0.438)$

之间，因此可以认为自相关函数在 $K > 1$ 之后截尾，因此我们选用 AR(1)模型进行数据拟合。复制 C2:C20 的数据，将之以值复制的形式复制到 D3:D21 的单元格，并在 D1 中填入标志项“Z(-1)”。选择“工具”菜单的“数据分析”子菜单，双击“回归”选项，弹出回归分析对话框。按图附-40 所示的方式填写对话框。然后单击“确定”按钮，即可得到 AR(1)模型的估计结果。

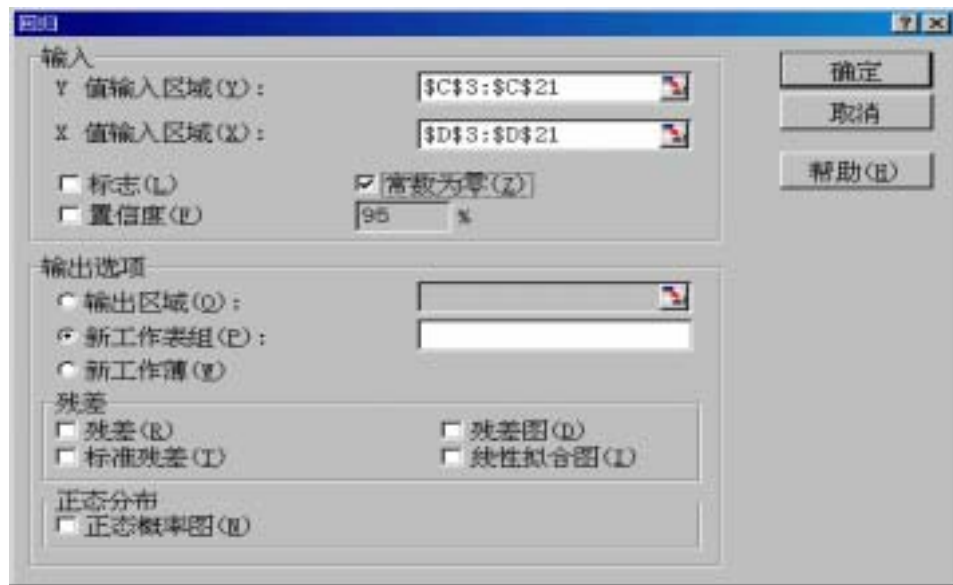


图 附-40

(三) 结果分析：按以上操作步骤，可得到图附-41 所示 AR(1)模型

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	回归统计								
4	Multiple R	0.93329							
5	R Square	0.87104							
6	Adjusted R Square	0.81548							
7	标准误差	13.0237							
8	观测值	19							
9									
10	方差分析								
11		df	SS	MS	F	Significance F			
12	回归分析	1	20621.4	20621.4	121.577	3.63113E-09			
13	残差	18	3053.09	169.616					
14	总计	19	23674.5						
15									
16	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%	
17	Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	
18	X Variable 1	1.06284	0.09639	11.0264	1.9E-09	0.860332697	1.26535	0.86033	1.26535

图 附-41

结果。因此，零均值化模型的估计结果是 $\hat{Z} = 1.06284 * Z(-1)$ ，还原成上证指数，最终的时间序列模型是：上证指数估计值-上证指数的平均值=1.06284(前一天上证指数-上证指数平均值)。

十四、季节变动时间序列的分解分析

(一) 简介：分解分析法是分析时间序列常用的统计方法。季节

时间序列是趋势变动(T)、季节变动(S)、随机变动(I)综合影响的结果,分解过程要从原始序列中消除随机变动,然后分别识别出季节变动和趋势变动的变化模式。下面结合具体例子介绍在 Excel 中如何实现时间序列的分解分析。如图附-42 所示,表中 A1 至 B13 单元格是 1996 至 1998 年各季度某海滨城市旅游人口数(千人),试预测 1999 年各季度旅游人口数。

(二) 操作步骤:

1. 计算一次移动平均,消除随机波动。在 C3 单元格填入公式“=AVERAGE(B2:B5)”,然后用“填充柄”将公式复制到 C4:C11 单元格。
2. 中心化移动平均数。在 D4 单元格输入公式“=AVERAGE(C3:C4)”,然后用“填充柄”将公式复制到 D5:D11 单元格。
3. 计算各个季节指数。在 E4 单元格输入公式“=B4/D4”,然后用“填充柄”将公式复制到 E5:E11 单元格。
4. 计算平均季节指数。在 F4 单元格中输入公式“=AVERAGE(E4,E8)”,然后用“填充柄”将公式复制到 F5:F7 单元格。

	A	B	C	D	E	F	G	H	I
1	季度	旅游人数	一次移动平均	中心化移动平均	季节指数	平均季节指数	调整季节指数	消除季节变动	时间
2	1996.1	77					0.670588435	114.8245272	1
3	1996.2	115	148.75				0.872016265	131.8782741	2
4	1996.3	298	157.75	153.25	1.94454	1.860272838	1.820272369	163.7117637	3
5	1996.4	105	168.75	163.25	0.64319	0.651123701	0.637122931	164.8033605	4
6	1997.1	113	178	173.375	0.65177	0.68532461	0.670588435	168.5087217	5
7	1997.2	159	184.25	181.125	0.87785	0.891178814	0.872016265	182.3360486	6
8	1997.3	335	193	188.625	1.77601		1.820272369	184.0383921	7
9	1997.4	130	201.5	197.25	0.65906		0.637122931	204.0422558	8
10	1998.1	148	210.25	205.875	0.71888		0.670588435	220.7016887	9
11	1998.2	193	216.5	213.375	0.90451		0.872016265	221.326147	10
12	1998.3	370					1.820272369	203.2662838	11
13	1998.4	155					0.637122931	243.2811512	12
14	1999.1								13
15	1999.2								14
16	1999.3								15
17	1999.4								16

