第一章　绪论

若想了解上帝在想什么，我们就必须学统计，因为统计学就是在测量他的旨意。

——弗洛伦斯·南丁格尔

第一节　统计与数据

一、什么是统计学

关于统计学的比较权威的定义来自《不列颠百科全书》，用一句话可概括为：统计学是关于收集和分析数据的科学和艺术。

统计学具有如下显著性质：

1. 统计学是研究数据的科学。统计学所研究的数据不是抽象的数和形，而是客观世界与现实生活中实际发生的数据，具体是指那些会经常发生变化的数据，它们往往是和偶然现象联系在一起的，或称随机会而发生变化的数据。如经济波动、股市行情、降水概率、彩票中奖号码等。

2. 统计学是以归纳推理为研究方法的科学。统计学是从样本证据出发，利用归纳推理的思想得出更一般的结论，从随机现象中寻求事物的本质，从个别得出一般的结论。

3. 统计学的研究结论带有不确定性。经典决定论给世界描绘出一张精美的具有因果关系的蓝图，19世纪，人们能够看到，世界可能是有规则的，但不服从自然的普通定律，随着统计学的出现和普及，哲人和世人都发现传统的决定论受到了侵蚀，一些新的定律往往是根据概率来表述的，所以统计学的研究成果往往带有不确定性，往往通过统计平均来描述客观世界。

4. 统计学的研究往往伴随新思想的产生。传统的演绎推理结论蕴含在前提之中，而以归纳推理为主的统计学结论具有发散性，但我们可以给出每一种结论对应的概率。当一个新的现象发生时，传统学说只能根据已有的理论外壳解释这种新的现象不是什么，而统计学可以说明它可能是什么，所以统计学是科学研究的原动力。

5. 统计学与大部分学科都有着密切联系，具有广阔的应用领域。随着统计学的发展和科学研究的需求，现在已经很难找到一个不应用统计学的学科领域。归纳目前世界各国统计学的应用，其领域可大略地概括为：精算，农业，动物学，人类学，考古学，审计学，晶体学，人口学，牙医学，生态学，经济计量学，教育学，选举预测和策划，工程，流行病学，金融，水产渔业研究，遗传学，地理学，地质学，历史研究，人类遗传学，水文学，工业，法律，语言学，文学，劳动力计划，管理科学，市场营销学，医学诊断，气象学，军事科学，核材料安全管理，眼科学，制药学，物理学，政治　学，心理学，心理物理学，质量控制，宗教研究，社会学，调查抽样，分类学，气象改善，博彩，……。

二、什么是数据

统计学的研究对象是数据。世界上各种现象无不可用数据来描述。以前我们往往把定性的东西作为理论分析和描述的对象，殊不知，它们也可用数据来表述，如气温、民族、性别、政治观点、满意度、爱情等。随着科学的进步，这些定性描述的现象一方面可以用定性数据来刻画和分析，另一方面迟早会表现为定量数据，如气温最早表现为人的感受：很热、

热、……、很冷，现在我们可用温度计来准确的刻画人们的主观感受。我们这个世界是可以用数据描述的世界，统计学是描述这个世界的科学。

根据客观世界的表现和人们的认知程度，数据又可分为以下不同类型：

1. 观测数据和试验数据

观测数据是指没有对观测对象进行任何人为因素控制条件下得到的客观数据，如GDP、交通流量、收入、年龄、性别、学历、天体测量、降水概率等。

试验数据是在对影响试验对象的相关因素进行人为控制条件下得到的数据，如减肥前后的体重、临床医学试验观测的人体体征、不同地质状况的农作物产量、产品销售试验等。

2. 定性数据和定量数据

定性数据是指用文字、符号、语言等描述的信息，又可具体分为定类数据和定序数据。定类数据是指对事物按照某种特征进行分类，如性别分为男和女；籍贯分为北京、上海、天津等。定序数据是指在对事物进行顺序上的定性描述，如满意程度描述为很不满意、不满意、……、很满意等；学历描述为高学历、中学历、低学历等。

定量数据是指对事物按照某种特征进行具体的数量描述，又可具体分为定距数据和定比数据，如GDP、固定资产投资、收入、年龄等。其中定距数据的特点是，两个数值之间可以计算差值，但它们的比值没有实际意义，常见的例子有摄氏温度、海拔高度、年份等。在定距数据中零点的选定只是为了方便或惯例，而不是通常所指的自然数或固定的零点。也就是说，零只是坐标上的一个点，并不表示该现象没有或不存在。例如，当某天的气温为0℃时并不是说这天就没有温度。定比数据则有绝对零点，并且两个数字间的比值具有实际意义，常见的例子有身高、体重等等。在实际应用中定距和定比数据常常不做区分而作为一个类别（定量数据）来使用。

3. 截面数据、时序数据和面板数据

截面数据是对多个事物在同一时期或时点上进行测量得到的数据，反映事物在相同时间或时点上的表现，如2008年年末我国各地区的GDP、某年全国各高等院校在校学生数等。

时序数据是对某个事物在不同时期或时点上进行测量得到的数据，反映该事物随时间变化的情况，如改革开放以来我国历年的GDP、恢复高考以来全国高校历年的大学生在校学生数等。

面板数据是对多个事物在不同时期或时点上进行测量得到的数据，反映多个事物随时间变化的情况，如1990-2008年全国30个省份的GDP、1990-2008年全国各高等院校在校学生数等。对于面板数据，如果只考虑某一时期或时点的情况，就是截面数据；如果只考虑某一事物的情况，就是时序数据。

第二节　统计学中的基本概念

一、确定性与随机性

从中学起，我们就知道自然科学的许多定律，例如物理中的牛顿三定律，物质不灭定律以及化学中的各种定律等等。但是在许多领域，很难用如此确定的公式或论述来描述一些现象。比如，人的寿命是很难预先确定的。一个吸烟、喝酒、不锻炼的人可能比一个很少得病、生活习惯良好的人活得长。

因此，可以说，活得长短是有一定随机性的。这种随机性可能和人的经历、基因、习惯等无数说不清的因素都有关系。但是从总体来说，我国公民的平均年龄却是非常稳定的。而且女性的平均年龄也稳定地比男性高几年。这就是规律性。一个人可能活过这个平均年龄，也可能活不到这个平均年龄，这是随机的。但是总体来说，平均年龄的稳定性，又说明了随

机之中有规律性。这种规律就是统计规律。

对于有随机性的事件，我们常常用概率来描述其发生的机会。例如在天气预报中会提到降水概率。大家都明白，如果降水概率是百分之九十，那就很可能下雨；但如果是百分之十，就不大可能下雨。显然，这种概率不可能超过百分之百，也不可能少于百分之零。换言之，概率是在0和1之间的一个数，说明某事件发生的机会有多大。

二、总体和样本

在统计应用中，统计方法可以分为描述统计方法和推断统计方法两大类。描述统计指的是用表格、图形和数字来概括、显示数据特征的统计方法。推断统计指的是从总体中抽取样本，并利用样本数据来推断总体特征的统计方法。推断统计的任务一般来说就是用样本统计量来推断参数。推断统计的方法非常丰富，一般以概率论与数理统计为理论基础，是现代统计学的主体内容。

总体是我们所研究的全部同类对象（事物）组成的集合。总体中的一个对象称为个体。比如，我们要研究某大学学生的体能状况，那么，该大学所有学生构成研究的总体，每个学生都是一个个体；我们想了解一批产品的质量，这批产品就是研究的总体，每一件产品都是一个个体；我们想了解某地区居民家庭消费状况，该地区的所有居民户构成研究的总体，每一个居民户都是研究的个体。这里所说的同类事物是指所研究的多个事物具有一定的同质性，比如上面说到的大学生、居民户等，我们不能把产品与居民户放到一起作为一个总体来研究。

根据总体中包含个体的数目是否有限，可将总体分为有限总体和无限总体两类。有限总体是指总体范围可以明确界定，个体数目有限的总体。比如，上面提到的三个总体都是有限总体。无限总体是指个体数目无限的总体。比如，在某一项科学试验中，所有进行的试验构成研究的总体，每一次试验可以看作一个个体，由于试验可以一直进行下去，试验次数是无限的，该总体是一个无限总体；我们想了解连续生产的某电子器件的质量，所有电子器件构成研究的总体，每一个电子器件是研究的个体，由于生产可以一直进行，个体数量是无限的，该总体也是一个无限总体。将总体区分为有限总体和无限总体，其意义在于不同类型的总体可以采取不同的数据获取和数据分析方法。现实中的大多数总体都是有限总体。

对于总体而言，我们总是希望了解其某些方面的特征。但是总体中个体的数量往往是巨大的，对每一个个体都进行逐一考察和了解显然不现实，对无限总体来说这也是不可能的。实际中采取的方法大多是从总体中抽取一部分个体进行研究，据此对总体特征进行推断。这就需要引入有关样本的概念。

样本是我们所研究的全部同类对象中部分对象组成的集合，或者说，样本是部分个体组成的集合，是总体的一部分。样本中所包含个体的数目成为样本容量。比如，从某地区的所有居民户构成的总体中随机抽出100户居民，这100个居民户构成一个样本，样本容量为100。可由此100户居民家庭消费状况推断该地区居民家庭的消费状况。

三、参数和统计量

参数与统计量是分别对应于总体和样本的两个概念。

参数是描述总体数量特征的一种概括性数字度量，是研究者想要了解的总体的某种特征。比如，某大学学生的平均身高与身高的标准差，某批产品的合格率，某地区某月所有居民户的平均消费额，某地区某月居民户的消费额与收入的相关系数等。由于总体数据通常是不知道的，所以这些反映总体某方面数量特征的参数就是一个未知的常数。正因为如此，我们才需要从总体中抽取样本，对总体参数进行估计和推断。

统计量是根据样本数据计算出来的描述样本某方面数量特征的一种概括性数字度量，它是样本的函数。比如，从某大学随机抽取的100名学生，得到一个样本，由其身高数据可以

计算出样本平均数和样本标准差；从一批产品中随机抽出50件产品进行质量检测，就可以得到样本合格率；从某地区抽取100个居民户，由其消费额与收入数据就可以计算消费额与收入的样本相关系数。由于样本是从总体中抽取出来的，所以统计量总是知道的，这样我们就可以根据样本统计量去估计总体参数。

为了区别和叙述方便，总体参数通常用希腊字母表示，比如，总体均值 μ，总体标准差 σ，总体比例 π；样本统计量通常用英文字母表示，比如，样本均值 $\bar{x}$，样本标准差 s，样本比例 p 等。

除了上面提到的用于估计总体参数的统计量外，还有一类常用的统计量，这类统计量是为了进行统计检验而构造出来的，而且作为样本的函数形式上往往较为复杂，比如 t 统计量，F 统计量，χ^2 统计量等。

四、变量和指标

1. 变量

变量是统计方法中一个非常重要和常用的概念。前面提到，总体是所研究的全部同类（同质性）对象（事物）组成的集合。变量就是说明这些同类事物某方面特征的名称。由于同类事物某方面特征总会表现出变化和差异，因此也就把这种特征的名称称为变量。变量的具体取值称为变量值。比如，表1-1为某班学生情况统计表，这里的“姓名”、“学号”、“性别”、“思想表现”和“成绩”都是说明该班学生某方面特征的名称，因此，也就是变量。表中的内容，比如“男”、“优秀”、“380”等就是变量值。

表 1-1　为某班学生情况统计表

姓名	*学号*	*性别*	*思想表现*	*成绩*
张三	01	男	良好	380
李四	02	男	优秀	340
王五	03	女	优秀	385
…	…	…	…	…

事实上，变量值是从不同方面对研究对象进行测量的结果，也就是统计数据。前面提到，按照测量的尺度与数据层次的不同，可将统计数据分为定类数据、定序数据和数值数据。相应的，我们也可以根据变量取值（数据）的类型把变量分为定类变量、定序变量和数值变量。说明事物类别的名称，也就是取值为定类数据的变量，称为定类变量；说明事物有序类别的名称，也就是取值为定序数据的变量，称为定序变量；说明事物数字特征的名称，也就是取值为数值数据的变量，称为数值变量。表1-1中的“性别”、 “思想表现”和“成绩”分别是定类变量、定序变量和数值变量。在这里“姓名”和“学号”只是研究对象的标记，但一般不参与统计分析，所以没有必要区分其类型。

现实中遇到的变量，多数是数值变量，大多数的统计方法也是处理和分析数值变量。根据数值变量的取值情况不同可以将数值变量分为离散数值变量和连续数值变量，简称为离散变量和连续变量。

离散变量是只能取有限个数值并且取值为整数的变量。比如，每分钟通过十字路口的车辆数，每隔1小时抽取的100件产品中不合格产品的数量，获得奖励的次数等，这些变量可能的取值是有限个整数，它们都是离散变量。

连续变量是可以在某区间中连续取值的变量。比如，年龄、体重、温度、销售额等都是连续变量。对于年龄这个变量，由于可以连续取值，因此理论上说是连续变量。但是实际中

我们说某人的年龄习惯于用整数，因此也可以当作离散变量处理。

在一些统计方法中也将变量称为元，比如多元统计分析，多元回归分析等，其中的元就是变量。

2. 统计指标

在实际工作中常常会使用统计指标的概念。统计指标一般有两种理解和使用方法：（1）指反映现象数量特征的概念例如年末人口数、商品销售额、劳动生产率等。（2）指反映现象数量特征的概念和具体数值，例如我国2008年的国内生产总值为300670亿元。

统计指标按其表现形式可分为总量指标、相对指标和平均指标。

总量指标也称为绝对数，指以绝对数形式表现现象规模和水平的统计指标。例如，2008全年入境旅游人数13003万人次；全年国内生产总值300670亿元；2008年末全国参加城镇基本养老保险人数为21891万人等等。绝对数可以分为时点数和时期数。时点数是描述某种现象在某一个特定时刻（某一瞬间或某一时点）数量表现的数据，例如，2008年年末全国总人口为132802万人。时期数是描述某种现象在某一个特定时间范围内所实现的成果的数据。例如，2008年我国全年各种运输方式完成货物运输周转量110301亿吨公里。区分数据是时点数还是时期数的方法之一是看其加总后的结果是否有意义。若有意义则该指标必定是时期数。反之，则必定是时点数。

相对指标是采用两个有联系的绝对数进行对比而得到的比值。也称为相对数，如产业结构比例、性别比、人口密度等等。平均指标也称为平均数，反映现象在某一时间或空间上的平均数量水平。例如职工的平均工资，平均考试成绩，等等。

第三节　常用统计软件简介

20世纪40年代以来，随着计算机技术的不断发展，专门用于数据收集和分析的统计软件也逐渐兴起和发展。功能强大的现代统计软件不仅可以替我们完成繁琐复杂的计算，还可以帮助我们全方位地分析与展示数据，充分挖掘出数据中蕴含的规律。

统计软件的发展使非专业人员应用复杂的统计方法成为可能，极大地促进了统计技术在各个领域的广泛应用，也反过来推动了统计科学的发展。可以说，现代统计学和统计软件已经密不可分了。

一、几种常用的综合统计软件

常用的统计分析软件包括SAS、SPSS、S-plus、R、Stata、Minitab等等。这些软件都能完成常用的统计方法，如描述统计、参数估计、假设检验、方差分析、回归分析、多元分析等，但不同的软件在功能、易用性、扩展性等方面又各具特色，下面我们分别加以简要介绍。

1. SAS

SAS过去是“Statistical Analysis System”的简称，由于其功能现已远远超出了统计分析的范围，“SAS”已经变成了一个单纯的商标。SAS系统具有非常强大的数据分析能力，可以同时处理多个数据集，有很多模块，功能非常全面；在数据分析和统计分析领域被誉为国际上的标准软件和最权威的优秀统计软件包，广泛应用于政府行政管理、科研、教育、生产和金融等不同领域。

SAS软件虽然也提供了许多菜单操作方式，但仍以编程为主，学习起来有一定困难，是最难掌握的统计软件之一。

2. SPSS

SPSS最早是“Statistical Package for Social Sciences”的简称，它也是最早的统计软件之一。该软件从17.0版本开始更名为“SPSS Statistics”，2009年IBM公司收购SPSS公司后，目

前的最新版本开始称为IBM SPSS Statistics。SPSS是一个组合式软件包，集数据处理、数据分析、统计图表生成及统计编程等诸多功能于一身，涵盖了统计学的所有常用的统计方法，广泛应用于通讯、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，是世界上应用最广泛的专业统计软件之一。

许多初学者都喜欢使用SPSS，因为它非常容易使用：用鼠标点击下拉菜单中的命令就能完成分析工作。当然，SPSS也提供了编程的操作方式。

3. S-Plus

许多人认为S-Plus是介于SAS和SPSS之间的一个软件，它也可以完成绝大部分统计分析，具有菜单式的操作界面，同时提供了强大的编程语言。你可以很容易地把自己编写的函数集成到S-Plus中去。S-Plus的绘图能力特别出色，灵活性强。

4. R

R是一套很像S-Plus的免费统计软件，其语法与图形功能几乎跟S-Plus一模一样，大多数的S-Plus程序也可在R上面顺利执行。R可以在R project的网页免费取得，不足之处是没有实现菜单式的图形用户界面，对于初学者来说学习起来较为困难。

这个软件的更新速度很快，比较容易找到最新的统计方法。

5. Stata

经济学和社会科学领域的许多学者喜欢使用Stata软件。这一软件也有菜单式的操作界面，同时提供了强大的编程能力，易学易用，扩展性强，更新速度快，很容易将自己编写或者网上下载的程序加入到软件中。

Stata的回归分析和回归诊断部分功能非常强大，几乎能估计统计学和计量经济学中的所有回归模型，而在多元统计分析方面的功能稍弱。Stata可以用菜单或程序做出高质量的图形，但完成后的图形不能再进行编辑。

6. Minitab

Minitab也是一个简单易学的统计软件，其统计功能和图形功能都比较全面，在统计学的教学中应用广泛。这一软件突出特色是提供的质量改进分析工具非常全面易用。

二、关于学习和使用统计软件的几点说明

1. 不管使用何种统计软件进行数据处理和分析工作，所涉及的工作都包括两部分内容：数据的准备以及对数据的统计分析。

数据的准备包括创建分析数据集和数据的预处理。创建分析数据集指的是将搜集的数据录入或者导入统计软件，按照统计软件的要求，建立相应的数据文件。包括确定变量及变量名称和类型、对数据进行编码、定义标签等，为进一步的分析做好准备。数据预处理包括检查数据是否有错误、是否有异常数据以及异常数据的处理、缺失数据的处理等。

统计分析过程包括：根据研究目的和数据类型选择相应的统计方法分析和呈现数据，估计统计模型，得出分析结果；如果需要的话，根据软件的输出结果对模型进行诊断和改进；将统计分析的结论应用到决策中或者给决策者提供决策参考，这是统计分析的最终目的。

2. 即使你只希望计算一个平均数，统计软件也可能有上百行的输出结果。因此，在使用统计软件进行数据分析时，不要不加分析地把软件的全部输出结果复制到分析报告中。对你研究的问题而言软件的大部分输出可能都是不必要的。

3. 每个软件生成的图表都有自己特定的格式。为了保持统计分析报告中的图表符合学术规范，使用几乎所有统计软件生成的图表都需要进行一些编辑工作。事实上，在规范的研究报告中，你根本看不出来相应的结果是使用哪个统计软件完成的。

4. 统计软件使得复杂的统计方法越来越容易应用，但也容易导致统计方法的误用。对相应的统计方法有透彻的理解是避免误用统计方法的根本保证。

本章小节

本章介绍有关统计的一些基本问题，主要包括以下几个要点：

1. 数据是对客观现象从不同方面进行测量的结果，是对客观现象的一种描述。

2. 统计学是关于数据的科学，是一门收集、整理、显示和分析数据的科学与艺术。

3. 透过数据发现规律是统计的本质。

4. 统计在众多领域可以为我们做许多事情！统计非常有用！

5. 统计数据有不同的类型，数据类型与所应用的统计分析方法有关。

6. 总体与样本、参数与统计量、变量、统计模型是统计学中的几个基本概念。

7. 统计软件可以帮助我们全方位地分析与展示数据，充分挖掘出数据中蕴含的规律。

8. 常用的统计软件有SPSS、SAS、S-PLUS、R、Stata、Minitab等。

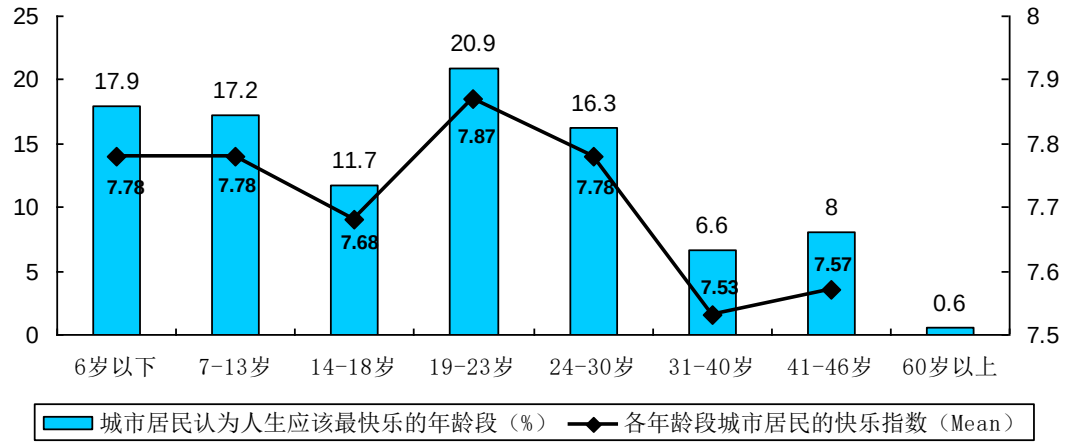
思考与练习

1. 结合实际谈谈你对统计的理解。
2. 举例说明总体、样本、参数、统计量的概念。
3. 指出表1-1中哪些是定类数据？哪些是定序数据？哪些是定量数据？
4. 举例说明定类变量、定序变量与定量变量的含义。
5. 利用统计软件分析数据需要经过哪几个步骤？

第二章　数据的搜集

零点研究咨询集团新近发布的一项关于城市居民“快乐指数”的调查结果显示：中国人的平均快乐指数为7.71。根据图2-1和图2-2，中国人的快乐感来源非常朴素：自己和家人身体健康、满意的经济收入、理想的工作构成快乐的三大主因。而生龙活虎的青年时期（19～23岁）是快乐感最强烈的时期，快乐指数高达7.87分。