图 附-42

5. 计算调整后的季节指数。为了让季节指数的总平均为 1,必须对季节指数加以调整。在 G4 单元格中输入公式“=F4/AVERAGE(\$F\$4:\$F\$7)”,然后用“填充柄”将公式复制到 G5:G7 单元格。G4:G7 就是最终计算出的季节指数,按 G4:G7

给出的 4 个季度的季节指数，将季节指数填充到 G2:G13 的其它单元格。

6. 消除旅游人数序列中的季节变动。在 H2 单元格中输入“=B2/G2”，然后用“填充柄”将公式复制到 H3:H13 单元格。则 H 列就是消除季节变动之后的旅游人数时间序列。
7. 对消除季节变动的旅游人数进行回归分析。在 I 列填入时间序号 1 至 15，如图附-42 所示。选择“工具”菜单的“数据分析”子菜单，双击“回归”选项，弹出回归分析对话框。按图附-43 所示的方式填写对话框。然后单击“确定”按钮，即可得到剔除了季节波动的时间序列的线性趋势模型。估计结果如图附-44 所示，其中 B35 单元格是线性趋势模型的截距，B36 单元格是斜率。

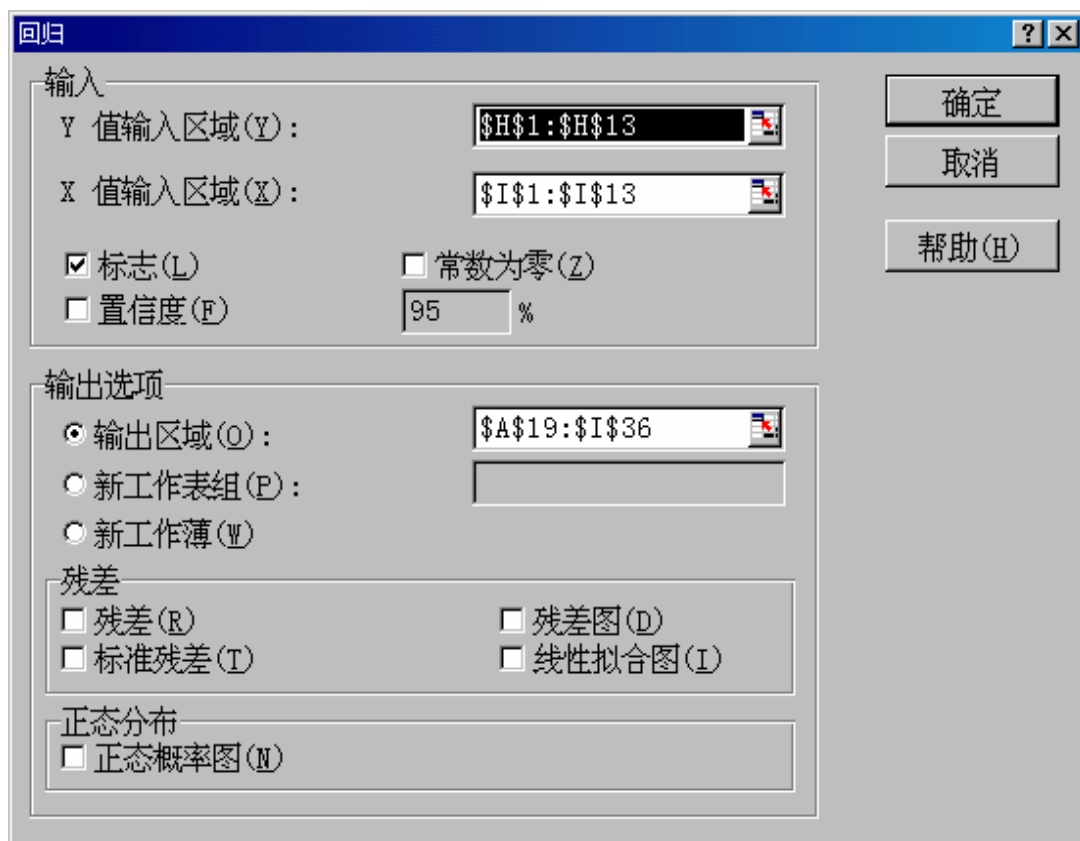


图 附-43

	A	B	C	D	E	F	G	H	I
19	SUMMARY OUTPUT								
20									
21	回归统计								
22	Multiple R	0.95448							
23	R Square	0.91103							
24	Adjusted R Square	0.90213							
25	标准误差	11.762							
26	观测值	12							
27									
28	方差分析								
29		df	SS	MS	F	Significance F			
30	回归分析	1	14166.4156	14166.4156	102.4	1.42572E-06			
31	残差	10	1383.44549	138.344549					
32	总计	11	15549.8611						
33									
34		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
35	Intercept	118.864	7.2390081	16.4199564	1.5E-08	102.7346797	134.99	102.7347	134.994
36	时间	9.95318	0.98358751	10.1192649	1.4E-06	7.761612654	12.145	7.761613	12.1448

图 附-44

8. 预测。在 G14:G17 单元格中分别填入刚才计算出的四个调整后的季节指数，在 B14 单元格中输入公式“=(\$B\$35+I14*\$B\$36)*G14”，然后利用“填充柄”将公式复制到 B15:B17 单元格，B14:B17 单元格中就是 1999 年各个季度旅游人数的预测值，如图附-45 所示。

	A	B	C	D	E	F	G	H	I
1	季度	旅游人数	一次移动平均	中心化移动平均	季节指数	平均季节指数	调整季节指数	消除季节变动	时间
14	1999.1	166.4773					0.67059		13
15	1999.2	225.1622					0.87202		14
16	1999.3	488.1278					1.82027		15
17	1999.4	177.1935					0.63712		16

图 附-45

(三) 结果分析：以上步骤完成了整个季节时间序列的分析和预测过程。使用了分解分析的方法，能将时间数列的各个影响因素都分解出来，由这种方法得到的预测模型和预测结果都比直接使用回归分析要更为可靠合理。参看以上分析步骤，用类似的方法还可以进行月份时间序列、双循环变动时间序列等的分解分析和预测。

附表 1. Excel 统计函数一览表

函数名称	函数功能
AVEDEV	返回一组数据与其均值的绝对偏差的平均值，用于评测这组数据的离散度。
AVERAGE	返回指定序列算术平均值。
AVERAGEA	计算参数清单中数值的算术平均值。不仅数字，而且文本和逻辑值（如 TRUE 和 FALSE）也将计算在内。
BETADIST	返回 Beta 分布累积函数的函数值。Beta 分布累积函数通常用于研究样本集合中某些事物的发生和变化情况。
BETAINV	返回 beta 分布累积函数的逆函数值。即，如果 $\text{probability} = \text{BETADIST}(x, \dots)$ ，则 $\text{BETAINV}(\text{probability}, \dots) = x$ 。beta 分布累积函数可用于项目设计，在给定期望的完成时间和变化参数后，模拟可能的完成时间。
BINOMDIST	返回一元二项式分布的概率值。函数 BINOMDIST 适用于固定次数的独立实验，实验的结果只包含成功或失败二种情况，且成功的概率在实验期间固定不变。例如，函数 BINOMDIST 可以计算三个婴儿中两个是男孩的概率
CHIDIST	返回 X^2 分布的单尾概率。 X^2 分布与 X^2 检验相关。使用 X^2 检验可以比较观察值和期望值。例如，某项遗传学实验假设下一代植物将呈现出某一组颜色。使用此函数比较观测结果和期望值，可以确定初始假设是否有效。
CHIINV	返回 X^2 分布单尾概率的逆函数。如果 $\text{probability} = \text{CHIDIST}(x, ?)$ ，则 $\text{CHIINV}(\text{probability}, ?) = x$ 。使用此函数比较观测结果和期望值，可以确定初始假设是否有效。

CHITEST	返回独立性检验值。函数 CHITEST 返回 X^2 分布的统计值及相应的自由度。可以使用 X^2 检验确定假设值是否被实验所证实。
CONFIDENCE	返回总体平均值的置信区间。置信区间是样本平均值任意一侧的区域。例如，如果通过邮购的方式订购产品，依照给定的置信度，可以确定最早及最晚到货的时间。
CORREL	返回单元格区域 array1 和 array2 之间的相关系数。使用相关系数可以确定两种属性之间的关系。例如，可以检测某地的平均温度和空调使用情况之间的关系。
COUNT	返回参数的个数。利用函数 COUNT 可以计算数组或单元格区域中数字项的个数。
COUNTA	返回参数数组中非空值的数目。利用函数 COUNTA 可以计算数组或单元格区域中数据项的个数。
COVAR	返回协方差，即每对数据点的偏差乘积的平均数，利用协方差可以决定两个数据集之间的关系。例如，可利用它来检验教育程度与收入档次之间的关系。
CRITBINOM	返回使累积二项式分布大于等于临界值的最小值。此函数可以用于质量检验。例如，使用函数 CRITBINOM 来决定最多允许出现多少个有缺陷的部件，才可以保证当整个产品在离开装配线时检验合格。
DEVSQ	返回数据点与各自样本均值偏差的平方和。
EXPONDIST	返回指数分布。使用函数 EXPONDIST 可以建立事件之间的时间间隔模型，例如，在计算银行自动提款机支付一次现金所花费的时间时，可通过函数 EXPONDIST，确定这一过程最长持续一分钟的发生概率。
FDIST	返回 F 概率分布。使用此函数可以确定两个数据系列是否存在变化程度上的不同。例如，分析进入高校的男生、女生的考试分数，确定女生分数的变化程度是否与男生不同。

FINV	返回 F 概率分布的逆函数值。如果 $p = \text{FDIST}(x, \dots)$ ，则 $\text{FINV}(p, \dots) = x$ 。在 F 检验中，可以使用 F 分布比较两个数据集的变化程度。例如，可以分析美国、加拿大的收入分布，判断两个国家是否有相似的收入变化程度。
FISHER	返回点 x 的 Fisher 变换。该变换生成一个近似正态分布而非偏斜的函数。使用此函数可以完成相关系数的假设检验
FISHERINV	返回 Fisher 变换的逆函数值。使用此变换可以分析数据区域或数组之间的相关性。如果 $y = \text{FISHER}(x)$ ，则 $\text{FISHERINV}(y) = x$ 。
FORECAST	根据给定的数据计算或预测未来值。此预测值为基于一系已知的 x 值推导出的 y 值。以数组或数据区域的形式给定 x 值和 y 值后，返回基于 x 的线性回归预测值。使用此函数可以对未来销售额、库存需求或消费趋势进行预测。
FREQUENCY	以一系列垂直数组返回某个区域中数据的频率分布。例如，使用函数 FREQUENCY 可以计算在给定的值集和接收区间内，每个区间内的数据数目。由于函数 FREQUENCY 返回一个数组，必须以数组公式的形式输入。
FTEST	返回 F 检验的结果。F 检验返回的是当数组 1 和数组 2 的方差无明显差异时的单尾概率。可以使用此函数来判断两个样本的方差是否不同。例如，给定公立和私立学校的测试成绩，可以检验各学校间的差别程度。
GAMMADIST	返回 γ 分布。可以使用此函数来研究具有偏态分布的变量。伽玛分布通常用于排队分析。
GAMMAINV	返回 γ 分布的累积函数的逆函数。如果 $P = \text{GAMMADIST}(x, \dots)$ ，则 $\text{GAMMAINV}(p, \dots) = x$ 。使用此函数可以研究出现分布偏斜的变量。
GAMMALN	返回伽玛函数的自然对数， $\Gamma(x)$ 。