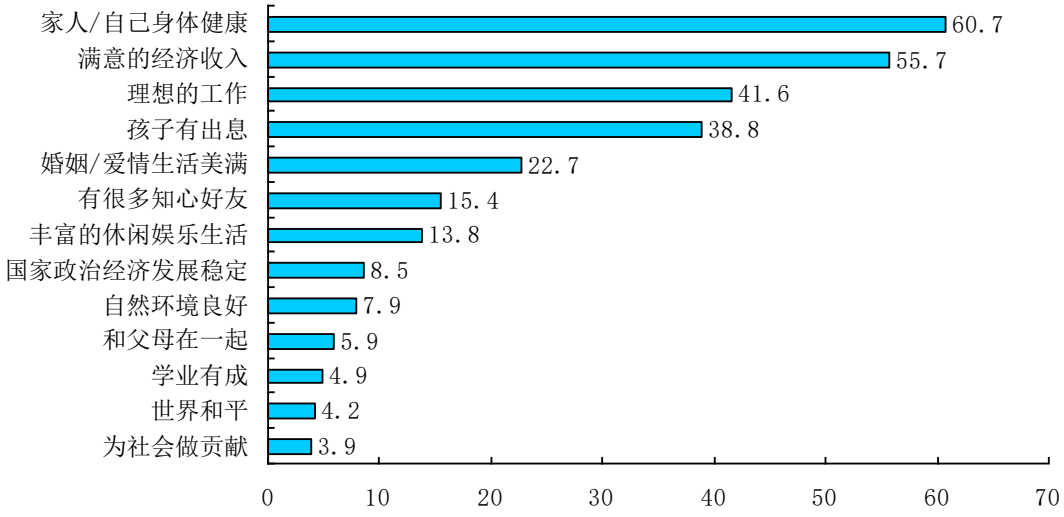
总体看来，中国城市居民还是比较快乐的，在十分制下，各年龄段居民的快乐指数均值在7.5分以上，但年龄段之间呈现多个波峰波谷：13岁以前是“快乐首峰”，14～18岁进入第一个“快乐低谷”，19～23岁达到一生的“快乐巅峰”，随后逐渐下降，而在31～40岁进入“快乐谷底”，步入老年后略有回升。同样，从各年龄段人群心理上最认可的“快乐时期”来看：19～23岁年龄段仍然拔得头筹。



资料来源：零点研究咨询集团2008年5月进行的“快乐指数”小主题调查

图2-1　城市居民认为应该最快乐的年龄段以及各年龄段居民的实际快乐指数

快乐之源在哪里？老话说：身体是革命的本钱，那么，“身体”与“革命”则是快乐的本钱，首先，快乐感最容易从“身体健康”、“事业有成经济优裕”中获得；其次，“家庭和睦”、“广交朋友”等小生活圈要素构成快乐之源的第二圈层；再次，“国际稳定”、“社会和平”等大社会要素带来的快乐感是间接的。



资料来源：零点研究咨询集团 2008 年 5 月进行的“快乐指数”小主题调查

图 2-2　城市居民快乐感的来源（%）

另外，这项研究显示，中国的父母把孩子视为重要的快乐来源，45.1%的壮年和 49.2%的中年人都表示孩子有出息，他们会感到快乐。可见在中国父母的快乐中有很大一部分是来自孩子，在父母的眼中，孩子的成功就是对他们最大的回报。然而“孩子们”并没有同等程度地认为父母是他们的快乐之源。此次调查中，最年轻（18～23 岁）的受访者中，仅有 11%的人认为和父母在一起是他们的一大快乐来源。而已为人父母的中年人群，同样较少视“与父母一起”为快乐来源。他们更多地从休闲娱乐生活和与朋友的交往中获得快乐。在这里，我们还是应该为中国人难能可贵的“高快乐指数”而鼓掌，即使有“青少年裂变期”的心理困惑、即使只是“小家子气”的快乐感、即使快乐来源的代际反差仍很明显，但至少别人问我们“你快乐吗”？我们能够大喊一声“我很快乐”！

此次调查采用多阶段随机抽样方式于 2008 年 5 月针对北京、上海、广州、武汉、沈阳、西安、成都 7 个城市共 1883 名 18～60 岁居民的入户访问，数据结果已根据各地实际人口规模予以加权处理，在 95%置信度下本次调查抽样误差为±2.76%。

这样的抽样调查，我们常在报刊上见到，但其中所表示的意义是什么？例如什么是多阶段随机抽样？95%的置信水平下调查的抽样误差为±2.76%是什么含义？数据是怎样搜集的等等？这些问题就是本章要研究的。

第一节　数据的来源

一、一手数据和二手数据

一位睿智的统计学家说过，世上有两种数据：好数据和坏数据。从统计数据本身的来源看，最初都是来源于直接的调查或实验。然而，从使用者的角度看，数据主要来源于两种渠道：一是直接调查和科学实验；二是别人的调查或实验。

1. 一手数据

一手数据的来源一是来自调查或观察、二是来自实验。调查是取得社会经济数据的重要手段，其中有统计部门进行的调查，也有其他部门或调查机构为特定的目的而进行的调查，

如市场调查、民意调查等；实验是取得自然科学数据的主要手段。本节着重讨论取得社会经济数据的主要方式和方法。

2. 二手数据

对于大多数数据使用者来说，亲自去做调查，搜集一手数据往往是不可能或不必要的。它们往往会利用已经有的，由别人直接调查得到的二手数据。

二手数据主要是公开出版或公开报道的数据，有些是尚未公开出版的数据。在我国，公开出版或报道的商务与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媒介。二手数据通过以下渠道获得：（1）国家统计资料。国家公开的一些规划、计划、统计报告和统计年鉴。（2）行业协会信息资料。行业协会经常公布发表一些行业销售情况，生产经营情况及专题报告。（3）图书资料。从图书馆或其它渠道获得的一些出版物、专业杂志、报纸所提供的信息资料。（4）计算机信息网络。从国际联机数据网络和国内数据库获取有关数据。（5）国际组织，国际商业组织定期发布大量市场信息资料。

使用二手数据之前，有必要先对二手数据进行评价：应注意数据的含义、计算口径和计算方法，以避免误用或滥用；要注意二手数据的时间性，不能用过时的数据；引用二手数据时应充分弄清楚这些数据所载信息之来源和可靠程度，引用时，应注明数据的出处，以尊重他人的劳动成果。

二、常用的统计调查方式

统计调查是取得社会经济数据的主要来源，也是获得直接统计数据的重要手段。实际中常用的调查方式主要有抽样调查、普查、统计报表等。

1．普查

（1）普查的定义

普查是指一个国家或地区为详细地了解某项重要的国情、国力而专门组织的一次性、大规模的全面调查，主要用来收集某些不能够或不适宜用定期的全面调查报表收集的信息资料，如人口普查、农业普查、经济普查等。世界各国一般都定期地进行各项普查，以便掌握有关国情、国力的基本数据。

（2）普查的特点

相对于其他调查方式，普查具有以下显著特点：

①. 普查通常是一次性或周期性的。由于普查涉及面广、调查单位多，需要耗费大量的人力、物力、财力和时间，通常需要间隔较长的时间进行一次。如 1790 年美国举行了第一次人口普查。这次普查范围为当时的 16 个州约 400 万人。由于没有统一印制的普查表，许多市镇、农村的行政区界不清，交通不便，原计划 9 个月完成，实际用了 18 个月。这次普查被认为是世界上第一次近代人口普查。从 1930 年起，美国每十年举行一次人口普查。我国的人口普查从 1953 年到 2000 年共进行了 5 次，分别在 1953 年、1964 年、1982 年、1990 年和 2000 年进行。目前我国的人口普查已规范化、制度化，国务院规定每 10 年进行一次人口普查，即每逢“0”年进行人口普查。

②. 普查一般需要规定统一的标准时点。标准时点是对被调查对象登记时所依据的统一时点。调查数据应反映被调查对象在这一时点的状况，避免调查数据的重复或遗漏。如我国第五次人口普查的标准时点为 2000 年 11 月 1 日 0 时，第二次全国经济普查的标准时点是 2008 年 12 月 31 日。

③. 普查数据的准确性、标准化程度均较高，可以进行统计汇总、分类比较，反映社会总体的一般特征，它可以为抽样调查或其他调查提供基本依据。

④. 普查的调查项目较少，适用范围较狭窄，调查资料缺乏深度。

（3）普查的组织方式

实际中通常采用以下两种普查组织方式：一是组织专门的普查机构，即配备大量普查人员，对调查单位进行直接的登记，如人口普查等；二是利用调查单位的原始记录和核算资料，颁发调查表，由登记单位填报，如物资库存普查等。这种方式比第一种组织方式简便，适应于内容比较单一、涉及范围较小的情况。特别是为了满足某种紧迫需要而进行的“快速普查”，就可以采用这种方式。

2．抽样调查

（1）抽样调查定义

抽样调查是为了根据样本调查结果推断总体数字特征，从总体中随机抽取一部分单位作为样本而进行的调查。

抽样调查是现代推断统计的核心。因为无论是对总体的参数估计还是假设检验，都是以测定样本得到的样本指标数值，即统计量的值为依据。抽样调查依据从总体中抽取的样本单元是否遵循随机原则，可以分为概率抽样和非概率抽样。

世界上最早采用抽样技术研究社会经济问题的是法国数学家拉普拉斯，他曾在 1800 年利用人口出生资料来估计当时法国人口总数。1903 年，第九届国际统计学会建议各国运用抽样技术。1934 年统计学家奈曼发表了有关“代表性抽样的两种方法”的论文是现代抽样调查理论发展史上的里程碑。第二次世界大战后，联合国对抽样调查的发展和推广起了积极作用，1947 年联合国统计司设立“统计委员会统计抽样分会”。随着抽样技术的不断发展和完善，抽样调查在各个领域得到了广泛的应用，例如人口调查、经济调查、社会调查、环境调查、环境资源调查等等。

在我国，随着统计调查制度的不断改革和完善，1994 年经国务院批准同意实施以下统计调查模式：建立以必要的周期性普查为基础，以经常性的抽样调查为主体，同时辅之以重点调查、科学估算等多种方法综合运用的统计调查方法体系。

抽样调查是实际中应用最广泛的一种调查方式。与其他调查方式相比，它具有以下明显的特点：

①. 经济性。这是抽样调查的一个最显著优点。由于调查的样本单位通常是总体单位中的很小一部分，调查的工作量小，因而可以节省大量的人力、物力、财力和时间。

②. 时效性强。抽样调查可以迅速、及时地获得所需要的信息。由于工作量、调查的准备时间、调查时间、数据处理时间等都可以大大缩减，从而提高数据的时效性。与普查等全面调查相比，抽样调查可以频繁地进行，随着事物的发生和发展及时取得有关信息，以弥补普查等全面调查的不足。例如，在两次人口普查之间各年份的人口数据都是通过抽样调查取得的。

③. 适应面广。抽样调查可以获得更广泛的信息，它适用于对各个领域、各种问题的调查。从适用范围和问题来看，抽样调查可用于调查全面调查能够调查的现象，也能调查全面调查所不能调查的现象，特别是适合对一些特殊现象的调查，如产品质量检验、农产品实验、医药的临床实验等。从调查的项目和指标来看，抽样调查的内容和指标可以更详细、更深入，能获得更全面、更广泛和更深入的数据。

④. 准确性高。只要抽样调查的方法运用得当，无论调查对象是什么，我们都可以得到具有代表性的非常精确的估计值。我们知道，包括普查在内的各种调查总会有误差存在，即调查数据与总体客观真值之间的差异。同普查相比，由于抽样调查无论在调查人员的素质、调查时间和调查范围等方面所产生误差的几率都小于普查，因此通过抽样调查能够得到高质量的统计数据。

3．统计报表

统计报表是按照国家有关法规规定，自上而下地统一布置、自下而上地逐级提供基本统计报表的统计报告制度，它可以经常地、定期地搜集反映国民经济和社会发展基本情况，为

各级政府和有关部门制定国民经济和社会发展计划以及检查计划执行情况服务。

统计报表按其性质和要求不同，有如下几种分类：

（1）按报表内容和实施范围不同，分为国家统计报表、部门统计报表和地方统计报表。

（2）按报送周期长短不同，分为日报、旬报、季报、半年报和年报。

（3）按填报单位不同，分为基层统计报表和综合统计报表。

第二节　抽样调查

抽样是通过抽取总体中的部分单元，搜集这些单元的信息，对总体进行推断的手段。如前所述，根据不同的抽选样本方法，有两种类型的抽样：概率抽样和非概率抽样。

一、概率抽样

概率抽样是指严格按随机原则[1]从总体中抽取部分单位构成样本，以此推断总体数量特征。在概率抽样中，若总体中每一个单位被抽中的概率相等，则称为等概率抽样；若总体中至少有一个单位被抽中的概率与其他单位不相等，则称为不等概率抽样。概率抽样的基本抽样组织方式主要有简单随机抽样、分层抽样、系统抽样、整群抽样、多阶段抽样等。

1. 简单随机抽样

简单随机抽样作为讨论概率抽样自然的出发点，并不是由于它被广泛应用（实际上它应用的并不广泛），而是由于它的方法最简单，并且是许多复杂方法的基础。

（1）简单随机抽样的定义

简单随机抽样也称纯随机抽样，是抽样中最基本的随机抽样方法。从含有 N 个单元的总体中直接抽取 n 个单元组成样本，并且每个样本单位被抽中的概率相等（从而每个可能样本被抽中的概率也相同），这种抽样方法就是简单随机抽样。

简单随机抽样分为重复抽样和不重复抽样。在重复抽样中，每次抽中的单位仍放回总体，总体中的单位可能不止一次被抽中。不重复抽样中，抽中的单位不再放回总体，总体中的单位只能抽中一次。社会调查通常采用不重复抽样。

（2）简单随机抽样的抽样方法

使用简单随机抽样，首先需要将总体中每个单位编成从 1~N（N 是总体单位数）的编号，然后根据总体大小选择抽签法或者随机数法抽取所需的样本单位。

①. 抽签法

抽选简单随机样本的一种方式是抽签法。当总体不大时可以利用抽签法进行简单随机抽样，即用均匀同质的材料制作 N 个签并将其充分混合，然后一次抽取 n 个签，或一次抽取一个签但不放回，直至抽够 n 个签为止，这 n 个签上所示的号码便是入样的单元号。

②. 随机数法

我们可以使用随机数表或者计算机软件来获得随机数以完成样本单位的抽选。这里我们只介绍随机数表的方法。随机数表(table of random numbers)，是一张由 0~9 的随机数字组成的表，这些表经过精心编制和验证，以保证每一位数、每一对数以及由此类推的更多位数最终以同样的频率出现在表中。在每次抽取时，每个数字都有相等的机会被抽中。较大的随机数字表是兰德公司编制的一百万数字的表，表 2-1 是这个随机数表中的一部分。例如，假定从 2000 名调查对象中随机抽取 30 名做样本，利用随机数表进行其抽样的步骤如下：先将 2000 名调查对象，由 0001~ 2000 连续编号；然后确定与调查对象编号位数相同选取号码的位数，即在此为 4 位数；确定随机数表中选取号码开始点，开始点可以任意设定。为便于说

[1] 所谓随机原则是指在抽取样本时，总体中每个单位都有一个已知的、非零的被抽取的概率。

明，我们以表 2-1 中第一行的前 4 位数开始；在第一行中从第一个 4 位数开始，按行从左至右选取 30 个 4 位数的号码构成所需样本。从表 2-1 第一行中读取的四位数字依次为：1116，4363（舍去），1875，0613，7674（舍去），2632（舍去），0751……，直到取足 30 个数。

注意由于总体中总共只有 2000 个单位，所以，选号过程中凡遇到大于 2000 的号码（2001~9999）都要跳过不选，而继续后面的选取，不重复抽样时遇到重复的号码也舍去。

表 2-1 兰德公司的随机数表（部分）

11164	36318	75061	37674	26320	75100	10431	20418	19228	91792
21215	91791	76831	58678	87054	31687	93205	43685	19732	08468
10438	44482	66558	37649	08882	90870	12462	41810	01806	02977
36792	26236	33266	66583	60881	97395	20461	36742	02852	50564
73944	04773	12032	51414	82384	38370	00249	80709	72605	67497
49563	12872	14063	93104	78483	72717	68714	18048	25005	04151
64208	48237	41701	73117	33242	42314	83049	21933	92813	04763
51486	72875	38605	29341	80749	80151	33835	52602	79147	08868
99756	26360	64516	17971	48478	09610	04638	17141	09227	10606
71325	55217	13015	72907	00431	45117	33827	92873	02953	85474
65285	97198	12138	53010	94601	15838	16805	61004	43516	17020
17264	57327	38224	29301	31381	38109	34976	65692	98566	29550
95639	99754	31199	92558	68368	04985	51092	37780	40261	14479
61555	76404	86210	11808	12841	45147	97438	60022	12645	62000
78137	98768	04689	87130	79225	08153	84967	64539	79493	74917
62490	99215	84987	28759	19177	14733	24550	28067	68894	38490
24216	63444	21283	07044	92729	37284	13211	37485	10415	36457
16975	95428	33226	55903	31605	43817	22250	03918	46999	98501
59138	39542	71168	57609	91510	77904	74244	50940	31553	62562
29478	59652	50414	31966	87912	87154	12944	49862	96566	48825
96155	95009	27429	72918	08457	78134	48407	26061	58754	05326
29621	66583	62966	12468	20245	14015	04014	35713	03980	03024
12639	75291	71020	17265	41598	64074	64629	63293	53307	48766
14544	37134	54714	02401	63228	26831	19386	15457	17999	18306
83403	88827	09834	11333	68431	31706	26652	04711	34593	22561

资料来源：Rand Corporation, A Million Random Digits with 100,000 Normal Deviates，1955, The Free Press.

简单随机抽样是最基本的抽样方法，其他随机抽样方法都是在它的基础上发展起来的，它的数学性质简单，理论也最为成熟。由于编制抽样框及抽取样本可能过于分散等原因，简单随机抽样在实际中具体实施起来有一定困难。另外，该方法没有利用其他辅助信息以提高估计的效率，所以在大规模的调查中，很少直接采用简单随机抽样，一般是将这种方法与其他抽样方法结合在一起使用。

2. 分层抽样

分层抽样是根据辅助信息将总体划分为若干个子总体，或称为层，然后分别从每个层中随机抽选样本。有两个原因促使我们进行分层抽样，一是组织管理方便，便于操作；二是有辅助信息表明总体的某些个体有明显的差异，将相近的化为一层，将会提高对总体推断的精度。例如，为了研究某个城市职工收入水平，可将该市全体职工分为机关、企业、事业等三个类型，然后再从各种类型的职工中抽取一定数量的职工组成样本。通过划类分层，增大了

各类型中单位间的共同性，使各类型内部的差异必定小于总体的差异，从各组中抽取的样本单位，其代表性较强。同时，各类型都有一定的单位入选，就可能使样本的结构更近似于总体的结构。因此，分层抽样在对总体参数进行估计的同时，还能对每层的参数进行估计，不仅研究了总体，还了解了总体内部某些信息。

分层的好处是由于各层中的样本量是由抽样者所控制，而不是由抽样过程随机决定的。层的样本量常常被确定与各层的规模成比例，当然也可以进行非比例分配。因此，分层抽样又分按比例分层抽样和不按比例分层抽样两种：（1）按比例分层抽样是根据各种类型或层次中的单位数目占总体单位数目的比重来抽取子样本的方法。（2）不按比例分层抽样是指在对各层抽取样本单位时，不是根据各层单位数占总体单位数的比重来确定各自的样本单位数目，而是根据具体研究的需要来确定。在实际抽样中，有的层在总体中的比重太小，其样本量就会非常少，此时采用该方法，主要是便于对不同层次的子总体进行专门研究或进行相互比较。如果要用样本资料推断总体时，则需要先对各层的数据资料进行加权处理，调整样本中各层的比例，使数据恢复到总体中各层实际的比例结构。

3. 整群抽样

整群抽样是首先将总体分为 N 个群或组，每个群或组中包含若干单位。然后按某种方式从总体中随机抽取 n 个群，对被抽中的所有单位进行全面调查。例如对自动流水线上大批量生产的产品进行质量检验，可以按时间间距将产品分群，每隔若干小时抽 15 分钟的产品，这 15 分钟出产的全部产品即为一群。然后对抽中的 15 分钟的产品逐个进行检验。此外，也可以按照现成的集合体为群。例如，库存大批成箱的零件，1 箱即为一群，对农村居民进行某种调查，1 个自然村即可为一群等等。

整群抽样的特点是，样本单位比较集中，容易集中力量进行调查，因此便于组织与管理，也节省了调查时间和费用；整群抽样不需要所有总体单位的抽样框；由于样本单位不能均匀的分布在总体各个部分，所以样本的代表性要差一些。

整群抽样与分层抽样的区别是，分层抽样要求组之间的差异较大，而各组内部差异较小；整群抽样要求各群之间的差异较小，而各群内部的差异性要大。换句话说，分层抽样是用代表不同组的子样本来代表总体中的组分布；整群抽样是用被抽中的群代表总体，再通过被抽中群所有单位的分布来反映总体样本的分布。

4. 系统抽样

系统抽样是将 N 个总体单元按某种顺序排列起来，在规定的范围内随机抽取一个单元作为样本的第一个单元，即起始单元，然后按一套确定的规则抽取其他样本单元，最终构成样本的一种抽样方法。当我们将总体中 N 个单元按一定顺序排列后，按简单随机抽样的方式抽取一个起始单元的编号 r，然后按照一个确定的间距 k 选取逐个抽取样本，直到抽满所需的样本量为止。这种系统抽样称为等距抽样。k 称为抽样间距。系统抽样在实际工作中应用广泛，特别是在大规模抽样调查中（如城乡居民住户抽样调查，人口抽样调查，农业产量的抽样调查，产品质量抽样检查等）都普遍采用系统抽样。

在对总体各单位进行排队时，所排定的顺序可能与所研究的项目无关，也可能有内在的联系。例如，职工按姓氏笔画排队、从中选取部分职工进行免费体检时，职工的顺序与研究目标没有任何关系；如果研究的是职工的收入，职工按受教育程度升序排队，则在总体中职工的收入会呈现出由低到高的整体趋势。在后一种情况下需要特别的抽样技术来降低误差，这里不再具体讲解。系统抽样的优点：（1）抽取样本简便易懂；（2）样本单元在总体中分布比较均匀，有利于提高估计的精度。系统抽样的缺点：（1）若单元的排列存在周期性的变化，样本的代表性可能很差；（2）系统抽样的方差估计较为困难。

系统抽样可以看成是一种特殊的整群抽样，或者看成是一种分层抽样。但在系统抽样的过程中，一旦初始单元确定，整个样本就确定了，这是系统抽样有别于其他抽样方法的一大

特点。但是由于系统抽样的所有样本对应于总体全部单元的一个特定顺序排列，所以并不是完全意义上的简单随机抽样。

5. 多阶段抽样

多阶段抽样是先从总体中随机地抽取若干初级单位，再从初级单位中抽取若干二级单位，……如此下去，直至抽取所要调查全部单位的抽样方法。如在一个省的范围内，要从100 多万农户中抽取 5000 户进行受教育情况、文化水平的调查研究，显然无法一户一户直接抽取。可行的办法是：第一阶段，在全省所有县中随机抽取若干个县；第二阶段，在抽中的县中随机抽取若干个乡；第三阶段，在抽中的乡中随机抽取若干个村；最后，在所有被抽中的村中再随机抽取农户，并使这些农户总数达到预定的样本单位数。我国国家统计局进行的许多全国性的调查都是多阶段抽样。如全国的农产量调查、城乡居民的住户调查、中小工商企业调查等由于样本单位遍布在全国各地，显然不可能直接一次抽到所需要的样本，只能分几个阶段来逐级抽取。