GEOMEAN	返回正数数组或数据区域的几何平均值。例如，可以使用函数 GEOMEAN 计算可变复利的平均增长率。
GROWTH	根据给定的数据预测指数增长值。根据已知的 x 值和 y 值，函数 GROWTH 返回一组新的 x 值对应的 y 值。可以使用 GROWTH 工作表函数来拟合满足给定 x 值和 y 值的指数曲线。
HARMEAN	返回数据集合的调和平均值。调和平均值与倒数的算术平均值互为倒数。
HYPGEOMDIST	返回超几何分布。给定样本容量、样本总体容量和样本总体中成功的次数，函数 HYPGEOMDIST 返回样本取得给定成功次数的概率。使用函数 HYPGEOMDIST 可以解决有限总体的问题，其中每个观察值或者为成功或者为失败，且给定样本区间的所有子集有相等的发生概率。
INTERCEPT	利用已知的 x 值与 y 值计算直线与 y 轴的截距。截距为穿过 known_x 担和 known_y 担数据点的线性回归线与 y 轴交点。当已知自变量为零时，利用截距可以决定因变量的值。例如，当所有的数据点都是在室温或更高的温度下取得的，可以用函数 INTERCEPT 预测在 0 时金属的电阻。
KURT	返回数据集的峰值。峰值反映与正态分布相比某一分布的尖锐度或平坦度。正峰值表示相对尖锐的分布。负峰值表示相对平坦的分布。
LARGE	返回数据集里第 k 个最大值。使用此函数可以根据相对标准来选择数值。例如，可以使用函数 LARGE 得到第一名，第二名，或第三名的得分。
LINEST	使用最小二乘法计算对已知数据进行最佳直线拟合，并返回描述此直线的数组。因为此函数返回数值数组，故必须以数组公式的形式输入。
LOGEST	在回归分析中，计算最符合观测数据组的指数回归拟合曲线，并返回描述该曲线的数组。由于这个函数返回一个数组，必须以数组公式输入。

LOGINV	返回 x 的对数正态分布累积函数的逆函数，此处的 $\ln(x)$ 是含有 mean (平均数) 与 standard-dev (标准差) 参数的正态分布。如果 $p = \text{LOGNORMDIST}(x, \dots)$ 那么 $\text{LOGINV}(p, \dots) = x$ 。使用对数正态分布可以分析经过对数变换的数据。
LOGNORMDIST	返回 x 的对数正态分布的累积函数，其中 $\ln(x)$ 是服从参数为 mean 和 standard_dev 的正态分布。使用此函数可以分析经过对数变换的数据。
MAX	返回数据集中的最大数值。
MAXA	返回参数清单中的最大数值。文本值和逻辑值 (如 TRUE 和 FALSE) 也作为数字来计算。函数 MAXA 与函数 MINA 相似。
MEDIAN	返回给定数值集合的中位数。中位数是在一组数据中居于中间的数，换句话说，在这组数据中，有一半的数据比它大，有一半的数据比它小。
MIN	返回给定参数表中的最小值。
MINA	返回参数清单中的最小数值。文本值和逻辑值 (如 TRUE 和 FALSE) 也作为数字来计算。
MODE	返回在某一数组或数据区域中出现频率最多的数值。跟 MEDIAN 一样，MODE 也是一个位置测量函数。
NEGBINOMDIST	返回负二项式分布。当成功概率为常数 probability_s 时，函数 NEGBINOMDIST 返回在到达 number_s 次成功之前，出现 number_f 次失败的概率。此函数与二项式分布相似，只是它的成功次数固定，试验总数为变量。与二项分布类似的是，试验次数被假设为自变量。
NORMDIST	返回给定平均值和标准偏差的正态分布的累积函数。此函数在统计方面应用范围广泛 (包括假设检验)
NORMINV	返回给定平均值和标准偏差的正态分布的累积函数的逆函数。
NORMS	返回标准正态分布的累积函数，该分布的平均值

DIST	为 0，标准偏差为 1。可以使用该函数代替标准正态曲线面积表。
NORMSINV	返回标准正态分布累积函数的逆函数。该分布的平均值为 0，标准偏差为 1。
PEARSON	返回 Pearson（皮尔生）乘积矩相关系数， r ，这是一个范围在 -1.0 到 1.0 之间（包括 -1.0 和 1.0 在内）的无量纲指数，反映了两个数据集合之间的线性相关程度。
PERCENTILE	返回数值区域的 K 百分比数值点。可以使用此函数来建立接受阈值。例如，可以确定得分排名在 90 个百分点以上的检测候选人。
PERCENTRANK	返回特定数值在一个数据集中的百分比排位。此函数可用于查看特定数据在数据集中所处的位置。例如，可以使用函数 PERCENTRANK 计算某个特定的能力测试得分在所有的能力测试得分中的位置。
PERMUT	返回从给定数目的对象集合中选取的若干对象的排列数。排列可以为有内部顺序的对象或为事件的任意集合或子集。排列与组合不同，组合的内部顺序无意义。此函数可用于彩票计算中的概率。
POISSON	返回泊松分布。泊松分布通常用于预测一段时间内事件发生的次数，比如一分钟内通过收费站的轿车的数量。
PROB	返回一概率事件组中落在指定区域内的事件所对应的概率之和。如果没有给出 upper_limit，则返回 x_range 内值等于 lower_limit 的概率。
QUARTILE	返回数据集的四分位数。四分位数通常用于在销售额和测量值数据集中对总体进行分组。例如，可以使用函数 QUARTILE 求得总体中前 25% 的收入值。
RANK	返回一个数值在一组数值中的排位。数值的排位是与数据清单中其它数值的相对大小（如果数据清单已经排过序了，则数值的排位就是它当前的位置）。
RSQ	返回根据 known_y's 和 known_x's 中数据点计算得出的 Pearson 乘积矩相关系数的平方。详细内容，

	则参阅函数 REARSON。R 平方值可以解释为 y 方差与 x 方差的比例。
SKEW	返回分布的偏斜度。偏斜度反映以平均值为中心的分布的不对称程度。正偏斜度表示不对称边的分布更趋向正值。负偏斜度表示不对称边的分布更趋向负值。
SLOPE	返回根据 known_y's 和 known_x's 中的数据点拟合的线性回归直线的斜率。斜率为直线上任意两点的重直距离与水平距离的比值, 也就是回归直线的变化率。
SMALL	返回数据集中第 k 个最小值。使用此函数可以返回数据集中特定位置上的数值。
STANDARDIZE	返回以 mean 为平均值, 以 standard-dev 为标准偏差的分布的正态化数值。
STDEV	估算样本的标准偏差。标准偏差反映相对于平均值 (mean) 的离散程度。
STDEVA	估算基于给定样本的标准偏差。标准偏差反映数值相对于平均值 (mean) 的离散程度。文本值和逻辑值 (如 TRUE 或 FALSE) 也将计算在内。
STDEVP	返回以参数形式给出的整个样本总体的标准偏差。标准偏差反映相对于平均值 (mean) 的离散程度。
STDEVP A	计算样本总体的标准偏差。标准偏差反映数值相对于平均值 (mean) 的离散程度。
STEYX	返回通过线性回归法计算 y 预测值时所产生的标准误差。标准误差用来度量根据单个 x 变量计算出的 y 预测值的误差量。
TDIST	返回学生氏-t 分布。T 分布用于小样本数据集的假设检验。使用此函数可以代替 t 分布的临界值表。
TINV	返回指定自由度的学生氏-t 分布的逆函数。
TREND	返回一条线性回归拟合线的一组纵坐标值 (y 值)。即找到适合给定的数组 known_y's 和 known_x's 的直线 (用最小二乘法), 并返回指定数组 new_x's 值在直线上对应的 y 值。

TRIMMEAN	返回数据集的内部平均值。函数 TRIMMEAN 先从数据集的头部和尾部除去一定百分比的数据点，然后再求平均值。当希望在分析中剔除一部分数据的计算时，可以使用此函数。
TTEST	返回与学生氏-t 检验相关的概率。可以使用函数 TTEST 判断两个样本是否可能来自两个具有相同均值的总体。
VAR	估算样本方差。
VARA	估算基于给定样本的方差。不仅数字，文本值和逻辑值（如 TRUE 和 FALSE）也将计算在内。
VARP	计算样本总体的方差。
VARPA	计算样本总体的方差。不仅数字，文本值和逻辑值（如 TRUE 和 FALSE）也将计算在内。
WEIBUL L	返回韦伯分布。使用此函数可以进行可靠性分析，比如计算设备的平均故障时间。
ZTEST	返回 z 检验的双尾 P 值。Z 检验根据数据集或数组生成 x 的标准得分，并返回正态分布的双尾概率。可以使用此函数返回从某总体中抽取特定观测值的似然估计。

附表 2 . Excel 数据分析工具一览表

“ F - 检验： 双 样 本 方 差 分 析 ” 分析工具	此分析工具可以进行双样本 F - 检验，用来比较两个样本总体的方差。例如，可以对参加游泳比赛的两个队的时间记分进行 F- 检验，查看二者的样本方差是否不同。
“ t - 检验： 成对双样本均值 分析 ” 分析工具	此分析工具及其公式可以进行成对双样本学生氏 t - 检验，用来确定样本均值是否不等。此 t - 检验并不假设两个总体的方差是相等的。当样本中出现自然配对的观察值时，可以使用此成对检验，例如对一个样本组进行了两次检验，抽取实验前的一次和实验后的一次。

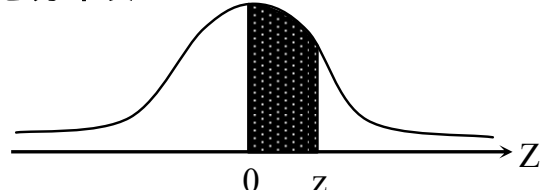
“ t - 检验： 双样本等方差假设 ” 分析工具	此分析工具可以进行双样本学生氏 t - 检验。此 t - 检验先假设两个数据集的平均值相等，故也称作齐次方差 t - 检验。可以使用 t - 检验来确定两个样本均值实际上是否相等。
“ t - 检验： 双样本异方差假设 ” 分析工具	此分析工具及其公式可以进行双样本学生氏 t - 检验。此 t - 检验先假设两个数据集的方差不等，故也称作异方差 t - 检验。可以使用 t - 检验来确定两个样本均值实际上是否相等。当进行分析的样本组不同时，可使用此检验。如果某一样本组在某次处理前后都进行了检验，则应使用“成对检验”。
“ z - 检验： 双样本均值分析 ” 分析工具	此分析工具可以进行方差已知的双样本均值 z - 检验。此工具用于检验两个总体均值之间存在差异的假设。例如，可以使用此检验来确定两种汽车模型性能之间的差异情况。
“ 抽 样 分 析 ” 分析工具	此分析工具以输入区域为总体构造总体的一个样本。当总体太大而不能进行处理或绘制时，可以选用具有代表性的样本。如果确认输入区域中的数据是周期性的，还可以对一个周期中特定时间段中的数值进行采样。例如，如果输入区域包含季度销售量数据，以四为周期进行取样，将在输出区域中生成某个季度的样本。
“ 傅立叶分 析 ” 分析工具	此分析工具可以解决线性系统问题，并能通过快速傅立叶变换（FFT）分析周期性的数据。此工具也支持逆变换，即通过对变换后的数据的逆变换返回初始数据。
“ 回 归 分 析 ” 分析工具	此工具通过对一组观察值使用“最小二乘法”直线拟合，进行线形回归分析。本工具可用来分析单个因变量是如何受一个或几个自变量影响的。例如，观察某个运动员的运动成绩与一系列统计因素的关系，如年龄、身高和体重等。在操作时，可以基于一组已知的体能统计数据，并辅以适当加权，对尚未进行过测试的运动员的表现作出预测。
“ 描 述 统 计 ” 分析工具	此分析工具用于生成对输入区域中数据的单变值分析，提供有关数据趋中性和易变性的信息。