多阶段抽样的阶数可以是任意的，但因为阶数越多，设计就越复杂，估计也就更困难。在多阶段抽样的各阶中可以使用任何一种抽样方法，因此多阶段抽样的主要优点之一是灵活。

抽样调查的组织方式完全取决于调查研究的目的要求、调查对象的特点和客观的条件。凡是能够最经济、最省时而又能够满足预期精确度和可靠性的组织方式，便是一种好的组织方式，这也是抽样设计的最根本的原则。

二、非概率抽样

非概率抽样是用主观的（非随机的）方法从总体中抽选样本。概率抽样的优势在于可以使用偏差很小或无偏的估计量，并且可以对样本估计值的精确度进行估计。而使用非概率抽样，用样本推断总体存在很大风险，因为从总体中抽选样本的方式可能会导致出现较大的偏差。既便如此，各种形式的非概率抽样仍在实践中被广泛采用，最主要的原因是节省费用和快速简便。以下介绍几种常用的非概率抽样方法，即方便抽样、判断抽样、配额抽样和滚雪球抽样。

1. 方便抽样

方便抽样即任意抽样，是根据调查者的方便性，以无目标、随意的方式进行的抽样调查活动。例如，常见的在街头拦截式访问和随意的入户访问就是方便抽样的常见形式。方便抽样是所有抽样技术中花费最小的（包括经费和时间），抽样单元是可以接近的、容易测量的、并且是合作的。在某些调查测试中，方便抽样会取得快速有效的结果。方便抽样的问题是无法知道所抽取的样本代表目标总体的程度，试图利用这类样本的结果对一般总体进行推断是危险的。

2. 判断抽样

判断抽样是调查者根据主观经验和判断从总体中选取有代表性的样本的方法，其抽样精度取决于抽样者的经验。判断抽样与方便抽样一样，也无法知道所抽取的样本代表目标总体的程度。若调查者对问题的判断比较客观、准确，则判断抽样样本的代表性要优于方便抽样。增强调查者判断的客观性和准确性的主要途径是积累判断经验和多向专家咨询。

3. 配额抽样

配额抽样也称定额抽样，是非概率抽样方法中最常用的一种抽样方法。配额抽样是根据调查者认为较重要的一些变量将总体分类，并确定各类（层）单位的样本数额（即指定每一类中的配额），然后从各层或各类中主观地选取一定比例的样本的方法。如一个调查员可能被指派的配额是：6 个 40 岁以下的男性，7 个 40 岁以下的女性，8 个就业的男性，5 个就业的女性等。

配额抽样较之判断抽样加强了对样本结构与总体结构在“量”的方面的质量控制，能够

保证样本有较高的代表性。然而采用这种方法进行选取样本，仍难避免受调查员的选择偏好的影响。如果在严格控制调查员和调查过程的条件下，可使配额抽样获得与分层抽样相接近的结果。不过，配额抽样的被调查者不是按随机原则抽出来的，配额抽样注重的是样本与总体在结构比例上的表面一致性；而分层抽样必须遵守随机原则。分层抽样一方面要提高各层间的差异性与同层内的同质性，另一方面是为了照顾到某些比例小的层次，使得所抽样本的代表性进一步提高，误差进一步减小。

4. 滚雪球抽样

滚雪球抽样是先选择一组调查对象，通常是随机地选取的。访问这些调查对象之后，再请他们提供另外一些属于所研究的目标总体的调查对象，根据所提供的线索，选择此后的调查对象。这一过程会继续下去，形成一种滚雪球的效果。此抽样的主要目的是估计在总体中十分稀有的人物特征。

抽样调查的组织方式完全取决于调查研究的目的要求、调查对象的特点和客观的条件。凡是能够最经济、最省时而又能够满足预期精确度和可靠性的组织方式，便是一种好的组织方式，这也是抽样设计的最根本的原则。

三、抽样误差与非抽样误差

数据质量的好坏直接影响统计分析的结果，数据的真实性是统计工作的生命。调查数据的误差可以来自许多不同方面，归纳之可分为两大类：抽样误差和非抽样误差。

1. 抽样误差(sampling error)

抽样误差是由于抽取样本的随机性造成的样本值与总体值之间的差异，即由样本结果估计总体特征所产生的误差。用估计量的方差或者标准差表示。抽样调查中之所以会出现这样一种误差，是因为它观测的并非总体中的所有单元，而只是其中的样本。由于样本的特征和总体的特征不一定完全相同，当二者不一致时便会产生误差。只要采用抽样调查，该误差就不可避免。然而，抽样误差是能够计量且可以得到控制的（要求必须是概率抽样，这也正是人们愿意采用概率抽样的重要原因）。

抽样误差的大小，取决于以下因素：（1）总体内部的差异程度。在其他条件固定时，总体内部差异越大，抽样误差就越大；反之，抽样误差就小。（2）样本容量的大小。在其他条件固定时，样本越大，抽样误差越小。抽样误差常会随着样本大小（sample size）的增加而缩小，从图 2-3 可以看到，抽样误差虽然在开始时随样本增加而缩小，但在一定阶段后便稳定了下来。换句话说，在超过某一阶段后，努力减少抽样误差已不符合经济效益。所以，过了这一阶段，只要稍微降低一点精度，就可以省下可观的成本。（3）抽样的方式方法。不同的抽样方式方法，产生的抽样误差也有差异。以上三个因素除第一个因素外，其余两个都是人为决定的。因此，抽样误差可以创造条件加以控制，这就大大提高了抽样调查的应用价值，使它能够适应不同条件和不同精确度要求的调查研究。

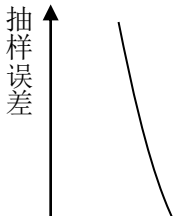


图 2-3 抽样误差与样本大小的关系

2. 非抽样误差(non-sampling error)

除了抽样误差，调查中还会出现由各种原因引起的与样本抽取无关的误差，即非抽样误差。非抽样误差不仅出现在概率抽样、非概率抽样中，也出现在全面调查和非全面调查中。与抽样误差不同的是，非抽样误差不能通过增大样本量得以控制，并且它对调查结果的影响非常大。在某些情形，正是由于非抽样误差没有得到有效的控制，造成调查结果的失真从而导致一项调查的失败。非抽样误差产生于抽样调查的各个环节，即调查方案设计、抽样设计、数据搜集、数据处理及分析等阶段。非抽样误差的特点：（1）在抽样调查中，非抽样误差不可能随着样本量的增大而变小；（2）容易造成估计量的有偏；（3）难以识别和测定；（4）非抽样误差的成因比较复杂。

3. 非抽样误差分类

非抽样误差可分为三类：抽样框误差、无回答误差和计量误差。

（1）抽样框误差

抽样调查中必须对两种总体做好区分。它们是目标总体和抽样总体。目标总体是作为调查研究对象的全体；抽样总体是从中抽选样本的总体。抽样框误差指目标总体和抽样总体不一致时产生的误差。理想的抽样框需要满足的是所有的抽样单位必须覆盖目标总体，对于较为简单的单阶段抽样，抽样框要求每个目标总体单位都应该对应着一个抽样单位，抽样单位必须相互独立，互不重叠，并且唯一地与目标总体相连接。

抽样框误差包括以下几种情形：a. 丢失目标总体单元。在这种情形抽样框(抽样总体)没能覆盖全部总体单元，它使总体总和估计偏低，同时也会造成均值(或比例)估计的偏倚。b. 包含非目标总体单元。在这种情形抽样框包含了一些不属于研究对象的即非目标总体单元。这种情况常造成总体总和估计的偏高。c. 复合连接。抽样框中的单元与目标总体单元不完全是一对一对应而是存在一对多或多对多的现象。这种情况称为抽样框(抽样总体)与目标总体存在着复合连接。d. 不正确的辅助信息。有些复杂抽样框还包含辅助消息(当采用分层抽样、不等概率抽样以及使用比估计或回归估计等情形)，如果这些辅助信息不完全或不正确，不仅不能提高抽样的效率，反而会降低估计的准确性，从而导致误差。

（2）无回答误差

无回答误差是指在调查中由于各种原因没有能够对被抽取的样本单位进行计量，没有获得有关样本单位的信息，而造成的偏误。很少有抽样调查的回答率达到 100%，大多数调查或多或少都会抽到搜集不到的数据。

无回答误差可以分为单位无回答和项目无回答。单位无回答是指被调查者没有参与或拒

绝接受调查。项目无回答是指被调查者虽然接受调查，但对其中的一些调查项目没有回答。无回答的原因多种多样，如地址有误；被调查者不在；无法与被调查者取得联系；被调查者拒访；调查者的失误等等。

降低无回答误差的措施：

①. 问卷设计。使问卷设计得更加有吸引力，引起被调查者的兴趣，并注意长度。

②. 充分利用调查组织者的社会权威性和社会影响力，激发被调查者的参与意识。

③. 确定准确的调查方位，使调查者容易找到被调查者。

④. 采取有助于消除被调查者冷漠、担心、怀疑的措施，如事先通知、调查前的说明、雇佣调查者熟悉的人做调查员等。

⑤. 做好调查员的培训，增强调查员的责任心，提高其访谈技巧。有经验的调查人员可以把调查中的无回答率降到最低。

⑥. 对调查过程进行监控。对不成功的调查及时总结，归纳经验。如拒绝调查是什么原因造成的，调查时间是否合适等。

⑦. 奖励措施。调查总要花费被调查者的时间和精力，适当的奖励是必要的。如入户调查中向被调查者赠送小礼品，对集体单位调查时许诺提供最后的调查报告或汇总结果等等。一些人接受调查并不是为了得到奖励，但是奖励措施会使对方觉得他们提供的信息很重要。

⑧. 多次调查。在一轮调查完成之后，针对无回答产生的原因，采取相应的措施，对无回答单元进行再次调查。无回答产生的原因有不在家、不方便等。在这两种情况下，多次调查都比较有效果。

⑨. 替换被调查单元。对于放弃的无回答者，需要替换单元，以便使接受的调查样本数不低于设计要求。替换的原则应该事先规定，例如入户调查中的“右手原则”，即用放弃户右边的第一户作为替代单元。替代原则的事先规定可以防止调查员自作主张，也便于事后检查。

（3）计量误差

并非所有记录在问卷上的数据都精确地测量了我们感兴趣的问题。这些不精确性可能是由问卷设计、调查员、被调查者和数据搜集方式有关的误差引起的。如问卷中某些问题措辞不当或问题编排顺序不当导致被调查者答案有偏；又如被调查者可能不愿意或不能精确地回答某些问题；或搜集数据的方式不当都可能导致计量误差的产生。总之，计量误差是指由于多种原因所造成的调查中获得的数据与其真值不一致的误差。

计量误差可以分为三类，一是问卷设计阶段产生的误差，二是调查阶段产生的误差，三是其他误差。问卷设计产生的误差，主要来自于不同措辞的不同表达，包括文字表达本身产生歧意，文字表达不够简练等；问卷设计阶段产生的误差还有相当一部分来自于不同问题的顺序和间隔；问卷设计阶段另外一个误差来源就是，问卷设计过长，导致访问者疲劳而产生数据失真的现象。调查阶段产生的误差来自于两个方面，一是访问员有意或无意导致数据失真，二是被调查者有意或无意导致数据失真。其他计量误差包括随机数字表的编制和使用，数据处理过程中（包括编码、录入）发生的误差等等。

减少计量误差的措施包括：

①. 调查设计方面。调查设计的质量与设计人员的能力密切相关。有能力的设计人员能够设计出更好的调查问卷和抽样程序，以减少由于设计不周所可能带来的计量误差。调查问卷设计出来后，应组织有关人员对问卷进行讨论。如果是大规模的调查活动，还应在正式调查之前进行预调查。

②. 现场准备方面。在收集数据之前，需要做许多准备工作，这些工作质量的好坏，对计量误差会产生直接的影响。主要的准备工作包括招聘访问员、对访问员进行培训、编写调查手册等。

③. 调查结果审核方面。审查是对调查质量进行控制的一道工序，也是减少计量误差的有效方法。审核的目的是要保证调查所得到的数据的完整性、一致性和有效性。

第三节　调查方案设计

为了及时准确地获得被研究现象的实际资料，在数据收集之前，应设计调查方案，调查方案设计的好坏直接影响到调查数据的质量。不同调查任务的调查方案在具体内容和形式上会有一定的差别，但基本结构大体是一致的。

一、调查方案的基本结构

1. 确定调查目的

调查目的是调查研究要解决的问题。只有在目的明确之后，才能有的放失地确定调查什么，向谁调查，采用什么方式方法调查。例如，我国 2008 年开展的第二次全国经济普查的主要目的是全面调查了解我国第二产业和第三产业的发展规模及布局；了解我国产业组织、产业结构、产业技术的现状以及各生产要素的构成；摸清我国各类企业和单位能源消耗的基本情况；建立健全覆盖国民经济各行业的基本单位名录库、基础信息数据库和统计电子地理信息系统。通过普查，进一步夯实统计基础，完善国民经济核算制度，为加强和改善宏观调控，科学制定中长期发展规划，提供科学准确的统计信息支持。

2. 确定调查对象和调查单位

调查对象是根据调查目的确定的调查研究的总体或调查范围。调查单位是构成调查对象的每一个单位，它是调查项目和指标的承担者或载体，是搜集数据、分析数据的基本单位。调查对象和单位所解决的是向谁调查，由谁提供所需数据的问题。2008 年第二次全国经济普查的对象是在我国境内从事第二产业和第三产业的全部法人单位、产业活动单位和个体经营户。行业范围包括：采矿业，制造业，电力、燃气及水的生产和供应业，建筑业，交通运输、仓储和邮政业，信息传输、计算机服务和软件业，批发和零售业，住宿和餐饮业，金融业，房地产业，租赁和商务服务业，科学研究、技术服务和地质勘查业，水利、环境和公共设施管理业，居民服务和其他服务业，教育，卫生、社会保障和社会福利业，文化、体育和娱乐业，以及公共管理与社会组织等。

3. 确定调查内容

调查内容是指需要调查的具体项目。通常以表格的形式来表现，称为调查表。它是用于登记调查数据的一种表格，一般由表头、表体和表外附加三部分组成。表头是调查表的名称，用来说明调查表的内容、被调查单位的名称、性质、隶属关系等；表体是调查表的主要部分，包括调查的具体项目；表外附加通常有填表人签名、填表日期、填表说明等内容组成。如第二次全国经济普查的主要内容包括单位基本属性、从业人员、财务状况、生产经营情况、生产能力、能源消耗、科技活动情况等。与第一次全国经济普查相比，为适应当前国家建设资源节约型、环境友好型社会对能源消耗统计的需求，第二次全国经济普查准备扩大主要能源品种和水的消费量的统计范围。

调查内容通常反映在调查表或调查问卷中，这是统计调查方案的核心。

4. 确定调查方式和方法

根据调查对象和调查内容，要确定采用什么组织方式和方法取得调查资料，在进行调研时，调查方式方法直接影响着调查的准确性、及时性和完备性。统计调查方式是指组织收集原始资料的形式，如普查、统计报表、抽样调查等方式。调查方法即调查者向被调查者收集

数据答案的方法，主要包括访问调查、邮寄调查、电话调查、电脑辅助调查等。例如，我国第一次经济普查规定对法人单位、产业活动单位采用普查的方式，对个体经营户采用普查辅助以典型调查等方式；具体收集数据一律采取访问调查。

5. 确定调查时间

调查时间包括两个方面的涵义：一是资料所属的时间；一是调查期限。对普查来说，调查时间是指普查的标准时点。例如，第二次全国经济普查的标准时点是 2008 年 12 月 31 日，时期资料为 2008 年度。调查期限是指调查工作进行的起讫时间(从开始到结束的时间)，例如，我国第二次经济普查规定，2007 年作为第二次全国经济普查的筹备阶段，主要是研究全国经济普查的总体方案；2008 年为普查的准备阶段，主要根据国务院的总体部署，组建国家及地方各级普查机构，开展宣传动员，进行普查试点，部署并落实普查方案，完成单位清查工作，为普查填报登记工作的开始做好准备；2009 年为普查登记、数据审核处理和普查结果发布阶段，2010 年为资料出版和利用普查结果开展的课题研究阶段。

6. 确定调查的组织实施计划

调查的组织实施计划是统计调查工作顺利进行的保证。因此，在实际的统计调查工作开始之前，要制定出详细的调查工作计划。一般调查的组织实施计划的具体内容包括：调查人员的选择、组织和培训；调查表格、问卷、调查员手册的印刷；必要调查工具的准备；调查经费的来源和预算等。

二、数据的调查方法

不论采取何种方式进行调查，在取得统计数据时，都有一些具体的收集方法。数据的调查方法归纳起来可分为询问调查和观察实验两大类，询问调查是调查者与被调查者直接或间接接触以获得数据的一种方法，具体包括访问调查、邮寄调查、电话调查、电脑辅助电话调查、座谈会、个别深入访谈等。

1. 访问调查

访问调查又称派员调查，它是调查者与被调查者通过面对面地交谈，从而得到所需资料的调查方法。访问调查的方式有标准式访问和非标准式访问两种。标准式访问又称结构式访问，它是按调查人员事先设计好的、有固定格式的标准化问卷或表格，有顺序地依次提问，并由被调查者做出回答。其优点是能够对调查过程加以控制，从而获得比较可靠的调查结果。非标准式访问又称非结构式访问，它事先不制作统一的问卷或表格，没有统一的提问顺序，调查人员只是给一题目或提纲，由调查人员和被调查者自由交谈，以获得所需的资料。市场调查和社会调查中常采用访问调查。

在访问调查中，调查者到人地生疏的地方搜集资料，且这些资料往往又是被调查者不愿意提供的，为顺利完成调查访问工作，调查者事前的准备工作非常重要，即所谓“凡事预则立，不预则废”。调查者事前的准备工作包括：注意自身仪容仪表；出发前须检查应携带的文件，如访员证、调查名册、参考资料、致被调查单位函、调查工作手册、调查表及必要工具等；预约并先了解访问对象；熟记调查表内重要问题及其填写方法；运用各种技巧以激发被调查者主动合作并力求获得最真实的答案；调查者也必须注意自身的安全等。

2. 邮寄调查

邮寄调查是通过邮寄或宣传媒体等方式将调查表或调查问卷送至被调查者手中，由被调查者填写，然后将调查表寄回或投放到预定收集点的一种调查方法。邮寄调查是一种标准化调查，其调查人员和被调查者没有直接的语言交流，信息的传递完全依赖于调查表。邮寄调查问卷或表格发放方式有邮寄、宣传媒介传送、专门场所分发三种。统计部门进行的统计报表及市场调查机构进行的问卷调查中经常使用邮寄调查。

3. 电话调查

电话调查是指调查者通过查找电话号码簿，用电话向被调查者进行询问，以达到搜集调

查资料的一种专项调查。20 世纪 80 年代中期，电话调查开始变得普遍且成为许多场合中调查方法的首选。迄今为止电话调查最大的优点在于它给我们提供了控制调查质量的机会。第二个优点是成本效益高，电话调查不需要差旅费，而且所需的调查时间比同样内容的访问调查少 10%~20%。第三个优点是速度快，电话调查的数据搜集和处理可以同时进行。电话调查的缺点是受访谈内容的复杂性和时间长度的限制较大。所以，每次电话调查时间不能过长；不能提过于复杂的问题；对挂断电话拒绝回答者很难做工作。

电话调查的基本步骤：（1）确定抽样方案；（2）确定抽样中使用的一组电话号码的方法；（3）为每个电话号码制定一张在抽样中使用的呼叫清单；（4）设计和规范问卷草稿；（5）草拟调查员使用的引导/选择单和回撤陈述；（6）聘用调查员和督导员，并制定访谈的具体进度和日程；（7）进行预调查并调整调查步骤和调查设备；（8）打印定稿的问卷和其他表格、清单；（9）培训调查员和督导员；（10）进行全监控的访谈；（11）对完成的问卷进行编辑、编码，并将数据转化成计算机可读的格式；（12）分析数据并撰写调查报告。

4. 电脑辅助电话调查

目前，电脑辅助电话调查已得到广泛应用，并已开发出了各种计算机辅助电话访问调查技术（Computer Assisted Telephone Interview），简称 **CATI**。

电脑辅助电话调查是一种调查员坐在计算机前进行的电话调查。在调查时，调查的问卷、答案都由计算机显示，整个调查的过程包括电话拨号、调查记录、数据处理等也都借助于计算机来完成。在这种调查中，问卷的使用和管理以及整个抽样过程都在计算机工作站的控制中。CATI 可用于控制抽样库的分布，甚至可为已经作好准备的调查员拨叫适当的电话号码；CATI 可以提供多种有关调查员效率的统计数据，为督导人员在人员决策方面提供帮助；CATI 也可以只用于控制问卷的管理而不控制抽样库。

上述电话调查的基本步骤，大多都适用电脑辅助电话调查的环境。

5. 座谈会

座谈会也称集体访谈法，它是将一组被调查者集中在调查现场，通过其对调查的主题发表意见，从而获取调查资料的方法。参加座谈会的人数不宜太多。通常有 6 至 10 人，并且是有关调查问题的专家或有经验之人。通过座谈会，调查者可以从一组被调查者那里获得所需的资料，而且，在彼此间交流的环境中，各个被调查者之间相互影响、相互启发、互相补充，并在座谈过程中不断修正自己的观点，从而有利于取得较为广泛、深入的想法和意见。

6. 个别深入访谈

个别深入访谈是一种一次只有一名被调查者参加的特殊的定性调查方法。调查人员运用大量的追问技巧，尽可能让被调查者自由发挥，不断深入被调查者的思想之中，努力发掘其行为的真实动机。所以个别深入访问常用于被调查者的动机研究，以发挥被调查者非表面化的深层意见。该方法最宜于研究较隐秘的问题如个人隐私或较敏感的问题。

此外，还有观察法和实验法。观察法是在一个真实或模拟的环境里，在被访者完全没有意识的条件下，对他的行为进行观察和分析。它包括直接观察（观察员直接观察被观察者的行为、活动，并记录下来）和简接观察（观察员不直接观察被观察者的行为、活动，而是通过能反映被观察者行动活动的东西达到观察的目的）。实验法是将自然科学中的实验应用到调查研究中的方法。即要求在一个受控制的环境下，使其他因素保持不变，研究所控制的变量对某一变量或某些变量的影响。实验法适用于测试各种广告的效果、研究商品相关因素之间的影响、研究品牌效应、研究各种促销方法的作用等。

三、调查问卷设计

问卷是用来搜集数据的一种工具，是调查者根据调查目的和要求所设计的，由一系列问题、备选答案、说明以及代码表组成的书面文件。问卷设计是根据调查目的和要求，将所需调查的问题具体化，使调查者能顺利地获取必要的信息资料，以便于统计分析的一种手段。

但是，设计一份完善的问卷并非一件轻而易举的事情，问卷设计人员除了要具备统计学、社会学、经济学、心理学、计算机软件等多方面的知识外，还需要掌握一定的技术，可以说，问卷设计是科学与艺术的结合。

问卷设计中根据调查方法的不同，可以把问卷分为访问调查问卷、座谈会调查问卷、邮寄调查问卷和电话调查问卷等；根据问卷的填写方式，可以把问卷分为自填式问卷和代填式问卷，自填式问卷主要适合于邮寄调查、宣传媒介发放的问卷调查等方式，代填式问卷主要适合于访问调查、座谈会调查以及电话调查等方式；根据回答问题的形式可分为“开放式问卷”和“封闭式问卷”。

问卷设计的好坏关系到数据的质量，因为在短短的数个问题中想了解被调查者的真正想法是不容易的，若问题的用语不当、顺序不对等都会影响资料搜集的可靠性，所以研究者在设计问卷时要非常小心，避免因问卷设计不当而破坏了整个调查。本节中，我们从问卷的基本结构，问卷中问题的设计，问卷中答案的设计，问卷中问题的顺序等几个方面来说明问卷设计的一些技巧。

1. 问卷的基本结构

问卷设计中，不同的调查问卷在具体结构、题型、措词、版式等设计上会有所不同，但在结构上一般都由开头部分、甄别部分、主体部分和背景部分组成。

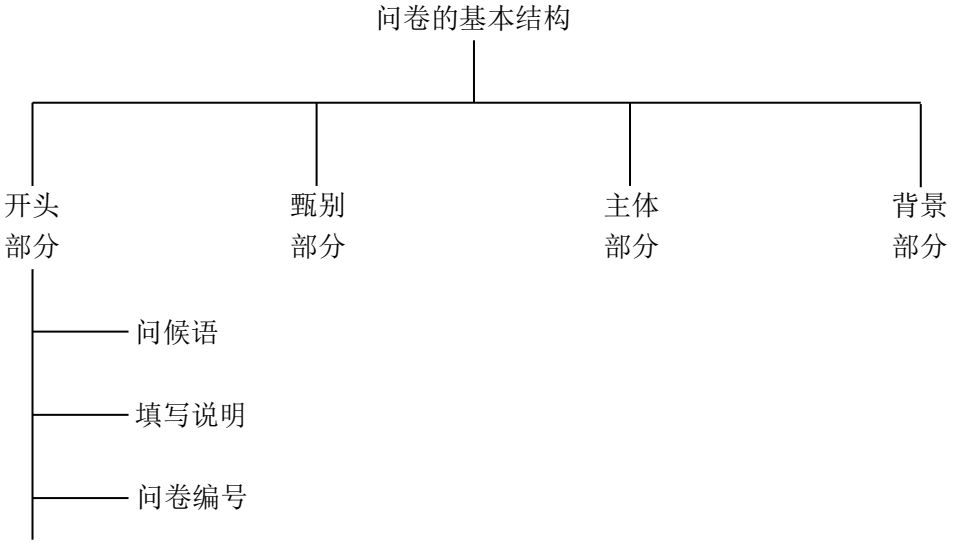


图 2-4 问卷的基本结构

（1）开头部分。开头部分包括问候语、填表说明和问卷编号等内容。

问候语是一封致被调查者的短信，用来向被调查者说明调查主办单位，组织或个人的身份，调查的目的和意义，调查的内容，对被调查者的希望和要求等。篇幅不宜过大，文字要简洁、准确，语气要谦虚、诚恳。在自填式问卷中，好的问候语能引起被调查者对于调查的重视，可以激发其参与调查的意识，消除其顾虑，最终圆满地完成调查，否则容易造成被调查者拒访。好的问候语的语气应亲切诚恳，用词简捷准确。如某高校对学生的一项调查中的问候语：

亲爱的同学：您好！

我们是统计学院 2009 级研究生，正在进行一项有关研究生对导师的看法的调查，目的

是更好的协调师生关系，以便从导师那里得到更多的指导和帮助。您的回答无所谓对错，只要真实反映您的情况和看法，就达到了这次调查的目的，希望您能积极参与。我们对您的回答完全是保密的。填写这份问卷大约耽搁您 15 分钟时间，请您谅解。谢谢您的支持与合作！

填表说明旨在向被调查者说明调查的目的和意义。有些问卷还有填表须知、交表时间、地点及其他事项说明等。问卷说明一般放在问卷开头,通过它可以使被调查者了解调查目的,消除顾虑,并按一定的要求填写问卷。问卷说明既可采取比较简洁、开门见山的方式，也可在问卷说明中进行一定的宣传，以引起调查对象对问卷的重视。在自填式问卷中要有详细的填表说明，如：请您在所选择答案的题号上画圈；对只许选择一个答案的问题只能画一个圈、对可选多个答案的问题，请在你认为合适的答案上画圈；需填写数字的题目在留出的横线上填写；对于表格中选择答案的题目，在所选的栏目内画勾；对注明要求您自己填写的内容，请在规定的地方填上您的意见等。

问卷编码是将问卷中的调查项目变成数字的工作过程，主要用于识别问卷、调查者、被调查者的姓名和地址等，以便于校对检查、更正错误。大多数市场调查问卷均需加以编码，以便分类整理，易于进行计算机处理和统计分析。所以，在问卷设计时，应确定每一个调查项目的编号，为相应的编码做准备。

（2）甄别部分。也称为过滤，是指通过设计一些问题先对被调查者进行过滤，筛选掉不符合条件的被调查者,然后得到满足条件的调查对象。甄别的目的是为确保被调查者合格，能够作为该市场调查项目的代表，从而符合调查研究的需要。例如：

请问您的年龄是:

A. 20 岁以下 .....终止访问

B. 20 岁—30 岁

C. 30 岁—40 岁

D. 40 岁—50 岁

E. 50 岁—60 岁

F. 60 岁以上 .....终止访问

Q2.您本人或您家里是否有从事下列行业的人员?

A.广告 / 市场研究

B.电视台 / 电台 / 报社 / 杂志社

C.酒类生产 / 销售 / 经营

有.....终止访问

无.....继续访问

（3）主体部分。包括了所要调查的全部问题，以及这些问题的所有可供选择的答案。是问卷的主体，是问卷最核心的组成部分。它主要是以提问的形式提供给被调查者，这部分内容设计的好坏直接影响整个调查的价值。

（4）背景部分。说明被调查者的一些主要特征，包括被调查者的性别、民族、职业、收入、文化程度、婚姻状况、家庭人口等。有的问卷还要求写出被调查者的姓名、地址和联系电话等。如果被调查者是单位，还需填写出厂名、店名、地址、负责人、主管部门、职工人数和固定资产原值等情况。以上项目哪些应该列入，应根据调查目的和要求而定。该部分通常放在问卷的最后。该部分所包含的各项问题，可使研究者根据背景资料对被调查者进行分类比较分析。如 2003 年“中国网络游戏调查问卷”中的个人背景资料：

姓名:

性别: (A) 男 (B) 女

婚姻: (A) 已婚 (B) 未婚

年龄: (A) 16 岁以下 (B) 16～18 岁 (C) 19～22 岁 (D) 23～25 岁 (E) 26—～30 岁 (F) 31～35 岁 (G) 35 岁以上

职业分布：

□国家机关工作人员 □企事业管理人员 □技术、生产人员 □学生

□农林牧渔人员 □商业、服务人员 □IT、信息产业人员 □文艺人员

□教育/传媒人员 □卫生医疗人员 □金融、保险人员 □军人

□公共服务人员 □无业 □其它

平均月收入：

□ 500 元以下 □ 501～1000 元 □ 1001～1500 元 □ 1501～2000 元

□ 2001～3000 元 □ 3001～5000 元 □ 5001～10000 元 □ 10000 元以上

□ 在读学生　　□ 无收入

2. 问卷中问题的设计

问卷所要调查的内容由若干个问题组成。如何科学地设计调查的问题在问卷设计中十分重要，正如乔治·盖洛普曾经说过的，没有什么会跟对问题的选择、措词一样困难和重要。问卷中问题的设计应注意以下几个问题：

（1）提问的内容尽可能短。问题中应该坚决摒弃多余的修饰词，提问的内容尽可能地短，这样可以节省调查时间，若问题比较复杂，最好将其分为几个问题来问。例如：

长问题："我国越来越多的人去国外旅游。您曾经去别的国家旅游过吗？如果去过，您也许是为了欣赏风光才去的。那么，别国的风光对您决定出国旅游有多重要？"

短问题："您出国旅游过吗？如果去过，那里的风光对您决定去旅游有多重要？"

（2）用词要确切通俗，避免提不具体的问题。问卷中的用词要确切、通俗，要容易被人理解，应避免使用过于专业的术语；设计的问题要适合所有被调查者；提问目的要明确，避免模棱两可。例如：避免使用通常、常常、一般等字眼，这些词一般很难界定其程度。如：

Q1：您通常几点上班？

这是一个非常不明确的问题。不知其是指何时离家还是在办公地点何时正式开始工作？问题应改为：

Q1：通常情况下,您几点离家去上班？

又如：

Q1：您对本餐厅是否满意？

□1.满意

□2.一般

□3.不满意

该问题过于笼统，很难达到预期效果，可将其具体化，分为三个层面：

	满意	一般	不满意
Q1：您对本餐厅饭菜质量是否满意？	□	□	□
Q2：您对本餐厅环境设施是否满意？	□	□	□
Q3：您对本餐厅服务态度是否满意？	□	□	□

（3）一项提问只包含一项内容。一个问句最好只问一个要点。一个问句中如果包含过多询问内容，会使被调查者无从答起，给统计处理也带来困难。如：“您和您家里人对现有住房条件是否满意？”这个问题的毛病在于“您和您家里人”，即被调查者不是一个人，包括了被调查者的家里人，而每个人由于年龄、性格、职业、文化程度等的不同，对住房条件的要求是不一样的。这一问题至少应该分成两个问题，如：“您对现有住房条件是否满意？”和“您觉得您家里人对现有住房条件是否满意？”

（4）避免诱导性提问。问卷设计中，应避免诱导性、暗示性的提问。如：“绝大多数饮用过明光牛奶的人都认为它口味纯正，您认为是这样吗？”诱导性提问会导致两个不良后果，一是被调查者不加思考就同意所诱导问题中暗示的结论；二是由于诱导性提问大多是引用权威或大多数人的态度，被调查者就会产生心理上的顺向反应。尤其对于一些敏感性问题，在诱导性提问下，被调查者不敢表达其真实想法。因此，诱导性、暗示性提问是调查中的大忌，它常常会引出与事实相反的结论。美国的一个民意调查，对一个问题的两种提法导致了两种不同的结果：一种问法是“艾森豪威尔将军说，陆军部和海军部应当合并为统一的作战部”，结果同意的比例为 49%；另外一种问法只是把艾森豪威尔的名字去掉，结果同意的比例为29%，两者之间有显著性的差异，权威者的姓名在这里产生了诱导的作用。因此，问卷设计中要坚持客观的态度，避免使用带有感情色彩的词语或带有权威性的语句。

（5）避免否定形式的提问。否定式的提问会影响到被调查者的思维，或容易造成相反意愿的回答。如：

Q：您不认为听到国歌不立正不是不对的吗？

□是

□不是

双重否定的问题难以让被调查者很快地理解，所以此问题应改为：

Q：您认为听到国歌不立正是否正确？

□是

□否

（6）避免敏感性问题。所谓敏感性问题，是指与个人或单位的隐私或私人利益有关而不便向外界透露的问题。问卷中要尽量避免提问敏感性问题或容易引起人们反感的问题，对

有些必须涉及敏感性问题的调查，应当在提问的方式上进行推敲，尽量采用间接询问的方式，用语也要特别婉转，以降低问题的敏感程度。

例如我们想知道某高校大学生中到底有多少人在考试中作弊？要调查这一问题，若直接询问你是否在考试中作弊？为了保护自己的隐私或出于其他目的，绝大多数学生都会作否定的回答或拒绝回答，因为涉及社会道德规范或个人隐私的问题往往是被调查者所不愿回答的，类似问题还有："您是否有偷税漏税行为？"、"您的收入是多少？"、"您的体重是多少？"等。遇到敏感性问题应采取迂回的问法（indirect question），才能探知被调查者真正的想法。例如：想探知学生是否有作弊行为，可试用以下问句：

Q1：有些同学考试作弊，您认为他们最主要的原因是

A 为了考试及格或取得更好成绩

B 别人都作弊，自己不抄太亏了

C 心存侥幸的心理

Q2：您同意他们的看法吗？

□同意

□不同意

总的来说，对于敏感性问题，若采用直接调查的方法，调查者将难以控制样本信息，得不到可靠的样本数据，为了得到敏感性问题的可靠的样本数据，还有必要采用一种科学可行的技术——随机化回答技术（Randomized Response Technique 简记为 RRT，它在调查中使用特定的随机化装置，使得被调查者以预定的概率 P 来回答敏感性问题。这一技术的宗旨就是最大限度地为被调查者保守秘密，从而取得被调查者的信任）。

3. 问卷中答案的设计

问卷中的回答项目是针对提问项目所设计的答案。由于问卷中的问题有不同的类型，所设计的答案类型对被调查者的回答要求也是不同的。

开放性问题是指对问题的回答未提供任何具体的答案，由被调查者根据自己的想法自由做出回答，属于自由回答型。开放型问题的优点：气氛轻松；适合于某些意愿调查；可以使被调查者充分表达自己的意见和想法；有利于被调查者发挥自己的创造性等。开放型问题缺点：资料整理有一定困难；资料难以量化，因此无法作深入的统计分析；容易"跑题"，导致成本提高等。封闭型问题是指对问题事先设计出了各种可能的答案，由被调查者从中选择。封闭型问题的优点：非常简明，深入到问题的实质，比较经济。封闭性问题的缺点：答案比较简单，有时难以完全覆盖现实中的复杂情况。

市场调查中的问卷大都由封闭型的问题组成，封闭型问题答案的设计方法主要有：两项选择法、多项选择法、顺序选择法、评定尺度法、双向列联法等。

（1）二项选择法：二项选择法也称二分法，提出的问题只有两种答案可以选择，"是"

或“否”,“有”或“无”等。这两种答案是对立的、排斥的，被调查者的回答非此即彼，不能有更多的选择。例如“您家里现在有空调吗?”答案只能是“有”或“无”。这种方法的优点是易于理解，可迅速得到答案，便于统计。但被调查者没有进一步阐明细节和理由的机会，难以反映被调查者意见与程度的差别，了解的情况也不够深入。这种方法适用于相互排斥的两项择一的问题，以及询问较为简单的事实型问题。这种问题的形式如下：

您是否购买了笔记本电脑?

A、是　B、否

（2）多项选择法。为了充分表达要求和意愿，有些问题还需要采用多种选择，即采用选择多个答案的问题，统计出多个答案的重要性的差别。多项选择题是从多个备选答案中选择一个或几个，它是各种调查问卷中采用最多的一种类型。如：

Q1：你买山地车的目的是（可多选　）

A、经济条件许可　B.用于代步工具　C、便于去郊外旅游，锻炼身体　D、别人有你也想有，赶时髦　E、作为礼物送给亲人朋友　F、其它

Q2：您的家庭是否拥有下列物品？（请在选项上打“√”）。

□汽车

□信用卡

□第二栋房子

□高级音响

□数码像机

□摄影机

□影碟机

（3）顺序选择法。顺序选择法的问题列出若干个答案，要求被调查者按其重要性或记忆的先后顺序将它们一一排列。顺序选择法一是可以对全部答案排序；另一是可以对其中的某些答案排序。对所选择的答案数量可以进行限制，也可以不进行限制。例如某高校学生就业情况调查表中的问题：

Q：您在找工作的过程中遇到的主要问题是（请您依次排序）（　　）

A 专业不对口　B 没有本地户口　C 缺乏社会关系　D 招聘信息不足　E 性别歧视　F 其他

（4）评定尺度法。也称量表法，量表是一种工具，是将一些主观的、抽象的概念定量化。在问卷的设计中，调查者要考虑被调查者对问题的回答是否便于进行量化统计和分析。如果使用问卷的调查结果是一大堆难以统计的定性资料，那么要从中得到规律性的结论则十分困难。在调查中，量表通常是将态度量化的工具，量表的方式很多，如：

问题：　请您为我校行政管理人员的整体素质打分

A、5　　B、4　　C、3　　D、2　　E、1

（5）双项列联法。这种方法是将两种不同的问题综合一起，通常用表格的形式来表现，它可以提供单一类型问题无法提供的信息，还可以节省问卷的篇幅。例如某高校有关课程设置情况调查表中的问题答案设计：

请您对以下各题予以客观评价，并在适当的□中打「√」	非常满意	比较满意	一般	不太满意	非常不满意
1.你对所学专业的满意程度	□	□	□	□	□
2.你对每学期课程安排的满意程度	□	□	□	□	□
3.你对所学专业实践教学的满意程度	□	□	□	□	□
4.你对我校教师教学质量的满意程度	□	□	□	□	□
5.你对本专业教师教学方法的满意程度	□	□	□	□	□
6.你对我校教师使用多媒体的满意程度	□	□	□	□	□
……	□	□	□	□	□

4. 问题顺序的设计

设计问题的顺序应注意以下几点：

（1）问题的安排应具有逻辑性。问题应该按照逻辑顺序一个接一个。最好采用从一般到具体的提问方式，因为这是多数人表现出来的思维方式，一般性问题置前，特殊性问题置后。问卷的设计要给人以上下连贯的逻辑性，使问卷形成一个相对完善的小系统。如：

A. 请问您最近买过什么牌子的口香糖?

B. 您喜欢什么味道的口香糖?

C. 您是否看过口香糖的广告?

D. 在看过的广告中印象最深的一幅画面是什么?

以上几个问题设计紧凑连贯，从而能获得较完满的回答。

（2）问题的安排应先易后难、由浅入深。按照先易后难的基本原则，把简单的、受人关注的、有趣的和容易回答的问题放在前面，复杂的、较难的问题放在后面，开放性问题和敏感性问题放在最后。

在调查问卷中，设计好第一个问题很重要，第一个问题应与调查的目的直接相关，它要适用于所有被调查者。如果第一个问题不适用，被调查者会觉得调查与自己没有关系。第一个问题也应设计得容易回答。敏感性问题问得太早，被调查者会不愿意回答，如果将其置于一份很长的问卷的最后，由于被调查者的疲劳也会影响答案的质量。所以，敏感性问题应该在被调查者感到轻松以及与其他问题的联系最有意义的地方出现。被调查者的背景资料通常用于分组，以便对搜集到的信息进行比较。在住户调查以及许多社会调查中，反映被调查者本人或家庭背景资料的问题通常放在问卷的最后。

（3）问卷主体部分的问题通常按过滤性、热身性、容易性、困难性的顺序进行排列。此外，即使是一份很成功的调查问卷，也必须经历实践的考验，所以在问卷初步设计完成时，为了观察包括问卷整理在内的调查过程各个阶段，需要进行预调查。预调查是对正式调查的一次“预演”，是从头至尾进行一次小规模的调查演习，包括数据处理和分析。所以，问卷设计是一个反复的逐渐完善的过程，只有这样才能够达到调查的终极目的。

小　结

1. 从统计数据本身的来源看，主要有两种渠道：一是直接的调查和科学实验，对使用者来说，这是数据的直接来源，称之为一手数据或原始数据；二是来源于别人调查或实验的数据，对使用者来说，这是数据的间接来源，称之为二手数据或间接数据。

2. 统计调查是取得社会经济数据的主要来源，也是获得直接统计数据的重要手段。实际中常用的统计调查方式主要有抽样调查、普查、统计报表等，实际工作中，采用何种调查方式，必须根据调查研究的目的和具体条件进行选择。

3. 根据不同的抽选样本方法，有两种类型的抽样：概率抽样和非概率抽样。

概率抽样的基本抽样组织方式主要有简单随机抽样、分层抽样、系统抽样、整群抽样、多阶段抽样。

4. 非概率抽样是用主观的（非随机的）方法从总体中抽选样本。非概率抽样方法有方便抽样、判断抽样、配额抽样和滚雪球抽样。

5. 抽样误差是由于抽取样本的随机性造成的样本值与总体值之间的差异,即由样本结果估计总体特征所产生的误差。调查中还会出现由各种原因引起的与样本抽取无关的误差，即非抽样误差。

6. 不论统计调查采取何种方式进行调查，在取得统计数据时，都有一些具体的数据收集方法。数据的收集方法归纳起来可分为询问调查和观察实验两大类，询问调查是调查者与被调查者直接或间接接触以获得数据的一种方法，具体包括访问调查、邮寄调查、电话调查、电脑辅助调查、座谈会、个别深入访谈等。

7. 为了及时准确获得被研究现象的实际资料，在数据收集之前，必须设计调查方案，调查方案设计的好坏直接影响到调查数据的质量。在设计调查方案中，问卷设计是使调查者能顺利地获取必要的信息资料,以便于统计分析的一种手段。问卷设计是科学与艺术的结合。

8.数据质量的好坏直接影响统计分析的结果，数据的真实性是统计工作的生命。调查数据的误差可以来自许多方面，归纳之可分为两大类：抽样误差和非抽样误差。

思考与练习

1. 简述普查的概念、特点。
2. 简述抽样调查的特点及适用范围。
3. 概率抽样有哪些基本抽样组织方式？
4. 简述简单随机抽样的优、缺点。
5. 简述非概率抽样与概率抽样的区别。
6. 简述统计数据的具体搜集方法。
7. 简述调查方案的基本结构。
8. 问卷的结构怎样？
9. 问卷设计应遵循哪些原则？
10. 应当如何设计问卷中的问题与答案？
11. 设计问卷中的问题顺序应注意的问题？
12. 非抽样误差的种类有哪些？
13. 为研究在校大学生的学习或生活情况设计一份调查方案。
14. 科南调查公司过去两个月接受了 3 个调查项目，各个客户的具体要求如下：
（1）电视台最近在黄金时段新播出一部电视连续剧，电视台有关部门想了解该电视剧的收视率情况及电视观众对其的评价。

（2）某某月刊是一本发行量很大的杂志，该杂志社想了解其读者群体在年龄结构、收入水平、文化水平等方面的特征及他们对该杂志的改进意见。
（3）某生产厂家在推出一款新手机之前，打算了解消费者对于该手机可能具有的偏好、认知度以及对该产品价格方面的意见。
请你为科南调查公司就以上各个调查项目选择合适的搜集数据方式并说明你的理由。

第三章　数据的描述

正如第一章所述，数据可以分为定性数据和定量数据：定性数据是指用文字、符号、语言等描述的信息；定量数据是指对事物按照某种特征进行具体的数量描述。在第二章，我们讨论了收集数据的方法。

数据搜集完成后，我们需要从中分析有用的信息。从数据文件中可以直接浏览数据，但随着数据量的增加，很难只从表面理解其包含的全部意义。因此，有必要使用其他方法帮助我们从数据中提取信息，并转化成可利用的形式。描述统计方法就是完成这些工作的有力工具。数据的描述统计方法通常包括三种：为数据作合适的统计图；为数据制作恰当的统计表；根据数据计算有代表性的数值。

第一节　统计图与统计表

一、频数分布与统计分组

频数分布（分布数列，Frequency Distribution）反映的变量的取值在各个组中的分布状况，是归纳与总结数据的一种重要方式。依据研究的目的将数据分成若干组，统计各个组中数值的个数，就可以得到频数分布。例如，按照考试成绩把学生分为优、良、中、及格、不及格五个等级，再汇总各组中的学生人数，就可以得到考试成绩的频数分布。定性数据和定量数据的频数分布是绘制很多统计图表的基础。