“ 排位和百分比排位 ” 分析工具	此分析工具可以产生一个数据列表,在其中罗列给定数据集中各个数值的大小次序排位和相应的百分比排位。用来分析数据集中各数值间的相互位置关系。
“ 随机数发生器 ” 分析工具	此分析工具可以按照用户选定的分布类型,在工作表的特定区域中生成一系列独立随机数字。可以通过概率分布来表示主体的总体特征。例如,可以使用正态分布来表示人体身高的总体特征,或者使用双值输出的伯努利分布来表示掷币实验结果的总体特征。
“ 相 关 系 数 ” 分析工具	此分析工具及其公式可用于判断两组数据集 (可以使用不同的度量单位) 之间的关系。可以使用 “ 相关系数 ” 分析工具来确定两个区域中数据的变化是否相关,即,一个集合的较大数据是否与另一个集合的较大数据相对应 (正相关); 或者一个集合的较小数据是否与另一个集合的较小数据相对应 (负相关); 还是两个集合中的数据互不相关 (相关性为零)。
“ 协方差 ” 分析工具	此分析工具及其公式用于返回各数据点的一对均值偏差之间的乘积的平均值。协方差是测量两组数据相关性的量度。可以使用协方差工具来确定两个区域中数据的变化是否相关,即,一个集合的较大数据是否与另一个集合的较大数据相对应 (正协方差); 或者一个集合的较小数据是否与另一个集合的较小数据相对应 (负协方差); 还是两个集合中的数据互不相关 (协方差为零)。
“ 移 动 平 均 ” 分析工具	此分析工具及其公式可以基于特定的过去某段时期中变量的均值,对未来值进行预测。移动平均值提供了由所有历史数据的简单的平均值所代表的趋势信息。使用此工具可以预测销售量、库存或其它趋势。
“ 直 方 图 ” 分析工具	在给定工作表中数据单元格区域和接收区间的情况下,计算数据的个别和累积频率,用于统计有限集中某个数值元素的出现次数。例如,在一个

	有 20 名学生的班级里，可以确定以字母打分（如 A、B-等）所得分数的分布情况。直方图表会给出字母得分的边界，以及在最低边界与当前边界之间某一得分出现的次数。出现频率最多的某个得分即为数据组中的众数。
“ 指 数 平 滑 ” 分析工具	此分析工具及其公式基于前期预测值导出相应的新预测值，并修正前期预测值的误差。此工具将使用平滑常数 a ，其大小决定了本次预测对前期预测误差的修正程度。
“ Anova :单 因素方差分析 ” 分析工具	此分析工具通过简单的方差分析(anova)，对两个以上样本均值进行相等性假设检验（抽样取自具有相同均值的样本空间）。此方法是对双均值检验（如 t-检验）的扩充。
“ Anova :可 重复 双 因素 分 析 ” 分析工具	此分析工具是对单因素 anova 分析的扩展，即每一组数据包含不止一个样本。
“ Anova :无 重复 双 因素 分 析 ” 分析工具	此分析工具通过双因素 anova 分析（但每组数据只包含一个样本），对两个以上样本均值进行相等性假设检验（抽样取自具有相同均值的样本空间）。此方法是对双均值检验（如 t-检验）的扩充。

附录二 常用统计表

表 1 正态分布表

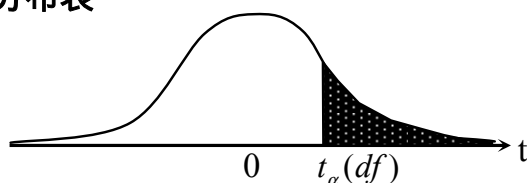


如 $P(0 < Z < 0.24) = 0.0948$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981

2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

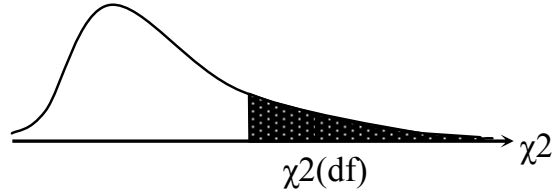
表 2 t 分布表

如 $df=5, P(t > 2.015048) = 0.05$ 

Df \ p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.619
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
∞	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