1. 定性数据的频数分布

在这部分，我们利用例3.1中的普通居民家庭购买笔记本电脑品牌的数据来说明如何构建和解释定性数据的频数分布问题。

【例3.1】目前笔记本电脑已经越来越多地进入了普通老百姓的家中，其中，联想、惠普、三星、华硕和索尼五个品牌受到了普通老百姓的喜爱。本数据展示了50个家庭购买这五种品牌笔记本电脑的样本。

为了构造该数据的频数分布，需要统计每种品牌笔记本电脑的购买次数，如表3-1所示。联想出现19次，惠普出现8次，三星出现5次，华硕出现13次，而索尼出现5次。频数分布说明了在50个家庭购买的笔记本电脑样本中，5种品牌是如何分布的。从表3-1可以看出，联想排在第一位，华硕排在第二位，惠普排在第三位，三星和索尼并列第四。频数分布也展示了这5种笔记本电脑在老百姓心目中的受欢迎程度。

表 3-1　5 种品牌笔记本电脑购买次数的频数分布表

品牌	频数	频率
联想	19	38%
惠普	8	16%
三星	6	12%
华硕	13	26%
索尼	4	8%
合计	50	100%

用每一组的频数除以总的频数n得到的相对数称为频率，常用百分数表示。也称为相对频数或百分数频数。表3-1给出了5种品牌笔记本电脑购买情况的频数和频率分布。

2 定量数据的频数分布

定量数据分组相对于定性数据略为复杂。定量数据的分组要确定三个要素：组数、组距和组限。具体步骤如下。

（1）确定组数

一般来说，定量数据可以分成5-15或20个组。为了更好地展示数据的变异情况，如果数据较少，用5或6组就可以；如果数据较多，可以使用较多的组数。另外，我们还可以根据经验公式确定分组组数，分组组数K应满足：

$$2^K > n，即 K = 1 + \frac{\lg(n)}{\lg(2)} \quad (3\text{-}1)$$

【例3.2】某一个会计师事务所，对其一个包含20个客户的样本，完成年终审计所需的天数如下：12，15，20，22，14，14，15，27，21，18，19，18，22，33，16，18，17，23，28，13。试确定分组的组数。

由于审计天数样本相对较少（n=20），可用5组构建频数。因为$k = 1 + \lg(n)/\lg(2) = 5.32$，因此，可以确定为5组。

（2）确定组距

定量数据频数分布的第二步是确定组距。一般来说，建议每组的组距相同。确定组距的方法是：先找出定量数据中的最大值和最小值，根据第一步中确定的组数得到近似组距。即

近似组距=（数据最大值-数组最小值）/组数 （3-2）

然后将组距确定为大于近似组距的一个整数就可以了。在实际中组距一般为5或10的倍数。例如，近似组距为9.38，可以取整为10，这样以10作为组距在构建频数时更方便。对于审计天数样本数据，最大值是33，最小值是12。因为组数为5，所以由式（3-2），可以计算出近似组距为（33-12）/5=4.2。因此，在频数分布中以5天作为组距。

（3）确定组限

组限的选择必须保证每一个数据属于且只属于一组。下组限是分配到该组的数据的最小值，上组限是分配到该组的数据的最大值。在构建定性数据的频数分布时，不需要规定组限，因为每一个数据会自动落入独立的组中。而对于定量数据，组限就是必要的，以确定每个数据的归属。

关于组限的确定有三种情况。

①. 上下组限间断，互不重叠。例如，对审计天数样本，我们对第一组选择10天为下组限、14天为上组限，该组在表3-2中标记为10～14。最小数据12包含在10～14组。然后，对下一组选择15天为下组限、19天为上组限，依此类推，直到获得全部的5个组：10～14、15～19、20～24、25～29和30～34。最大数据33包含在30～34组。

表 3-2　审计天数数据的频数分布表

组限	频数	频率
10～14	4	20%
15～19	8	40%
20～24	5	25%
25～29	2	10%
30～34	1	5%
合计	20	100%

②. 上下组限重叠，上组限不在内。例如，上面的例子中组限可以设为10～15，15～20，20～25，25～30，30～35。这时15这个数值要归入15～20这一组，依此类推。对于连续变量往往使用这种方式。

③. 使用开口组。即最小组用“…以上”，最大组用“……以下”表示。例如在对考试

成绩进行分组时第一组常用“60以下”来表示。在数据中存在特别大或者特别小的数值时常使用开口组。

相邻两组的下组限之差就是组距。一旦确定了组数、组距和组限，通过统计属于每一组的数据项的个数就可以得到频数分布。例如，【例3.2】的数据显示，有4个数（12、14、14和13）属于10～14组，因此10～14组的频数是4。对于剩余几组进行同样计算，就可以得到审计天数的频数分布表3-2。该表显示，最频繁发生的审计天数处于15～19天这一组，在20个审计天数中，有8个属于这一组；只有一次审计需要30天或更多时间。

在频数分布的基础上，定量数据的另一种汇总方式是累积频数分布。频数的累计可以从小到大进行，也可以从大到小进行，我们这里只给出从小到大累积的情况，这时称为小于式的累积频数分布，表示的是小于或等于每一组上组限的数据个数，而不是每一组的频数。表3-3给出了审计天数的累积频数分布。

以“小于或等于24”的组为例来说明累积频数分布的确定。这一组的累积频数是数据中小于或等于24的所有组的频数之和。对表3-2中的频数分布来说，组10～14、15～19和20～24的频数之和为4+8+5=17，它表明有17个数据小于或等于24。因此，该组的累积频数是17。另外，表3-3中的累积频数表明有4次审计在14天内完成，有19次审计在29天内完成。

表 3-3　审计天数数据的累积频数分布表

审计天数	频数	累积频数	累积百分数频数
小于等于 14	4	4	20%
小于等于 19	8	12	60%
小于等于 24	5	17	85%
小于等于 29	2	19	95%
小于等于 34	1	20	100%

用累积频数除以数据总数，可以计算出累积频率频数分布。累积频率分布显示，有85%的审计在24天内完成，有95%的审计在29天内完成。

二、列联表

列联表是归纳两个变量数据的方法。我们通过例3.3对农民工月收入水平和收入等级自我评价的数据来说明列联表的用途。

【例3.3】某县100名外出打工农民工组成的一个样本，搜集他们的月收入水平和文化程度数据。其中，收入水平是一个定量变量，其变化范围是1000-4900元；文化程度是一个定性变量，等级类别有小学及以下、初中和高中及以上。

该数据的列联表如表3-4所示。左边和顶部的栏目规定了两个变量的组别。在左边，纵向栏目（小学及以下、初中和高中及以上）对应着文化程度变量的三个组。在顶部，横向栏目（1000-2000元，2000-3000元，3000-4000元和4000-5000元）对应月收入水平变量的四个组。样本中的每个农民工都给出了月收入水平和文化程度。因此，样本中的每个农民工都与列联表中的某一列和某一行的交叉单元相联系。我们只须简单地计算出属于交叉分组表每个单元的农民工数就可以完成列联表的构建。

从表3-4可以看出，样本中的绝大部分农民工（19人）具有初中文化程度且月收入水平在3000-3900元之间，只有1个农民工具有高中及以上的文化程度且月收入水平在1000-1900元之间，其他频数也可以进行类似的解释。另外，我们注意到列联表的右边和底部分别给出了农民工文化程度和月收入水平的频数分布。从右边的频数分布中，可以看到有25个人文化程度为小学及以下、49个为初中、26个为高中及以上。类似地，底部显示出月收入变量的频数分布。

表 3-4　100 名农民工月收入水平和收入等级的列联表

文化程度	月收入水平（元）				合计
	1000-2000	2000-3000	3000-4000	4000-5000	
小学及以下	12	13	0	0	25
初中	12	16	19	2	49
高中及以上	1	8	9	8	26
合计	25	37	28	10	100

把表3-4中的项目转换成行百分数或列百分数，能够提供变量间关系的其他信息。我们以行百分数来加以说明。表3-4中的每个频数除以对应的行合计数，所得到的结果显示在表3-5中。从表3-5可以看出，小学及以下的农民工的收入集中在1000-2000元和2000-3000元两个组；而高中及以上的农民工的收入则主要集中在2000-3000元、3000-4000元和4000-5000元的组中。因此，可以看到较高的文化程度与较高的收入等级相联系。

表 3-5　每一个收入等级类的行百分比（%）

文化程度	月收入水平（元）				合计
	1000-2000	2000-3000	3000-4000	4000-5000	
小学及以下	48.0	52.0	0.0	0.0	100
初中	24.5	32.7	38.8	4.0	100
高中及以上	3.8	30.8	34.6	30.8	100

三、常用统计图

1．条形图和圆形图

对定性数据描述的统计图主要有条形图（Bar Chart）和圆形图（Pie Chart）。它们主要用来描绘变量中各类观测值的频数分布和比例。

条形图是用垂直或水平的条形来描述数据分布的一种方法。条形图一般用横轴表示对数据进行分组的标记，纵轴表示次数、百分比等。绘制条形图就是将固定宽度的长条放在每一组的标记上，将这个长条的高度向上延伸直至达到该组的频数（或者频率）。对于定性数据，应将这些长条分开，以说明每一组都是相互独立的。图3-1是品牌笔记本电脑购买次数条形图。该图展示了联想、惠普、三星、华硕和索尼五个品牌笔记本电脑受老百姓欢迎的程度。

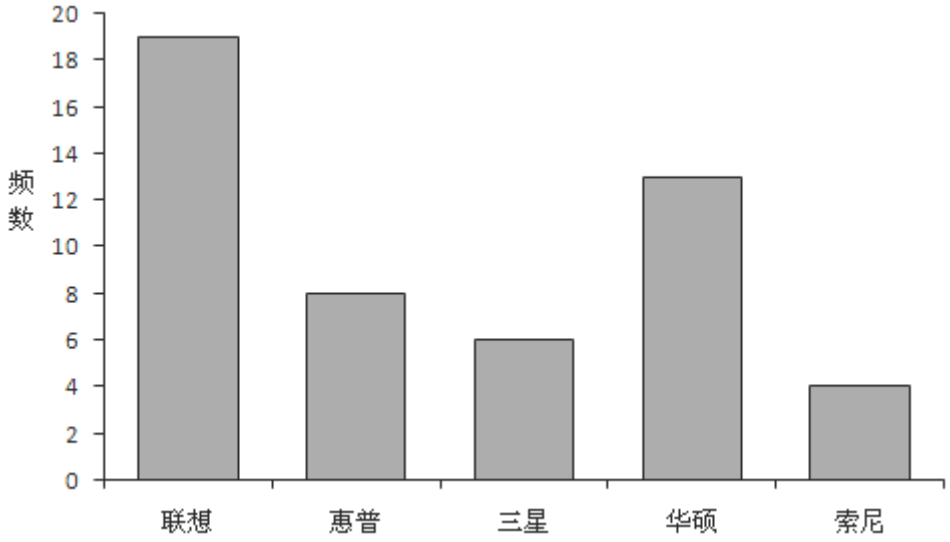


图 3-1　品牌笔记本电脑购买次数条形图

圆形图也称为饼图。整个圆形的面积代表所研究数据的整体，每一个扇形代表每一个子集所占的百分比。圆形图给出了部分相对于整体的比例。圆形图在商业研究中使用广泛，尤其适合于描述预算、市场份额和时间及资源的分配等等。画出圆形图的关键是确定每一个子集相对于整体的比例。图3-2是品牌笔记本电脑购买次数的圆形图。从图3-2中可以看出5种品牌笔记本电脑的购买比例。从另一个角度看，如果已知总的购买次数，就可以推算各种品牌笔记本电脑的购买次数。

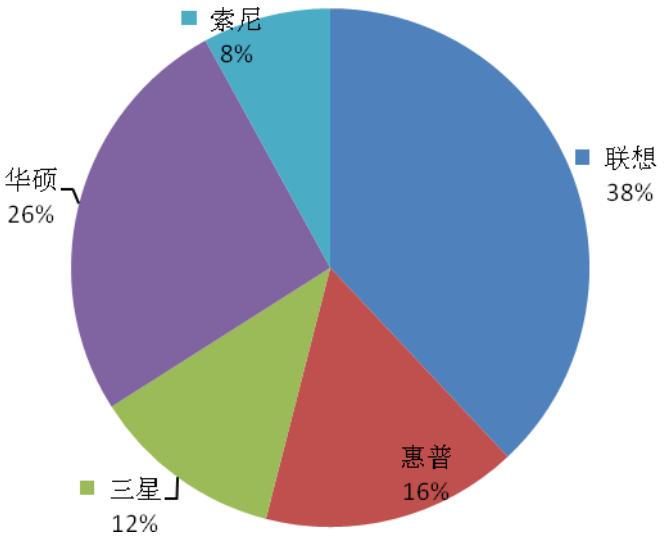


图 3-2　品牌笔记本电脑购买次数圆形图

注意，（1）条形图中的纵轴和圆形图中的扇形面积可以是频数或频率。（2）在频数分布中，组别数与数据中的类别通常是一致的。例3.1的数据只涉及5种品牌笔记本电脑，每一种笔记本电脑定义为一个独立的频数分布组。在实际应用问题中，市场上笔记本品牌有很多，其中大部分的类别只有很少的购买次数。在这种情况下，为了简化数据结构，可以把频数较小的组合并为“其他”组。频率小于5%的组通常这样处理。（3）在频数分布中，频数的总和等于观测值数目；在频率分布中，百分比的总和等于100%。

2．直方图

直方图（Histogram）是定量数据最常用的图形描述方式。其横轴是标注的通常是定量变量的各组界限；纵轴是各个分组对应的频数或频率。构建直方图时，每组的次数用一个长方形绘制，长方形的底在横轴上，以组距为底，以每组相应的频数或频率为高。下面用例3.2的数据来说明。

图3-3是审计天数的直方图。该图中最大频数的组由15～20天这一组的长方形表示，长方形的高度表示这一组的频数是8。注意，直方图中邻近的长方形是相互连接的，长方形之间没有间隔。这是直方图与与条形图的区别之一。

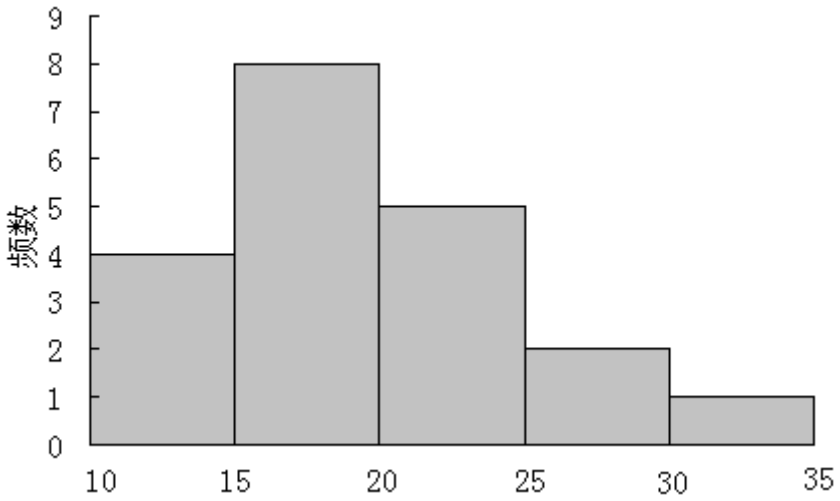
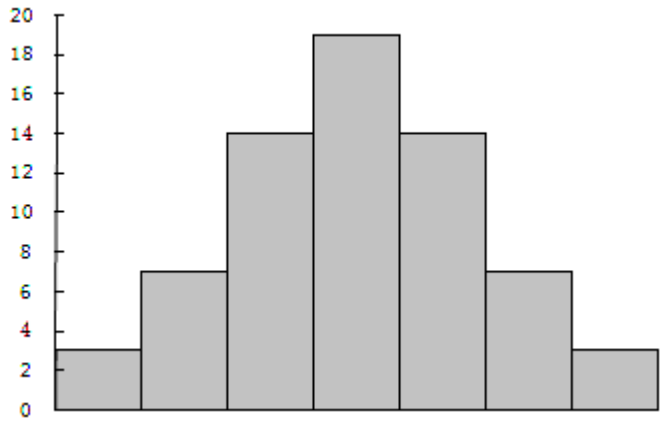
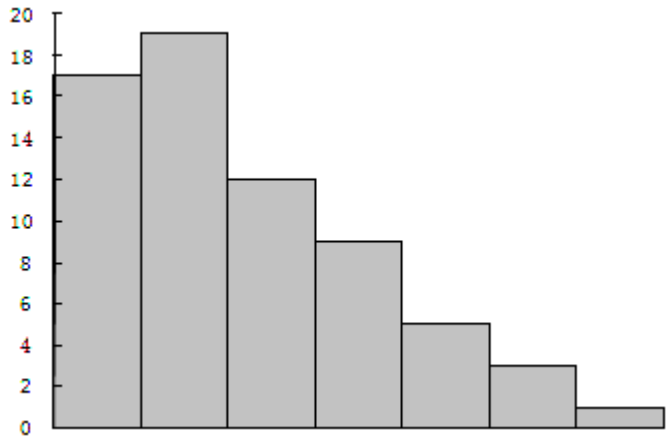


图 3-3　审计天数的直方图

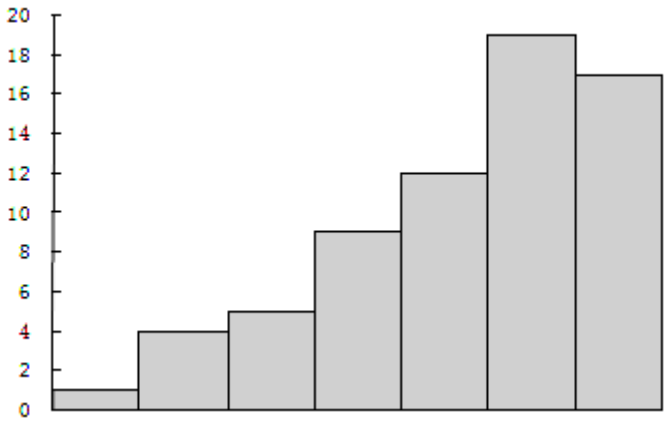
直方图的一个重要应用是用来描述数据分布的形态。图3-4是三个直方图。图3-4a显示的是一个对称的直方图。在对称直方图中，左尾和右尾形状相同。在实际应用中得到数据的直方图不会绝对对称，很多直方图是近似对称。例如，人的身高和体重数据通常都是近似对称的。如果图形尾部向右延伸一些，则称图形为右偏。图3-4b显示的直方图有一定程度的右偏。右偏的典型例子是商品房住宅价格数据，其原因主要是少数昂贵住宅造成右尾偏斜。如果图形的尾部向左延伸一些，则称图形为左偏。图3-4c显示的直方图有一定程度的左偏。学生考试成绩是这种直方图的典型应用。由于考试成绩大多数分布在70%以上，只有极少数的成绩很低。



（a）对称



（b）右偏



（c）左偏

图 3-4　呈现不同偏度水平的直方图

3．折线图

折线图(Frequency polygon)也称频数多边形图，是在直方图的基础上，把直方图顶部的中点(组中值)用直线连接起来，再把原来的直方图抹掉得到的。折线图的两个终点要与横轴相交，具体的做法是第一个矩形的顶部中点通过竖边中点（即该组频数一半的位置）连接到横轴，最后一个矩形顶部中点与其竖边中点连接到横轴。组数越多，组距就越小，折线图就越光滑，逐渐形成一条平滑的曲线，这就是频数分布曲线。

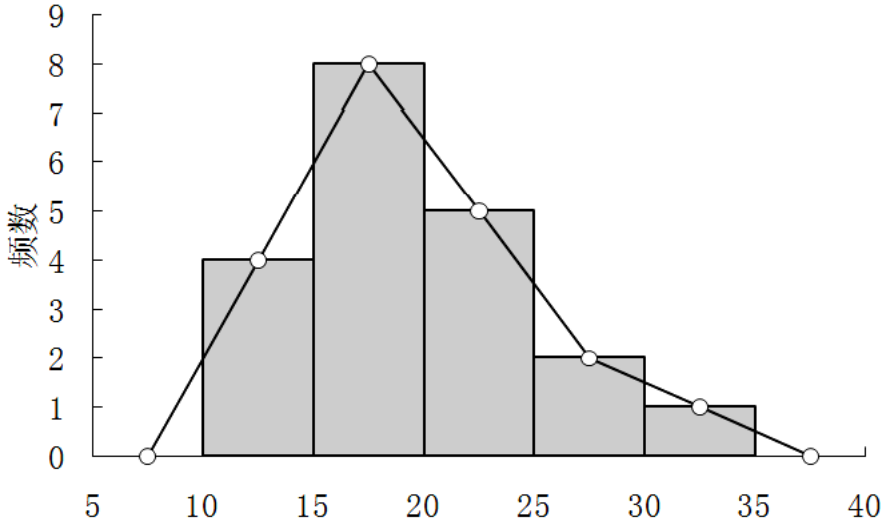


图 3-5　审计天数折线图

4．茎叶图

茎叶图（Stem-and-leaf plot）可以同时展示数据的分布形状和原始数据。茎叶图主要用于显示未分组的原始数据的分布。由“茎”和“叶”两部分构成，其图形是由数字组成的。通常以数据的高位数值作树茎，低位数字作树叶，树叶上只保留一位数字。树叶的竖列要对齐，以计算各组的次数。下面以例3-4说明茎叶图的使用。

【例3.4】50个学生的统计学课程考试得分数据。茎叶图如图3-6所示。

统计学得分 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
2.00	4 .	13
3.00	4 .	667
3.00	5 .	003
2.00	5 .	89
5.00	6 .	01144
5.00	6 .	57799
7.00	7 .	0001344
8.00	7 .	55666788
6.00	8 .	011344
5.00	8 .	57789
3.00	9 .	022
1.00	9 .	7

Stem width: 10

Each leaf: 1 case(s)

图 3-6　统计学课程考试得分茎叶图

在图3-6的右半部分的茎和叶中，左边的茎代表得分的第一位数字，右边的叶代表得分的第二位数字，处于该得分段的分数有几个则列出几个。为清楚起见，每5分被分成了两部分，例如，第一行的4代表40-44分，第二行代表45-49分。图3-6中左半部分频数Frequency表示该分数段的分数个数。例如，第一行表示在40-44分这一组中，有两个同学位于这一组。

通过茎叶图我们可以清楚的看出观测值的分布。同时，茎叶图还保留了原始数据。

注意茎叶图中计量单位对结果的影响。图3-6中读出的第一个数字为4.1，乘以stem width (10)，说明实际代表的为41。如果stem width 标注为100，则实际代表410。

由于每一个观测值都会在茎叶图中占据一个空间，所以，当变量的观测值很多时，茎叶图的效果就不是太好了，这时最好直接采用直方图。

5．线图

线图(Line Chart)主要利用线形的升降起伏来表现描述的变量在一段时期内的变动情况，主要用于显示时间数列的数据。通常横轴表示时间，纵轴表示另一个变量。横轴上每一个时间都对应纵轴上变量的一个取值。