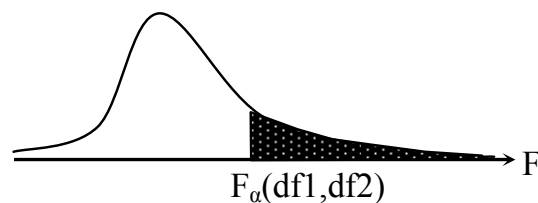
表 3 χ^2 分布表

如当 $df=5$, $P(\chi^2 > \boxed{12.83250}) = 0.025$



df\p	.950	.900	.500	.100	.050	.025	.010	.005
1	0.00393	0.01579	0.45494	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.10259	0.21072	1.38629	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.35185	0.58437	2.36597	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.711072	1.06362	3.35669	7.77944	9.48773	11.14329	13.27670	14.86026
5	1.14548	1.61031	4.35146	9.23636	11.07050	12.83250	15.08627	16.74960
6	1.63538	2.20413	5.34812	10.64464	12.59159	14.44938	16.81189	18.54758
7	2.16735	2.83311	6.34581	12.01704	14.06714	16.01276	18.47531	20.27774
8	2.73264	3.48954	7.34412	13.36157	15.50731	17.53455	20.09024	21.95495
9	3.32511	4.16816	8.34283	14.68366	16.91898	19.02277	21.66599	23.58935
10	3.94030	4.86518	9.34182	15.98718	18.30704	20.48318	23.20925	25.18818
11	4.57481	5.57778	10.34100	17.27501	19.67514	21.92005	24.72497	26.75685
12	5.22603	6.30380	11.34032	18.54935	21.02607	23.33666	26.21697	28.29952
13	5.89186	7.04150	12.33976	19.81193	22.36203	24.73560	27.68825	29.81947
14	6.57063	7.78953	13.33927	21.06414	23.68479	26.11895	29.14124	31.31935
15	7.26094	8.54676	14.33886	22.30713	24.99579	27.48839	30.57791	32.80132
16	7.96165	9.31224	15.33850	23.54183	26.29623	28.84535	31.99993	34.26719
17	8.67176	10.08519	16.33818	24.76904	27.58711	30.19101	33.40866	35.71847
18	9.39046	10.86494	17.33790	25.98942	28.86930	31.52638	34.80531	37.15645
19	10.11701	11.65091	18.33765	27.20357	30.14353	32.85233	36.19087	38.58226
20	10.85081	12.44261	19.33743	28.41198	31.41043	34.16961	37.56623	39.99685
21	11.59131	13.23960	20.33723	29.61509	32.67057	35.47888	38.93217	41.40106
22	12.33801	14.04149	21.33704	30.81328	33.92444	36.78071	40.28936	42.79565
23	13.09051	14.84796	22.33688	32.00690	35.17246	38.07563	41.63840	44.18128
24	13.84843	15.65868	23.33673	33.19624	36.41503	39.36408	42.97982	45.55851
25	14.61141	16.47341	24.33659	34.38159	37.65248	40.64647	44.31410	46.92789
26	15.37916	17.29188	25.33646	35.56317	38.88514	41.92317	45.64168	48.28988
27	16.15140	18.11390	26.33634	36.74122	40.11327	43.19451	46.96294	49.64492
28	16.92788	18.93924	27.33623	37.91592	41.33714	44.46079	48.27824	50.99338
29	17.70837	19.76774	28.33613	39.08747	42.55697	45.72229	49.58788	52.33562
30	18.49266	20.59923	29.33603	40.25602	43.77297	46.97924	50.89218	53.67196

表 4 F 分布表

(A) $\alpha=0.05$ 如: $F_{0.05}(10,6)=\underline{4.0600}$

df2\df1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	∞
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	250.0951	251.1432	252.1957	254.3144
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4624	19.4707	19.4791	19.4957
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6166	8.5944	8.5720	8.5264
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7459	5.7170	5.6877	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.4957	4.4638	4.4314	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	<u>4.0600</u>	3.9999	3.9381	3.8742	3.8082	3.7743	3.7398	3.6689
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.3758	3.3404	3.3043	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.0794	3.0428	3.0053	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.8637	2.8259	2.7872	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.6996	2.6609	2.6211	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.5705	2.5309	2.4901	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.4663	2.4259	2.3842	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.3803	2.3392	2.2966	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3082	2.2664	2.2229	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2468	2.2043	2.1601	2.0658

(续前表)

df2\df1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	∞
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.1938	2.1507	2.1058	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1477	2.1040	2.0584	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1071	2.0629	2.0166	1.9168
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.0712	2.0264	1.9795	1.8780
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0391	1.9938	1.9464	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757	2.0960	2.0102	1.9645	1.9165	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	1.9842	1.9380	1.8894	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	1.9605	1.9139	1.8648	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9390	1.8920	1.8424	1.7330
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9192	1.8718	1.8217	1.7110
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9010	1.8533	1.8027	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.8842	1.8361	1.7851	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.8687	1.8203	1.7689	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045	2.0275	1.9446	1.8543	1.8055	1.7537	1.6376
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8409	1.7918	1.7396	1.6223
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7444	1.6928	1.6373	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.6491	1.5943	1.5343	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.5543	1.4952	1.4290	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.4591	1.3940	1.3180	1.0000

(B) $\alpha=0.01$ 如: $F_{0.01}(10,6)=\boxed{7.874}$

df2\df1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	∞
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6106.321	6157.285	6208.730	6260.649	6286.782	6313.030	6365.864
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.466	99.474	99.482	99.499
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.505	26.411	26.316	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.838	13.745	13.652	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.722	9.553	9.379	9.291	9.202	9.020
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.396	7.229	7.143	7.057	6.880
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.155	5.992	5.908	5.824	5.650
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.198	5.116	5.032	4.859
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.649	4.567	4.483	4.311
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.405	4.247	4.165	4.082	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	3.941	3.860	3.776	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.858	3.701	3.619	3.535	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.665	3.507	3.425	3.341	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.505	3.348	3.266	3.181	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.214	3.132	3.047	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.259	3.101	3.018	2.933	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.003	2.920	2.835	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.919	2.835	2.749	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.844	2.761	2.674	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.778	2.695	2.608	2.421