【例3.5】根据1990年以来中国的居民消费价格指数序列绘制的线图。

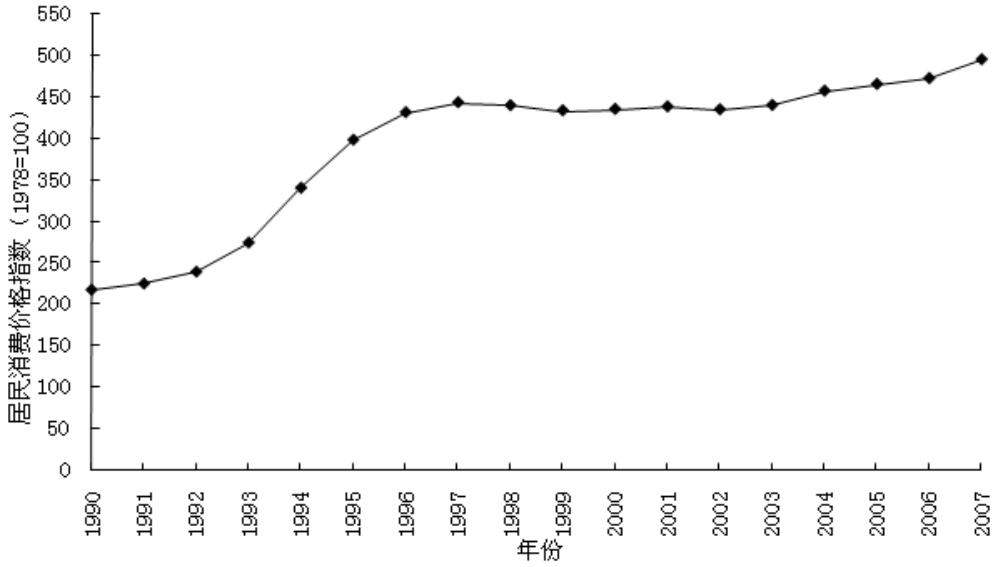


图 3-7　中国 1990-2007 年居民消费价格指数

图3-7展示了中国1990年以来的居民消费价格指数（以1978年为基期），其中横轴表示年份，纵轴表示价格指数。从该图可以看出，1990年以来，中国的居民消费价格指数大致呈上升趋势，期间各个时期具有一定的波动趋势。

四、绘制统计图应注意的问题

绘制一个恰当的统计图需要注意以下几个方面的问题：

第一，通过选择恰当的图形类型、刻度、长宽比例等，使图形能够准确反映数据中包含的信息。时间一般绘在横轴，指标数据绘在纵轴；长宽比例要适当，其长宽比例大致为10：7；一般情况下，纵轴数据下端应从“0”开始；数据与“0”之间的间距过大时，可以采取折断的符号将纵轴折断。

第二，图形要尽量简明。图形应该突出所要传达的信息，不必要的标签、背景、网格线、等会分散读者的注意力。

第三，图形应该有清楚的标题和必要的说明，明确图形的含义、计量单位、坐标轴代表的变量、资料来源等等。

第四，反复加工和修改是获得优秀统计图形的重要步骤。统计软件给出的统计图形没有多少可以不加修改而直接应用。

五、统计表

数据搜集后按照一定的要求进行整理、归类，并按照一定的顺序把数据排列起来，制成表格，便形成统计表（statistical table），它是由纵横交叉线条所绘制的表格来表现统计数据的一种形式。统计表能够：（1）说明研究对象之间的相互关系；（2）把研究对象之间的变化规律显著地表示出来；（3）把研究对象之间的差别显著地表示出来。这样便于人们分析和研究问题。

统计表的内容一般都包括表头、行标题、列标题、数字资料、单位和制表日期等。表头是表的名称，它要能简单扼要地反映出表的主要内容；行标题是指每一横行内数据的意义；列标题是指每一纵栏内数据的意义；数字资料是指各空格内按要求填写的数字；单位是指表格里数据的计量单位。在数据单位相同时，一般把单位放在表格的左上角。如果各项目的数据单位不同，可放在表格里注明；制表日期放在表的右上角，表明制表的时间；统计表为“开口式”，左右两端不封口；表的上下两条横线一般用粗线，其他线用细线，线条要少；很多

统计表都有“备注”或“附注”，以便对表格的内容进行补充说明。表3-7展示了2003年以来中国就业的基本情况，从表中我们可以对应分析统计表的各种要素。

表 3-7　中国基业基本情况

项　　目	2003	2004	2005	2006	2007
经济活动人口(万人)	76075	76823	77877	78244	78645
就业人员合计(万人)	74432	75200	75825	76400	76990
在岗职工人数(万人)	10492	10576	10850	11161	11427
国有单位	6621	6438	6232	6170	6148
城镇集体单位	951	851	769	726	684
其他单位	2920	3287	3849	4264	4595
城镇登记失业人数(万人)	800	827	839	847	830
城镇登记失业率　　(%)	4.3	4.2	4.2	4.1	4.0

注：本篇章就业人员合计 1990 年至 2000 年数据根据第五次人口普查资料重新调整，2001 年及以后数据根据人口变动抽样调查资料推算，因此，与相应年份的分地区、分登记注册类型、分行业资料的分项数据之和不一致。

第二节　数据描述的数值方法

在第一节，我们讨论了用统计图和统计表来描述数据的统计方法。除了图表外，还可以用数值方法来汇总数据。这一节我们将介绍描述数据的集中趋势、离散程度、分布形态的数值方法。

一、集中趋势

集中趋势（central tendency)是一组数据向其中心值靠拢的倾向和程度。测度集中趋势就是寻找数据一般水平的代表值或中心值。常用的集中趋势的测度指标有：算术平均数、调和平均数、几何平均数、众数、中位数、分位数等。

1. 算数平均数

（1）简单算数平均数

在集中趋势的数值度量中，最常用的就是平均数，统计上叫做均值（mean）。根据样本数据计算的平均数称为样本均值（sample mean），用 $\bar{x}$ 表示；根据总体数据计算的平均数称为总体均值，用 μ 表示。用 x_i 表示变量 x 的第 i 个观测值，则 n 个观测值的样本均值可以用式（3-3）计算。

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{3-3}$$

以前面【例3.2】中审计时间的数据为例，20次审计的平均审计时间为（12+15+……+13）/20=19.25天。

用N表示总体观测值的个数，则总体平均数的计算公式如式（3-4）所示。

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{3-4}$$

平均数的优点在于它利用了变量的每一个观测值，能够获得更多的信息。但另一方面，正因为使用了每一个观测值，所以平均数的计算相对麻烦，而且会受到极端值的影响。这是平均数的一个重要弱点。

（2）加权算术平均数

前面我们介绍的简单算数平均数，其计算公式中每个 x_i 都有相同的重要性或权重。实践中，有时每个观测值的重要性不一定相同，此时，计算平均数时就需要对不同的观测值赋予相应的权重，以这种方式计算的平均数称为加权平均数。

对于有*n*个观测值的样本，假设观测值 x_i 的权重为 w_i，则加权平均数的计算公式如下：

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3-5}$$

当数据来自总体时，总体的加权平均数用 μ 取代 $\bar{x}$。

【例3.6】大学生的平均学分绩点（GPA）。某大学生在结束了60个学分的课程学习后，有9个学分获得A，15个学分获得B，33个学分获得C，3个学分获得D。已知成绩为A、B、C、D时的学分绩点分别为4、3、2、1，计算这个学生的平均学分绩点。

在本例中，以每个等级的学分绩点作为观测值，则 $x_1 = 4$， $x_2 = 3$， $x_3 = 2$， $x_4 = 1$；以每个等级相应的学分数作为权重，则 $w_1 = 9$， $w_2 = 15$， $w_3 = 33$， $w_4 = 3$。利用式（3-5），可计算平均学分绩点如下：

$$\bar{x} = \frac{4\times 9 + 3\times 15 + 2\times 33 + 1\times 3}{9 + 15 + 33 + 3} = 2.5$$

加权平均数的计算结果表明，该学生的平均学分绩点为2.5。

2. 中位数

中位数（median）是对变量中心位置的另一种度量方法。将所有数据按升序（从小到大的顺序）排列时，位于中间位置的数值即为中位数。当观测值为奇数时，中位数就是位于中间的那个数值；当观测值为偶数时，则没有单一的中间数值，在这种情况下，定义中位数为中间两个观测值的平均数。

【例3.7】计算5个大学班级样本的班级人数的中位数。将这5个数值按升序排列如下：

32　　42　　42　　46　　54

由于*n*=5是奇数，则中位数就是中间的数据值，因此班级人数的中位数是42。这里的数据中有两个值为42，按升序排列时，对每个观测值应单独处理。

如果计算的样本班级增加为6个，那么按升序排列为：

32　　42　　42　　46　　54　　60

则找出中间的两个值42和46，中位数就是这两个数的平均数，即（42+46）/2=44。

对于对称分布，平均数和中位数是相等的。当数据为右偏时，平均数一般比中位数要大；当数据为左偏时平均数一般比中位数要小。

中位数的计算简单，对极端值不敏感，因为除了中间值，中位数并未利用其他观测值，但这样它就没有利用数据中的所有信息。

3. 众数

众数（mode）是数据中出现频率最高的数值。例如，上面5个大学班级人数组成的样本。出现次数超过一次的数值只有42，这个数据出现的频数为2，是出现频率最高的数值，所以它就是众数。

有时候所有的数值出现的次数都相等，这种情况下不存在众数。有时候出现频数最大的数值可能是两个或两个以上，在这种情况下存在多个众数。

众数是描述定性数据的重要指标。例如，对表3-1中5种品牌笔记本电脑的频数分布数据，众数（品牌电脑购买的最大频数）是联想。对于这种类型的数据，平均数和中位数显然是毫无意义的。众数能够提供给大家最感兴趣的数据，即品牌笔记本电脑的最大购买频数。

4. 四分位数

将所有数值按大小顺序排列并分成四等份，处于三个分割点位置的数值就是四分位数。最小的四分位数称为下四分位数，用Q1表示；中点位置的四分位数就是中位数，用Q2表示；最大的四分位数称为上四分位数，用Q3表示。此外，十分位数、百分位数也是常用的分位数。

计算四分位数时，首先需要确定四分位数的位置，然后根据位置计算相应的数值。假设数据个数为n，常用的方法认为四分位数的位置为：$(n+1)/4$，$2(n+1)/4$和$3(n+1)/4$[1]。如果四分位数的位置不是整数，则四分位数等于前后两个数的加权平均。

【例3.8】12名执业药师的收入如下：2710，2755，2850，2880，2880，2890，2920，2940，2950，3050，3130，3325。计算执业药师月收入数据的四分位数。

将收入数据按照升序排列。根据前面的公式，Q1为第（12+1）/4=3.25个数，Q2为第6.5个数，Q3为第9.75个数。

因此Q1=2850+(2880-2850)*0.25=2857.5； Q2=2890+(2920-2890)*0.5=2905； Q3=2950+(3050-2950)*0.75=3025。

5. 时间序列的描述统计：平均发展水平和平均发展速度

对时间序列进行描述统计分析时有一些特殊问题需要加以注意：计算平均发展水平时需要区分时间序列的类型；计算平均发展速度时一般使用几何平均的方法，下面我们分别加以介绍。

（1）平均发展水平

时间序列中每一个观测值称为发展水平。在时间序列分析中，通常把要研究的那个时间的发展水平称为报告期水平或计算期水平，把用来作为比较基础的时间的发展水平称为基期水平。基期和报告期不是固定不变的，可以随着研究的目的不同和研究时间的变更作相应的改变。

将不同时期的发展水平加以平均得到的平均数称为平均发展水平。平均发展水平既可以根据绝对数时间序列来计算，也可以根据相对数或平均数时间序列来计算。

①.根据绝对数时间序列计算平均发展水平。

在绝对数时间序列中，由于时期序列和时点序列的性质不同，两种序列序时平均数的计算方法也有较大差别。

由绝对数时期序列计算平均发展水平。由于绝对数时期序列中各个指标数值具有可加性，其数值大小与时期长短有着密切的关系，因此我们只需采用简单算术平均法，以时期项数除时期序列中各个数值之和即可求得序时平均数。计算公式为：

[1]四分位数所处位置的计算方法并不一致，Excel 中四分位数的位置分别为(n+3)/4，2(n+1)/4，（3 n+1)/4。按不同的公式计算结果可能会有差异，但数据量大时一般差别不大。

$$\bar{a}=\frac{a_1+a_2+.....+a_n}{n}=\frac{\sum_{i=1}^{n}a_i}{n} \tag{3-6}$$

式中，$\bar{a}$：平均发展水平；a_i：各时期发展水平（i=1,……,n）；n：发展水平的个数。

由绝对数时点序列计算平均发展水平。分为连续时点和间断时点两种情况。严格地说，时点序列中两个时点指标之间都是有间隔的。从这个意义上来说，时点序列都属于不连续的间断序列。但是，由于历史的原因，人们习惯上把逐日记录、逐日排列的时点序列称作连续时点序列。连续时点序列可按时期序列公式计算平均发展水平。间断时点序列可以先计算出两个点之间的平均数，再用相隔的时期长度加权计算总的平均数。计算公式为：

$$\bar{a}=\frac{\frac{a_1+a_2}{2}\times f_1+\frac{a_2+a_3}{2}\times f_2+.....\frac{a_{n-1}+a_n}{2}\times f_{n-1}}{f_1+f_2+.....f_{n-1}} \tag{3-7}$$

其中，f_i为时间间隔长度。

如果各时点之间的间隔相等，式（3-7）可简化为：

$$\bar{a}=\frac{\frac{a_1}{2}+a_2+a_3+.....+\frac{a_n}{2}}{n-1} \tag{3-8}$$

②. 根据相对数或者平均数时间序列计算平均发展水平时，其基本方法是先按构成相对数或者平均数的两个数值（分子和分母）的性质分别求出它们的平均发展水平，再将两者相除即可得到总的平均发展水平。用公式表示为：

$$\bar{c}=\frac{\bar{a}}{\bar{b}} \tag{3-9}$$

【例 3.9】中国 1991 年-2000 年的 GDP、年末总人口数见表 3-8。根据数据计算：①. 1991-2000 年的年均 GDP；②. 1991-2000 年的年均人口数，已知 1990 年年末人口数为 114333 万人；③. 计算 1991-2000 年的年人均 GDP。

表 3-8 中国 1991 年-2000 年的 GDP 和年末总人口数

年份	GDP，亿元 2000 年价格	年末人口数 万人
1991	37296.99	115823
1992	42555.87	117171
1993	48130.69	118517
1994	54195.15	119850
1995	59072.72	121121
1996	64861.84	122389
1997	70439.96	123626
1998	75944.61	124761
1999	81390.56	125786
2000	88228.10	126743

①由于 GDP 是时期序列，因此利用式（3-6），中国 1991-2000 年的年均 GDP 为：

$$\bar{a}=\frac{\sum_{i=1}^{n}a_i}{n}=62211.65\text{（亿元）}$$

②由于人口数据是时点序列且各时点的时间间隔相等，因此，根据式（3-8）可得 1991-2000 年的年均人口数为：

$$\bar{Y}=\frac{\frac{114333}{2}+115823+\Lambda+\frac{126743}{2}}{11-1}=120958.2\text{（万人）}$$

③根据式（3-9），1991-2000 年的年人均 GDP 为：

$$\bar{c}=\frac{\bar{a}}{\bar{b}}=\frac{62211.65}{120958.2}=0.5143\text{万元}$$

（2）平均发展速度

时间序列中的两个发展水平相比的结果称为发展速度，一般用百分数表示。报告期水平与某一固定基期水平相比得到的结果称为定基发展速度，用报告期水平与前期水平相比得到的结果称为环比发展速度；发展速度减去 100%得到的数值称为增长速度。根据发展速度可以计算出平均发展速度；平均发展速度减去 100%则可以得到平均增长速度。

平均发展速度一般采用几何平均的方法计算[2]，等于连续 n 个环比发展速度的乘积开 n 次方，见公式 3-10。几何平均法的实质是，现象从最初水平出发，每期都按平均发展速度发展变化，到报告期所达到的水平正好等于实际水平。

$$\bar{g}=\sqrt[n]{g_1\cdot g_2\cdot g_3\cdots\cdots g_n}=\sqrt[n]{\prod g}=\sqrt[n]{\frac{a_n}{a_0}}\qquad\text{（3-10）}$$

式中 n 为环比发展速度的个数，g_i 为各个环比发展速度，a_0 和 a_n 为基期和报告期的水平。

【例 3.10】计算 1991-2006 年中国 GDP 的定基和环比发展速度、增长速度以及平均发展速度和平均增长速度。（数据见表 3-9）。

表 3-9　1991-2006 年中国的 GDP 和平均发展速度、均增长速度

年份	GDP（90 年价格，亿元）	发展速度（%）		增长速度（%）	
		环比	定基	环比	定基
1990	18668	—	100	—	—
1991	20384	109.19	109.19	9.19	9.19
1992	23287	114.24	124.74	14.24	24.74
1993	26534	113.94	142.14	13.94	42.14
1994	30007	113.09	160.74	13.09	60.74
1995	33287	110.93	178.31	10.93	78.31
1996	36620	110.01	196.17	10.01	96.17
1997	40020	109.28	214.38	9.28	114.38
1998	43154	107.83	231.17	7.83	131.17
1999	46448	107.63	248.81	7.63	148.81

[2]在我国的统计学教科书中还经常介绍另外一种计算方法：高次方程法（或累计法）。按这种方法，平均发展速度应满足以下方程：$a_1+a_2+\cdots\cdots a_n=a_0\bar{g}+a_0\bar{g}^2+\cdots\cdots a_0\bar{g}^n$。也就是说，现象从最初水平出发，每期都按平均发展速度发展变化，计算所得的各期水平之和等于实际各期水平之和。解这个高次方程即可求出平均发展速度。虽然这种计算方法借助计算机并不难实现，但在实际中应用不多。

2000	50358	108.42	269.76	8.42	169.76
2001	54539	108.30	292.15	8.30	192.15
2002	59496	109.09	318.71	9.09	218.71
2003	65461	110.02	350.66	10.02	250.66
2004	72061	110.08	386.01	10.08	286.01
2005	79576	110.43	426.27	10.43	326.27
2006	88847	111.65	475.93	11.65	375.93

定基和环比发展速度、增长速度的计算结果见表 3-9。

GDP 平均发展速度为：$\bar{g}=\sqrt[16]{\frac{88847}{18668}}=110.24\%$；GDP 平均增长速度为：110.24%-100%=10.24%。

二、离散程度

除了度量数据的集中趋势外，我们往往还需要考虑数据的变异程度即离散程度。因为对于不同的数据来说，即使集中趋势的衡量指标相等，其具体的数据分布也会有差异。

1. 全距

全距（range）是一种最简单的变异程度的衡量指标。全距=最大值-最小值。例如，执业药师月收入数据的最大值为 3325，最小值为 2710，因此全距为 3325-2710=615。

虽然全距是最容易计算的变异程度的衡量指标，但却很少单独使用。其原因是，全距仅仅是基于最大最小两个观测值计算的，极易受到异常值的影响。例如，执业药师月收入数据中的最大值如果是 10000，在这种情况下，全距将为 10000-2710=7290，而不是 615。如此之大的全距将不能准确地描述执业药师月收入数据的变异程度，因为在 12 个数据中有 11 个数据都集中在 2710 到 3130 之间。

2. 四分位距

四分位距（Interquantile Range，IQR）是第三四分位数 Q_3 与第一四分位数 Q_1 的差值，IQR= Q_3 - Q_1，即四分位数间距是在中间的 50%的数据的全距。四分位数间距在一定程度上克服了异常值的影响。

对于执业药师月收入数据，四分位数 Q_3 =3025， Q_1 =2857.5。因此，四分位数间距等于 3025-2857.5=167.5。

3. 方差和标准差

（1）方差

方差（variance）的计算依赖于每个观测值 x_i 与均值之间的离差。如果数据来自总体，则离差平方的均值称为总体方差（population variance），总体方差用字母 σ^2 表示。对于具有 N 个观测值的总体，其总体方差可以定义为：

$$\sigma^2=\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} \qquad (3\text{-}11)$$

在实际应用中，我们往往需要分析样本数据，用样本数据方差来估计总体方差。用 s^2 表示样本方差，则其定义为：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad (3\text{-}12)$$

在计算样本方差时分母上用 n-1 是因为这一统计量是总体方差的无偏估计。

【例 3.11】根据执业药师月收入数据计算样本方差。

计算过程见表 3-10。

表 3-10　执业药师月收入数据的样本方差计算表

起始月薪 (x_i)	样本均值 $(\bar{x})$	离差 $(x_i-\bar{x})$	离差的平方 $(x_i-\bar{x})^2$
2850	2940	-90	8100
2950	2940	10	100
3050	2940	110	12100
2880	2940	-60	3600
2755	2940	-185	34225
2710	2940	230	52900
2890	2940	-50	2500
3130	2940	190	36100
2940	2940	0	0
3325	2940	385	148225
2920	2940	-20	400
2880	2940	-60	3600

因此，

$$\sum(x_i-\bar{x})^2=301850,\quad s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{301850}{11}=27440.91$$

（2）标准差

标准差（Standard Deviation）是方差的平方根，我们用 s 表示样本标准差，σ 表示总体标准差。则：

$$\text{样本标准差}\ s=\sqrt{s^2} \tag{3-13}$$

$$\text{总体标准差}\ \sigma=\sqrt{\sigma^2} \tag{3-14}$$

例如，已知执业药师月收入的样本方差 $s^2=27440.91$，则样本标准差 $s=\sqrt{s^2}=\sqrt{27440.91}=165.65$。

将方差转换为标准差的优势在于：方差的单位都是平方项，例如，执业药师月收入方差的单位为元 ^{2}，而标准差的单位为元。由于标准差和原始数据有相同的度量单位，因此更容易和平均数等其他统计量进行比较。

4. 离散系数

离散系数（Coefficient of Variation）主要用来度量数据相对于均值的离散程度。如果两个变量的计量单位不同，则只能通过离散系数来比较其离散程度。在计量单位相同的情况下，如果两组数据的均值相差悬殊，离散系数也可能比标准差等绝对指标更有意义。

离散系数是标准差与其相应的均值之比，表示为百分数。计算公式为：

$$\text{总体的离散系数}\qquad CV=\frac{\sigma}{\bar{X}}\times100\% \tag{3-15}$$

样本的离散系数　　$cv=\frac{s}{\bar{x}}\times 100\%$　　（3-16）

对于执业药师月收入数据，样本平均数为 2940，样本标准差为 165.65，因此离散系数为 $(165.65/2940)\times 100\%=5.6\%$，它说明样本标准差仅为样本平均数的 5.6%。

三、数据分布的形状的描述

前面介绍的几种对数据集中趋势和变异程度的度量方法，分布形态的度量也很重要。我们已经学过的直方图对分布形态提供了一种很好的图形描述。分布形态的数值度量的常用方法是偏度与峰度。

1. 偏态及其测定

偏度（Skewness）是对数据分布的对称性的衡量指标。对于单位个数为 N，均值为 μ，标准差为 σ 的总体，其偏度的计算公式为：

$$偏度=\frac{\sum_{i=1}^{N}(X_i-\mu)^3}{N\sigma^3}　　（3\text{-}17）$$

显然，如果数据分布是对称的，偏度的值会等于 0。如果这个指标不等于 0，则说明数据分布是不对称的。经验表明，如果偏度小于零，则数据一般呈左偏分布，在直方图的左侧有一条长尾巴；如果偏度大于零，则数据一般呈右偏分布，在直方图的右侧有一条长尾巴。对于左偏的数据，偏度是负数；对于右偏的数据，偏度是正值；如果数据是对称的，则偏度为零。