(续前表)

	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	∞
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.880	2.720	2.636	2.548	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.667	2.583	2.495	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.781	2.620	2.535	2.447	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.577	2.492	2.403	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.699	2.538	2.453	2.364	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.664	2.503	2.417	2.327	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.632	2.470	2.384	2.294	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.753	2.602	2.440	2.354	2.263	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.726	2.574	2.412	2.325	2.234	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.700	2.549	2.386	2.299	2.208	2.006
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.203	2.114	2.019	1.805
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.028	1.936	1.836	1.601
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.192	2.035	1.860	1.763	1.656	1.381
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.185	2.039	1.878	1.696	1.592	1.473	1.000

表 5 DW 检验上下界表 (5%的上下界)

n	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.60
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

续前表

(5%的上下界)

n	k = 2		k = 3		k = 4		k = 5		k = 6	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.26	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

注：n 为样本容量；k 为包括常数项的解释变量个数。

参考文献

- [1]. 郭志刚主编：《社会统计分析方法——SPSS 软件应用》，中国人民大学出版社，1999 年；
- [2]. 柯惠新、黄京华、沈浩：《调查研究中的统计分析法》，北京广播学院出版社；
- [3]. 何晓群：《现代统计分析方法与应用》，中国人民大学出版社；
- [4]. 王庆石、卢兴普主编：《统计学案例教材》，东北财经大学出版社，1999 年；
- [5]. 米红：《应用现代统计分析讲义》，厦门大学研究生院公共学位课试用教材；
- [6]. 钱伯海著：《经济学新论》，中国经济出版社，1999 年；
- [7]. 黄良文主编：《社会经济统计学原理》，中国统计出版社，1994 年；
- [8]. 王美今著：《经济预测与决策》，厦门大学出版社，1997 年；
- [9]. 马俊林、颜金锐、王钦：《现代统计学》，黑龙江朝鲜民族出版社；
- [10]. 茆诗松、王静龙：《概率论与数理统计》，华东师范大学出版社，1994 年；
- [11]. 袁荫棠：《概率论与数理统计》，中国人民大学出版社；
- [12]. 邱东：《多指标综合评价方法的系统分析》，中国统计出版社，1991 年；
- [13]. 王学民：《应用多元分析》，上海财经大学出版社，1999 年；
- [14]. 张寿、于清文：《计量经济学》，上海交通大学出版社；
- [15]. 李长风：《经济计量学》，上海财经大学出版社；
- [16]. 卢纹岱等：《SPSS for Windows 从入门到精通》，电子工业出版社；
- [17]. 袁卫等编著：《新编统计学教程》，经济科学出版社，1998 年；
- [18]. [美]V·F·夏普著，王崇德译：《社会科学统计学》，科学技术文献出

- 版社，1990 年；
- [19].施锡铨、范正绮编著：《数据分析方法》，上海财经大学出版社，1997 年；
- [20].张尧庭编著：《定性资料的统计分析》，广西师范大学出版社，1991 年；
- [21].刘振亚编著：《计量经济学教程》，中国人民大学出版社，1995 年；
- [22].童恒庆编著：《经济回归模型及计算》，湖北科学技术出版社，1997 年；
- [23].安希忠、林秀梅编著：《实用多元统计方法》，吉林科学技术出版社，1990 年；
- [24].张尧庭、方开泰：《多元统计分析引论》，科学出版社，1982 年；
- [25].方开泰著：《实用多元统计分析》，华东师范大学出版社，1989 年；
- [26].周光亚等：《多元统计方法》，吉林大学出版社，1988 年；
- [27].杨维权等：《多元统计分析》，高等教育出版社，1989 年；
- [28].陈希孺著：《数理统计引论》，科学出版社，1981 年；
- [29].王式安编：《数理统计》，北京理工大学出版社，1994 年；
- [30].许飞琼、曾玉平编著：《统计学》，中国统计出版社，1995 年；
- [31].贾俊平等编著：《市场调查与分析》，经济科学出版社，1999 年；
- [32].黄京华著：《用数据解读市场》，中国广播电视出版社，1998 年；
- [33].Kleinbaum,Kupper,Muller,Applied Regression Analysis and Other Multivariable Methods,PWS-KENT Publishing Company,1988;
- [34].Cox,D.R.(1972)"Regression models and life table."Journal of the Royal statistical Society, Series B 34:187-202.
- [35].Efron,B(1977)"The efficiency of Cox's likelihood function for censored data."Journal of the American Statistical Association 72:557-565.
- [36].Tuma,N.B.(1982)"Nonparametric and partially parametric approaches to event history analysis,"PP.1-60 in S.Leinhardt(ed.)Sociology Methodology 1982.San Francisco:Jossey-Bass.

- [37].(1976)"Rewards,resources and the rate of mobility:a nonstationary multivariate stochastic model."American Sociological Review 41:338-360.
- [38].Paul D.Allison(1992)"Event History AnalysisRegression for Longitudinal Event Data ",a SAGE UNIVERSITY PAPER,Series:Quantitative Applications in the Social Sciences.
- [39].Wall,F.J.,Statistical data analysis handbook, Donnelley & Sons Company, 1986;
- [40].Dillon, W.R. & M.Goldstein(1984) Multivariate Analysis: Methods and Applications. New York: Wiley .

(在本书编写过程中参阅了大量相关书籍及报刊杂志的文章和资料 , 限于篇幅 , 这里只将所参阅的主要文献列入。在此 , 对所有参考文献的作者表示真诚的感谢 !)