根据样本数据计算偏度时，考虑到统计量的数学性质，计算公式要复杂一些，这里我们不做介绍。但计算结果的含义是相同的。

2. 峰度及峰度系数

峰度（kurtosis）描述了分布中的尖峰程度。对于单位个数为 N，均值为 μ，标准差为 σ 的总体，其峰度的计算公式为：

$$峰度=\frac{\sum_{i=1}^{N}(X_i-\mu)^4}{N\sigma^4}-3　　（3\text{-}18）$$

如果数据呈正态分布，则峰度系数为 0；当数据的峰度系数大于 0 时，典型的情况下数据分布的中心会比均值为 μ、标准差为 σ 的正态分布高一些，而尾部则厚一些，称为尖峰分布；当数据的峰度系数小于 0 时，典型的情况下数据分布的中心会比均值为 μ、标准差为 σ 的正态分布矮一些，而尾部则窄一些，称为平峰分布。

根据样本数据计算峰度时，考虑到统计量的数学性质，计算公式要复杂得多，这里我们不做介绍。

3. 箱线图

箱线图（Boxplot）是基于数据位置统计量的一种图形描述方法。箱线图在不同软件中的绘制方法有一些细小的差别，这里介绍一下 SPSS Statistics 中的绘制方法。

（1）先根据三个四分位数 Q1、Q2、Q3 画出中间的盒子；

（2）由 Q3 至 Q3+1.5*IQR 区间内的最大值向盒子的顶端连线，由 Q1 至 Q1-1.5*IQR 区间内的最小值向盒子的底部连线；

（3）处于 Q3+1.5*IQR 至 Q3+3*IQR 或者 Q1-1.5*IQR 至 Q1-3*IQR 范围内的数据用圆点标出，称为离群值（outliers）；

（4）大于 Q3+3*IQR 或者小于 Q1-3*IQR 的用星号标出，称为极端值（Extremes）。

图 3-9 是执业药师月收入数据的箱线图。图中标注的数字 10 表示这个离群值对应的个体

的序号为 10。从图中可以看出数据是右偏的。

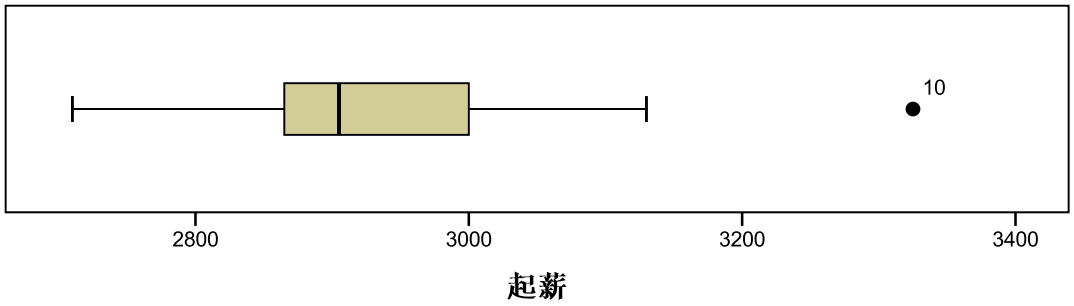


图 3-8　执业药师月收入数据的箱线图

4. *z*-分数

对于一个数据集，除了位置、变异程度和形态的度量外，有时还需要对数据项在数据集中的相对位置进行考察。这样的指标能帮助我们确定一个特殊的数据项是靠近数据集的中心，还是远离数据集。

我们可以根据平均数和标准差确定观测值的相对位置。假设一个样本包含 n 个观测值，其观测值用 x_i 表示，样本平均数为 $\bar{x}$，样本标准差为 s，则 x_i 的 *z*-分数计算公式为：

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (3\text{-}19)$$

z-分数（z-score）通常被称为标准得分（standard score），它可以解释为 x_i 距离平均数 $\bar{x}$ 的标准差的个数。例如，z_1 =1.2，表示 x_1 比样本平均数大 1.2 个标准差。当观测值大于平均数时，*z*-分数将大于零；当观测值小于平均数时，*z*-分数将小于零；*z*-分数等于零，则表示观测值等于平均数。

两个不同数据集的观测值如果具有相同的 *z*-分数，则可以说它们具有相同的相对位置，因为它们与平均数的距离都有相同个数的标准差。

表 3-11 给出了大学班级人数数据的 *z*-分数。已知样本平均数 $\bar{x} = 44$，样本标准差 $s = 8$。第 5 个观测值的 *z*-分数为-1.50，说明它距离平均数最远，且比平均数小 1.5 个标准差。

表 3-11　大学班级人数数据的 z-分数计算表

班级人数(x_i)	平均数的离差($x_i - \bar{x}$)	*z*-分数 $\frac{x_i - \bar{x}}{s}$
42	-2	-2/8=-0.25
54	10	10/8=1.25
42	-2	-2/8=-0.25
46	2	2/8=0.25
32	-12	-12/8=-1.50

第三节　用 SPSS 进行描述统计分析

一、用 SPSS 绘制统计图表

SPSS具有很强的制图功能，可以绘制多种统计图形。这些图形可以由各种统计分析过程产生，也可以直接由“图形”菜单产生。这里我们主要介绍由“图形”菜单中的“图表构建程序”（Chart Builder）生成图形的方法。

1. SPSS“图表构建程序”

利用“图表构建程序”可以通过简单的拖拉操作，利用预定义的图形“库”（Gallery Charts）或者“基本元素”（Basic Elements）绘制各种复杂的图形。其具体步骤如下：

第一步，在SPSS的“图形”菜单下选择“图表构建程序”选项，随即出现“图表构建程序”的界面。如图3-9所示，默认的是使用图形“库”。

图 3-9　“图表构建程序”第一步：“图表构建程序”界面

第二步：在图3-9的下半部分，选择图形种类。SPSS提供了9种图形可供选择。每一种都有若干类型。例如，要绘制条形图，选择“条”（bar），双击第一个图形或将第一个图形拖到图3-9的画布（canvas）区域，如图3-10所示。

图 3-10　“图表构建程序”第二步：选择图形种类

第三步：从图3-10中的“变量”列表中将所需要的变量拖到左边面板的坐标轴区域。这样一个基本的图形就做好了。注意，变量类型在这一步中非常重要，因为在“图表构建程序”中对于不同的图形，坐标轴区域适用的变量类型不同。可以对变量用右键选择相应的类型来暂时修改，以满足做图要求。

如果需要，还可以对图形进行修改，例如在“元素属性”对话框中设置图形元素的属性等。

下面我们利用一组饮料数据为例说明绘制条形图的过程。在SPSS中打开数据文件“饮料.sav”，在SPSS的“图形”菜单下选择“图表构建程序”选项，在“图表构建程序”的“库”选项卡中，选中其中的“条”形图；将图例中的简单条形图拖到画布区域；将定性变量“购买品牌”拖到*x*轴。其他选项采用默认设置，单击“确定”按钮即可得到图3-11所示的条形图。

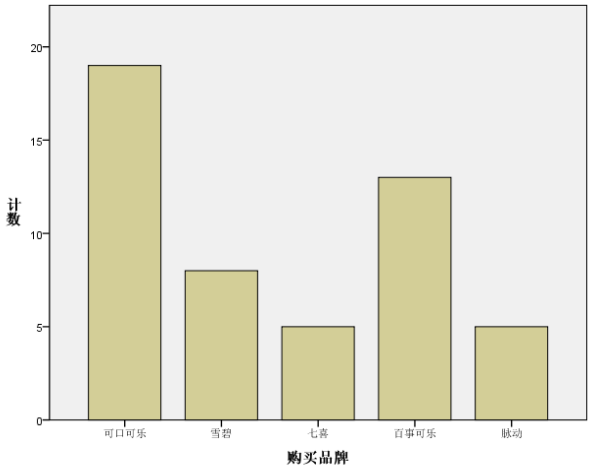


图 3-11　饮料购买次数的条形图

类似地，可以对该数据绘制饼图，只需选择“饼图/极坐标图”，其他选项采用默认设置，单击“确定”按钮即可。可以进行相应的修改在饼图上显示数值标签：双击图形进入编辑状态，在“元素”菜单中选择“显示数据标签”，并在弹出的属性对话框中设定显示“购买品牌”，即可以在饼图中显示出相应的标签。相关操作和图形如图3-12所示。

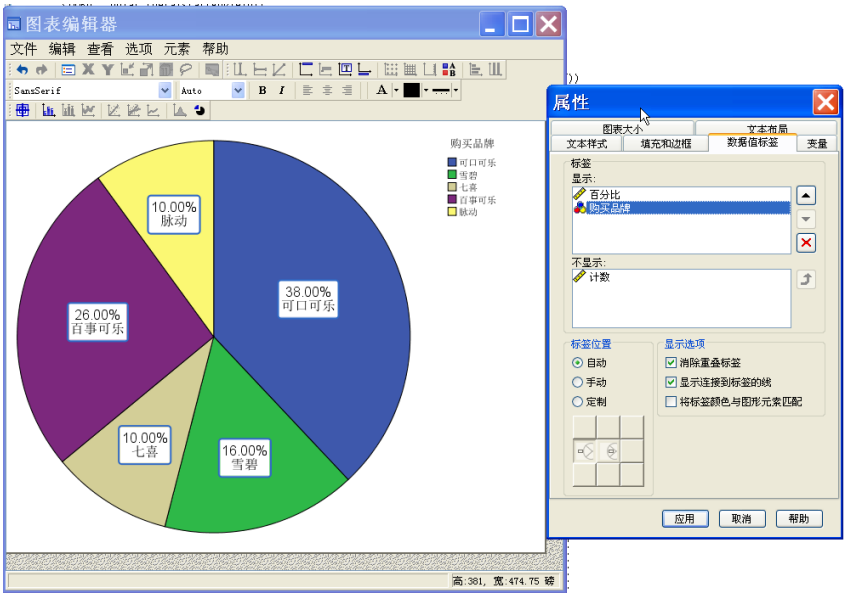


图 3-12　饮料购买比例的饼图及其编辑

其他图形如直方图、箱线图的绘制方法类似。可以分别使用“散点图”、“直方图”、“箱图”等选项进行绘图。

2. 用SPSS绘制直方图和茎叶图

在“分析”菜单下选择“描述统计”中的“探索”（Explore）过程可以完成直方图和茎叶图的绘制。

我们用“学生调查.sav”数据集加以说明。这一数据集中包含了对35名学生的调查结果，其中包括学生的性别、年龄、学习兴趣、概率成绩、统计成绩、身高、体重等变量。我们对学生的概率成绩做直方图和茎叶图。在“探索”对话框中，单击“绘制”按钮，选中“茎叶图”、“直方图”复选框，在输出结果中就会看到直方图（图3-13）和茎叶图（图3-14）。

绘制直方图时，SPSS中默认是自动设置各组界限的，这可能和我们的要求不一致。例如，如果我们希望以10分为组距对考试成绩进行分组，可以进行相应的修改：双击直方图进入编辑状态，再双击直方图的条形区域，会弹出相应的属性对话框。在属性对话框中对x轴进行设置，即可得到需要的结果（参见图3-13）。

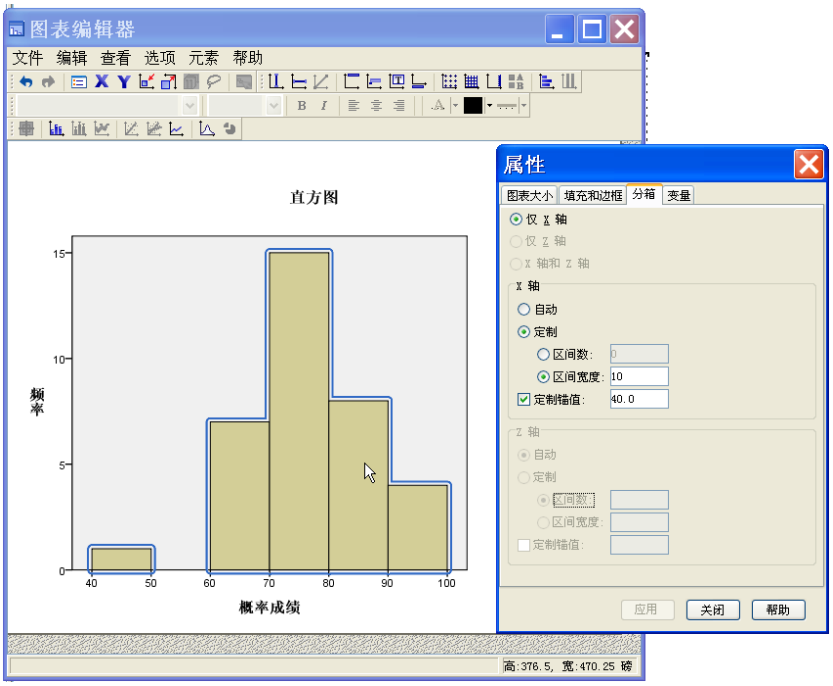


图3-13 SPSS绘制的直方图及其编辑

Frequency	Stem &	Leaf
1.00	Extremes	(=<49)
3.00	6 .	234
4.00	6 .	6788
6.00	7 .	112344
9.00	7 .	556788889
2.00	8 .	02
6.00	8 .	567778
4.00	9 .	0012

Stem width: 10
Each leaf: 1 case(s)

图 3-14　概率成绩的茎叶图

二、用 SPSS 计算描述统计量

1. “分析”菜单下的“描述统计”

SPSS中进行描述性统计分析的模块主要集中在“分析”菜单下的“描述统计”中。包括“频率”、“描述”、“探索”等模块。这些模块的功能既有自己的特色，又有所交叉，可以根据个人的习惯和数据分析的需要选择使用。

“频率”模块可以生成频数表，计算均值、方差、分位数、偏度、峰度等常用的描述统计指标，绘制饼图、条形图、直方图等。众数只有在这一模块才能直接计算。

“描述”模块可以计算均值、方差、偏度、峰度等指标，该模块还有个特殊功能就是将原始数据进行标准化以变量的形式存入数据集供以后的分析使用。

“探索”模块可以计算均值、方差、分位数、偏度、峰度等常用的描述统计指标，绘制茎叶图、直方图、箱线图，计算均值的置信区间，进行正态性检验等。“探索”模块的一个特色是可以先将目标变量按照某一个指标进行分组，然后对不同组别的数据分别进行描述统计分析。这一功能在实际中比较有用。

以上这些模块的使用方法比较类似，我们以“频率”模块为例加以说明。

如果我们需要计算“学生调查.sav”数据集中学生概率成绩的描述统计指标，相关操作如下：在SPSS中打开数据集；选择“分析”→“描述统计”→“频率”，进入“频率”对话框。将“概率成绩”选入“变量”框。在这里“频率表格”（频数分布表）没有太大意义，我们去掉相应的复选框；单击“统计量”按钮进入“统计量”对话框，选中需要的统计指标，单击“继续”按钮返回，单击“确定”按钮可得到相应的计算结果。相应的软件操作参见图3-15，结果见表3-12。

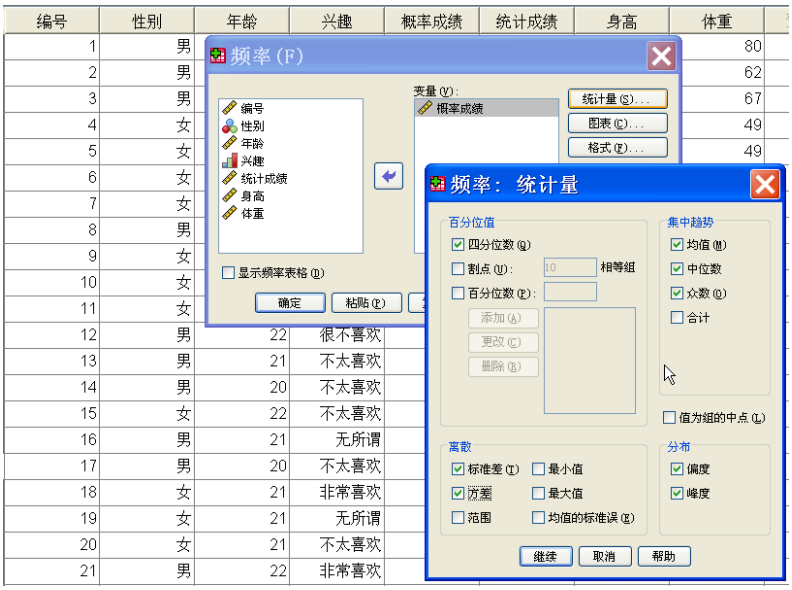


图 3-15 使用“频率”模块进行描述统计的相关操作

表 3-12 SPSS 对概率成绩的描述统计结果

N	有效	35
	缺失	0
均值		76.60
中值		77.00
众数		78
标准差		9.790
方差		95.835
偏度		-.485
偏度的标准误		.398
峰度		.344
峰度的标准误		.778
百分位数	25	71.00
	50	77.00
	75	86.00

2. 用SPSS进行分组汇总

在对数据进行描述统计分析时，经常需要统计定性变量的频数分布，根据多个定性变量编制统计表，对变量进行分组计算以分析不同组别之间的差异等等。“描述统计”中的“交叉表”模块可以绘制列联表；“分析”→“比较均值”→“均值”模块可以进行分组计算；“分析”→“表”模块可以完成很复杂的统计表和分组计算。这里我们只介绍“交叉表”和“均值”模块的使用方法。

（1）列联表的绘制。假设我们需要汇总“学生调查.sav”数据集中分性别和学习兴趣的人数。选择“分析”→“描述统计”→“交叉表”，把“性别”和“兴趣”变量分别设定为行变量和列变量，单击“确定”即可得到相应的列联表。相应的操作和结果参见图3-16

和表3-13。

图3-16 SPSS绘制列联表的相关操作

表 3-13　根据性别和兴趣绘制的列联表

性别* 兴趣 交叉制表

		兴趣					
		很不喜欢	不太喜欢	无所谓	比较喜欢	非常喜欢	合计
性别	男	2	5	4	1	4	16
	女	2	4	3	7	3	19
合计		4	9	7	8	7	35

（2）用SPSS进行分组计算。假设我们希望按学习兴趣分组，计算不同组别概率成绩的均值和标准差。选择“分析”→“比较均值”→“均值”，把变量“概率成绩”选入“因变量列表”框中，把变量“兴趣”选入“自变量列表”框中，单击“选项”选择需要计算的统计指标，即可得到需要的计算结果。相应的操作和结果参见图3-17和表3-14。

图 3-17 使用“均值”模块进行分组计算的相关操作

表 3-14　概率成绩的描述统计指标（按兴趣程度分组）

兴趣	均值	N	标准差
很不喜欢	66.00	4	12.728
不太喜欢	81.33	9	9.772
无所谓	73.00	7	9.092
比较喜欢	75.88	8	7.240
非常喜欢	81.00	7	6.633
总计	76.60	35	9.790

小　结

第一节介绍了如何用统计图和统计表组织和汇总数据的方法，使我们能够更容易地解释数据，理解其中的内涵。频数分布、条形图和饼图是汇总定性数据的方法。频数分布、直方图、茎叶图、累积频数分布是汇总定量数据的方法。列联表是汇总两个变量数据的方法。

第二节数据描述的数值方法介绍了概括数据分布的位置、变异程度和形态的描述统计指标。作为中心位置的描述指标，我们定义了均数、中位数和众数。接着我们介绍了全距、四分位距、方差、标准差和标准误差等度量变异程度或离散程度的指标。数据分布形态的基本度量是偏度、峰度。我们还介绍了箱线图，它对数据分布的位置、变异程度和形态提供了类似信息。

第三节介绍了使用SPSS进行绘制统计图和计算描述统计量的方法。

思考与练习

1. 某研究提供了目前最畅销汽车品牌的数据。这里列出的有雪铁龙、福特、大众、本田和丰田。表3-15列出了50次汽车购买的样本数据（数据文件：汽车品牌.sav）。

表 3-15　50 次汽车购买的数据

雪铁龙	大众	本田	福特	雪铁龙
福特	本田	雪铁龙	丰田	大众
雪铁龙	福特	丰田	雪铁龙	福特
大众	福特	大众	福特	丰田
大众	福特	福特	本田	丰田

雪铁龙	雪铁龙	雪铁龙	雪铁龙	大众
丰田	雪铁龙	本田	福特	福特
大众	福特	本田	丰田	本田
福特	雪铁龙	福特	福特	大众
丰田	雪铁龙	福特	福特	福特

（1）构建频数和频率分布。

（2）哪两款汽车最畅销？

（3）绘制条形图和饼图，说明汽车销售的品牌分布。

2. 对例3.8中执业药师的收入数据（数据文件：执业药师.sav），

（1）绘制收入的直方图和箱线图；

（2）根据（1）的结果分析数据分布的特征；

（3）计算均值、中位数和众数；

（4）根据数据分布的特征，哪个指标能更好地反映数据的一般水平？

3. 某地区平均每月外出就餐费用65.88元。一个样本提供了过去几个月该地区人们外出就餐费用的数据如表3-16（单位：元）（数据文件：就餐费用.sav）。

表3-16 外出就餐的费用

253	113	104	169	129
80	178	134	131	118
11	152	467	95	225
55	245	198	0	124
101	69	161	0	151

（1）计算平均数、中位数和众数；

（2）考虑（1）的结果，这些人外出就餐费用是否与当地人的平均费用相等；

（3）计算第一和第三四分位数；

（4）计算全距和四分位数间距；

（5）计算方差和标准差；

（6）计算数据的偏度，并评价数据分布的对称性。

（7）绘制数据的直方图，分析数据分布的特征。

4. 我国2001-2008年的GDP增长率如表3-17。根据数据计算2001-2008年我国的年均GDP增长率。

表3-17　2001-2008年我国的GDP增长率

2001	2002	2003	2004	2005	2006	2007	2008
8.3	9.1	10.0	10.1	10.4	11.6	13.0	9.0

5. 在SPSS软件中重复第三节中的例子，掌握软件的相关操作。

第四章　参数估计与假设检验

掌握参数估计和假设检验的基本思想是正确理解和应用其他统计推断方法的基础，后面将要学习的方差分析、非参数检验、回归分析、时间序列等统计推断方法都是在此基础上展开的。需要特别指出的是，所有的统计推断都要以随机样本为基础。如果样本是非随机的，统计推断方法就不适用了。

由于相关知识在先修课程中已经学习过，本章主要在回顾相关知识的基础上，补充讲解必要样本容量的计算、p 值、参数估计和假设检验方法的软件操作和结果分析等内容。

本章的主要内容包括：

（1）参数估计的基本思想和软件实现。

（2）简单随机抽样情况下样本容量的计算。

（3）假设检验的基本原理。

（4）假设检验中的 p 值。

（5）几种常用假设检验的软件实现。

第一节　参数估计

一、参数估计的基本概念

参数估计是指利用样本信息对总体数字特征作出的估计。例如，我们可以通过估计一部分产品的合格率对整批产品的合格率作出估计，通过调查一个样本的人口数来对全国的人口数作出估计，等等。

参数估计可以分为点估计和区间估计。

点估计是指根据样本数据给出的总体未知参数的一个估计值。对总体参数进行估计的方法可以有多种，例如矩估计法、极大似然估计法等，得到的估计量（样本统计量）并不是唯一的。例如我们可以使用样本均值对总体均值作出估计，也可以使用样本中位数对总体均值进行估计。因此，在参数估计中我们需要对估计量的好坏作出评价，这就涉及到估计量的评价准则问题。常用的估计量评价准则包括无偏性、有效性、一致性等。无偏性是指估计量的数学期望与总体参数的真实值相等；有效性的含义是，在两个无偏估计量中方差较小的估计量较为有效，方差越小越有效；一致性是指随着样本容量的增大，估计量的取值应该越来越接近总体参数。

样本的随机性决定了估计结果的随机性。由于每一个点估计值都来自于一个随机样本，所以总体参数真值刚好等于一个具体估计值的可能性极小。区间估计的方法则以概率论为基础，在点估计的基础上给出了一个置信区间，并给出了这一区间包含总体真值的概率，比点估计提供了更多的信息。置信区间是根据事先确定的置信度$1-\alpha$给出的总体参数的一个估计范围。由于所研究的问题不同，区间估计中使用的计算公式也有所不同。

在区间估计中，置信度为$100（1-\alpha）$%的含义是：在同样的方法得到的所有置信区间中，有$100(1-\alpha)$%的区间包含总体参数的真实值。也可以说，对于计算得到的一个具体区间，“这个区间包含总体真实值”这一结论有 $100(1-\alpha)$%的可能是正确的。但是说“总体参数有$100(1-\alpha)$%的概率落入某一区间”是不严格的，因为总体参数是非随机的。

二、抽样分布

1. 抽样分布的概念

抽样分布（Sampling Distribution）是指统计量的概率分布，是区间估计和假设检验的基本依据。从总体中抽取一个样本量为 n 的随机样本，我们可以计算出统计量的一个值；如果从总体中重复抽取样本量为 n 的样本，就可以得到统计量的多个值。统计量的抽样分布就是这一统计量所有可能值的概率分布。

注意抽样分布是统计量的分布而不是总体或样本的分布。在统计推断中总体的分布一般是未

知的，不可观测的（但常常被假设为正态分布）；样本数据的统计分布是可以直接观测的，最直观的方式是直方图，样本数据的统计分布可以用来对总体分布进行检验；抽样分布则一般利用概率统计的理论推导得出，在应用中也是不能直接观测的，其形状和参数可能完全不同于总体或样本数据的分布。

2. 样本均值的抽样分布

下面我们用一个假设的例子来说明抽样分布的含义。设一个总体含有 4 个个体，分别为 X_1=1、X_2=2、X_3=3、X_4=4。则显然有总体的均值 $\mu=\frac{\sum_{i=1}^{N}X_i}{N}=2.5$；方差 $\sigma^2=\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}=1.25$。总体的频数分布如图 4-1。

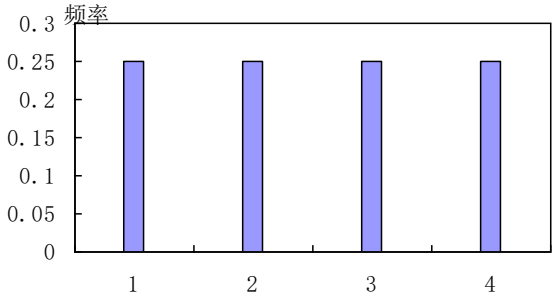


图 4-1　总体的频数分布

现从总体中抽取容量 n＝2 的简单随机样本，在重复抽样条件下，共有 4^2=16 个可能样本。所有的可能样本如表 4-1，各样本的均值如表 4-2，样本均值的频数分布（抽样分布）如图 4-2。从图 4-1 和图 4-2 我可以看出，在这个例子中样本均值的抽样分布和总体的分布是有很大差异的。

根据表 4-2 中的数据，16 个样本均值的算术平均数 $\mu_{\bar{x}}=\frac{\sum_{i=1}^{n}\bar{x}_i}{M}=\frac{1.0+1.5+\Lambda+4.0}{16}=2.5$；方差 $\sigma_{\bar{x}}^2=\frac{\sum_{i=1}^{n}(\bar{x}_i-\mu_{\bar{x}})^2}{M}=\frac{(1.0-2.5)^2+\Lambda+(4.0-2.5)^2}{16}=0.625$。容易验证，在这里样本均值抽样分布的均值等于总体的均值，方差等于总体方差的 1/n。

表 4-1　重复抽样条件下所有可能的 n =2 的样本（共 16 个）

第一个观察值	第二个观察值			
	1	2	3	4
1	1，1	1，2	1，3	1，4
2	2，1	2，2	2，3	2，4
3	3，1	3，2	3，3	3，4
4	4，1	4，2	4，3	4，4

表 4-2　16 个样本的样本均值

第一个观察值	第二个观察值			
	1	2	3	4
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5

4	2.5	3.0	3.5	4.0

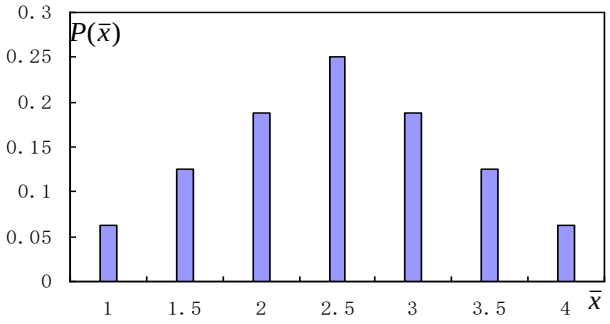


图 4-2　样本均值的抽样分布

一般的，关于样本均值的抽样分布有以下结论：当总体服从 $N(\mu,\sigma^2)$时，来自该总体的容量为 n 的样本的均值 $\bar{x}$ 也服从正态分布，$\bar{x}$ 的期望为 μ，方差为 σ^2/n。即 $\bar{x}\sim N(\mu,\sigma^2/n)$。

正态分布是一种对称的钟形分布，两个参数 μ 和 σ 共同决定了正态曲线的形状：参数 μ 决定了分布的中心位置，参数 σ 决定了曲线的陡峭程度（图 4-3）。

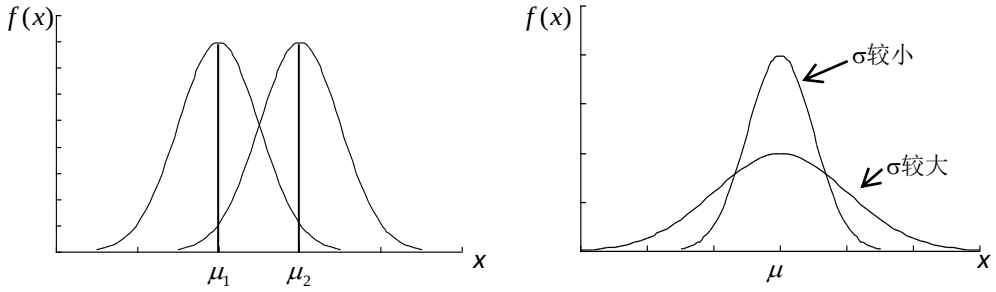


图 4-3　参数 μ 和 σ 对正态分布曲线形状的影响

当总体的分布不是正态分布时，样本均值抽样分布的形状如何呢？中心极限定理表明，从均值为 μ，方差为 σ^2 的一个任意总体中抽取容量为 n 的样本，当 n 充分大时，样本均值的抽样分布近似服从均值为 μ、方差为 σ^2/n 的正态分布。在应用中习惯上认为 $n\geqslant 30$ 时就可以认为是大样本了。

在统计学中，统计量的标准差被称为标准误（Standard Error）。简单随机抽样、重复抽样时，样本均值的标准误等于 $\sigma/\sqrt{n}$。这个指标在统计推断中经常用到，在对变量进行描述统计时统计软件一般会输出这一结果。

如果在简单随机抽样中采用的是不重复的方法，则样本均值抽样分布的方差略小于重复抽样的方差，等于 $\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}$。式中 $\frac{N-n}{N-1}$ 这一系数称为有限总体校正系数（Finite Population Correction Factor）。在实际应用中，当抽样比 $n/N<0.05$ 时可以忽略有限总体校正系数。

3. 统计推断中常用的概率分布

除了正态分布以外，在参数估计和假设检验中经常用到的概率分布还包括 t 分布、F 分布、χ^2 分布等等。

在根据样本比例对总体比例进行推断时也经常使用正态分布。用 p 表示样本比例，π 表示总体比例，则当 $np\geqslant 5$、$n(1-p)\geqslant 5$ 时，在简单随机抽样的情况下样本比例近似服从以下正态分

布：在重复抽样时 $p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$，在不重复抽样时 $p \sim N\left(\pi, \frac{\pi(1-\pi)}{n} \cdot \frac{N-n}{N-1}\right)$。

在根据样本均值推断总体均值时，在总体方差未知的情况下需要使用样本方差 s^2 来估计总体方差 σ^2，这时统计推断中需要使用 t 分布：随机变量 $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ 服从自由度为 n-1 的 t 分布。t 分布是类似于标准正态分布的一种对称分布，有一个称为自由度的参数，随着自由度的增大 t 分布会逐渐趋于标准正态分布（如图 4-4）。在大样本的情况下 t 分布可以使用标准正态分布代替。

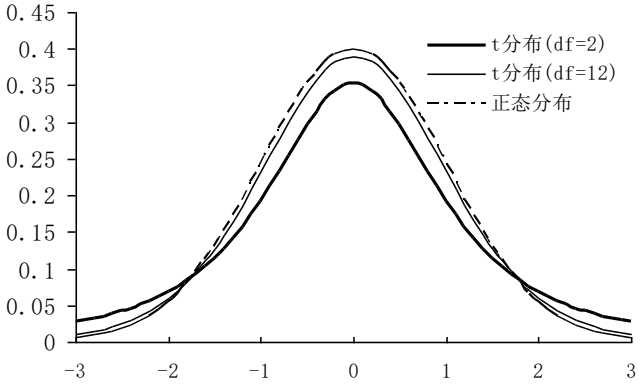


图 4-4　t 分布的形状

在对总体方差的推断和非参数检验中，相关检验统计量服从 χ^2 分布。χ^2 分布的自由度 n 决定了该分布的形状。该分布的变量值为正值，形状通常为不对称的右偏分布，但随着自由度的增大逐渐趋于对称（图 4-5）。

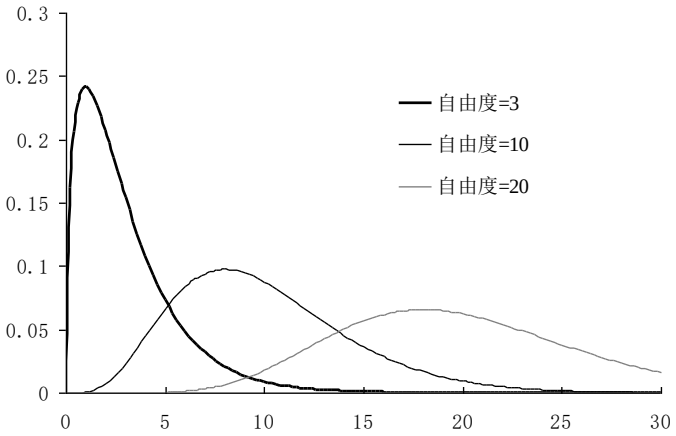


图 4-5　χ^2 分布的形状

在对两个总体方差的推断、方差分析和回归分析中，还要用到 F 分布。F 分布是由统计学家费希尔（R. A. Fisher）提出的，其形状由两个自由度参数 n_1 和 n_2 决定（图 4-6）。F 分布的变量值也大于 0，为右偏分布。

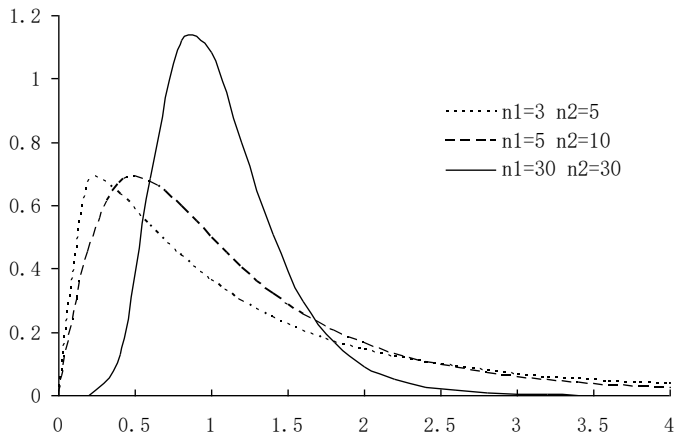


图 4-6　F 分布的形状

三、总体均值和比例的区间估计和软件实现

1. 单个总体均值的区间估计

如前所述，在总体正态、方差已知，或者总体分布未知，但为大样本的情况下，样本均值的抽样分布是正态分布或者可以用正态分布近似。设给定的置信度为 1-α，在重复抽样的情况下我们有：$P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \le Z_{\alpha/2}\right)=1-\alpha$，从而可以得出总体均值$\mu$的 100（1-$\alpha$）%的置信区间为：

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (4\text{-}1)$$

在实际应用中，总体方差一般是未知的，需要用样本方差来估计。在这种情况下需要按照 t 分布来计算置信区间，见式（4-2）。在手工计算的情况下，大样本时公式中的 t 值可以用 z 值来近似；使用统计软件进行计算时都是按照 t 值来计算的。

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \qquad (4\text{-}2)$$

在不重复抽样的情况下，我们有 $\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$，式（4-1）和式（4-2）需要做如下的修改：

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (4\text{-}3)$$

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (4\text{-}4)$$

2. 总体比例的区间估计

当 $np \geqslant 5$、$n(1-p) \geqslant 5$ 时，根据样本比例的抽样分布可知，重复抽样时总体比例的置信区间为式（4-5），不重复抽样时总体比例的置信区间为式（4-6）：

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad (4\text{-}5)$$

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} \qquad (4\text{-}6)$$

下面我们来看一个用 SPSS 计算置信区间的例子。

【例 4.1】儿童电视节目的赞助商希望了解儿童每周看电视的时间。表 4-3 是对 100 名儿童进行随机调查的结果（数据文件：看电视时间.sav）。（1）计算平均看电视时间 95%的置信区间；（2）计算每周看电视时间大于 30 小时的儿童比例的 95%的置信区间。

表 4-3　100 名儿童每周看电视的时间（小时）

39.7	19.5	34.7	27.0	41.3	15.1	20.5	31.3	18.3	17.0

21.5	29.9	15.0	16.4	36.8	23.4	24.1	28.9	23.4	24.4
40.6	46.4	23.6	39.4	35.5	19.5	29.3	31.2	20.6	34.9
15.5	31.6	38.9	38.7	27.2	26.5	14.7	15.6	28.4	24.0
43.9	20.6	29.1	9.5	21.0	42.4	13.9	32.8	29.8	32.9
33.0	38.0	28.7	20.6	19.7	38.6	37.1	17.0	15.1	23.4
21.0	21.8	29.3	21.3	22.8	23.4	32.5	11.3	43.8	30.8
15.8	23.2	20.3	33.5	30.0	37.8	24.4	26.9	29.0	27.7
27.1	22.0	36.1	23.0	22.1	26.5	22.9	26.9	30.2	25.2
23.8	35.3	21.6	35.7	30.8	22.7	24.5	21.9	26.5	50.3

解：（1）在 SPSS 软件中依次选择“分析”→“描述统计”→“探索”，对“时间”变量进行描述统计分析，可以得到表 4-4。从表中可以直接得到 95%的置信区间为[25.53，28.85]。

表4-4　SPSS输出的“时间”变量的部分描述统计结果

			统计量	标准误
时间	均值		27.191	.8373
	均值的 95% 置信区间	下限	25.530	
		上限	28.852	
	……		……	
	峰度		-.278	.478

（2）为了计算每周看电视时间大于 30 小时的儿童比例及其 95%的置信区间，我们先在 SPSS 中生成一个取值为 0 或 1 的变量，每周看电视时间大于 30 小时时取值为 1，否则取值为 0。具体操作为：选择“转换”→“计算变量”，在对话框中按照图 4-7 进行设置，即可生成名称为“大于 30”的新变量。对这个变量使用“探索”模块进行描述统计分析，结果如表 4-5。根据表 4-5，样本中每周看电视时间大于 30 小时的儿童比例为 33%。由于样本量 n=100，显然满足 $np \geq 5$、$n(1-p) \geq 5$ 的条件，因此可以按照正态分布来计算比例的置信区间。表 4-5 给出比例的 95%的置信区间为[0.2362, 0.4238]。

需要注意的是，在实际应用中，如果在计算置信区间时需要考虑有限总体校正系数，则需要根据软件的描述统计结果，按照式（4-3）、式（4-4）或者式（4-6）进行手工计算，软件不能直接提供最终的计算结果。

图 4-7　在 SPSS 中计算新变量

表4-5　SPSS输出的“大于30”变量的部分描述统计结果

			统计量	标准误
大于30	均值		.3300	.04726
	均值的 95% 置信区间	下限	.2362	
		上限	.4238	
	……		……	
	峰度		-1.491	.478

四、样本容量的计算

很显然，在抽样调查中样本容量越大抽样误差越小。由于调查成本方面的原因，在调查中我们总是希望抽取满足误差要求的最小的样本容量。下面我们根据置信区间的相关知识来看一下必要样本容量的计算问题。

1. 关于抽样误差的几个概念

（1）实际抽样误差

样本估计值与总体真实值之间的绝对离差称为实际抽样误差，用公式$|\hat{\theta}-\theta|$表示。由于在实践中总体参数的真实值是未知的，因此实际抽样误差是不可知的；由于样本估计值随样本而变化，因此实际抽样误差是一个随机变量。

（2）抽样平均误差

抽样平均误差反映了所有可能样本的估计值（$\hat{\theta}$）与相应总体参数（θ）的平均误差程度。计算公式为：

$$\sigma_{\hat{\theta}}=\sqrt{E(\hat{\theta}-\theta)^2} \qquad (4\text{-}7)$$

式(4-7)实际上就是统计量$\hat{\theta}$抽样分布的标准差，也就是一般所说的标准误(Standard Error)。例如在简单随机抽样时，重复抽样条件下有$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$，不重复抽样条件下有$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$。

我们通常所说的“抽样调查中可以对抽样误差进行控制”，就是指的抽样平均误差。影响抽样误差的因素包括：总体内部的差异程度，样本容量的大小，以及抽样的方式方法。

（3）最大允许误差（Allowable Error)

在参数估计中，我们一方面希望误差小一些，误差越小，估计的精确程度就越高；另一方面误差也不是越小越好，误差越小需要的样本量就越大，费用就越高。因此，需要确定一个允许的误差范围，这种在一定概率下抽样误差的允许范围，即为最大允许误差，也称为抽样极限误差。我们用E来表示最大允许误差，则有$|\bar{x}-\mu|\le E$，从而有$\bar{x}-E\le\mu\le\bar{x}+E$。可见，最大允许误差就是计算置信区间时样本均值（或样本比例）加减的量。置信区间和最大允许误差之间的关系是：置信区间=$\bar{x}\pm E$。最大允许误差是人为确定的，是调查可以容忍的误差水平。

2. 样本容量的计算

（1）必要样本容量的影响因素

①总体标准差。总体的变异程度越大，必要样本容量也就越大。

②最大允许误差。最大允许误差越大，需要的样本容量越小。

③置信度 1-α。要求的置信度越高，需要的样本容量越大。

④抽样方式　。其它条件相同，在重复抽样、不重复抽样；简单随机抽样与分层抽样等不同抽样方式下要求的必要样本容量也不同。

（2）简单随机抽样、重复抽样情况下，估计总体均值和估计总体比例时样本容量的确定方法

根据前面的分析，在简单随机抽样、重复抽样条件下，在估计总体均值时有 $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$，从而有

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{E^2} \qquad (4\text{-}8)$$

式中的总体方差可以通过以下方式估计：根据历史资料确定；通过试验性调查估计。

在估计估计总体比例时有 $E = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$，从而有：

$$n = \frac{Z_{\alpha/2}^2 \pi(1-\pi)}{E^2} \qquad (4\text{-}9)$$

式中的总体比例 π 可以通过以下方式估计：根据历史资料确定；通过试验性调查估计；取为0.5。这是因为 π =0.5 时 π (1- π)取得最大值。

（3）简单随机抽样、不重复抽样情况下，估计总体均值和估计总体比例时样本容量的确定方法

在不重复抽样时的必要样本容量要比重复抽样小一些。用 n_0 表示重复抽样时的样本容量，则不重复抽样时的必要样本容量如式（4-10）所示：

$$n = \frac{n_0}{1+\frac{n_0}{N}} \qquad (4\text{-}10)$$

下面我们来看几个例子。

【例 4.2】需要多大规模的样本才能在 90%的置信水平上保证均值的误差在±5 之内？前期研究表明总体标准差为 45。

解：$n = \frac{Z^2\sigma^2}{E^2} = \frac{1.645^2(45)^2}{5^2} = 219.2 \approx 220$ 。

注意在最后取整数时不是四舍五入，而是向上取整，这样才能保证满足对允许误差的要求。

【例 4.3】一家市场调研公司想估计某地区有电脑的家庭所占的比例。该公司希望对比例 π 的估计误差不超过 0.05，要求的可靠程度为 95%，应抽多大容量的样本（没有可利用的 π 估计值）？

解：已知 E=0.05，α=0.05，$Z_{\alpha/2}$=1.96，由于没有可利用的 π 的信息，我们将其取为 0.5。

$n = \frac{Z_{\alpha/2}^2\pi(1-\pi)}{E^2} = \frac{(1.96)^2(0.5)(1-0.5)}{(0.05)^2} = 384.16$，因此需要抽取 385 户居民进行调查。

【例 4.4】假设在【例 4.3】采用的是不重复抽样，并且已知该地区共有 1000 户居民，则应抽多大容量的样本？

解：$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{384.16}{1+\frac{384.16}{1000}} \approx 277.54$，因此需要抽取 278 户居民进行调查。

第二节　假设检验

一、假设检验的基本原理

假设检验是事先作出关于总体参数、分布形式、相互关系等的命题，然后通过样本信息来判断该命题是否成立的统计方法。

假设检验有着非常广泛的应用，可以说是最重要的推断统计方法。例如可以根据样本数据对以下各类问题进行检验：产品生产线工作是否正常？某种新生产方法是否会降低产品成本？治疗某疾病的新药是否比旧药疗效更高？厂商声称产品质量符合标准，是否可信？等等。

1. 假设检验的基本原理

假设检验的基本原理是：小概率事件在一次试验中几乎不会发生。如果对总体的某种假设是真实的（例如产品合格率≥95%），那么不利于或不能支持这一假设的事件 A（小概率事件，

例如抽样合格率=55%）在一次试验中几乎不可能发生的；要是在一次试验中 A 竟然发生了（抽样合格率=55%），就有理由怀疑该假设的真实性，拒绝提出的假设。
假设检验采用的是反证法的思想，从假设条件出发，如果能够推导出一个矛盾的结论，就可以证明假设条件不成立。
2. 假设检验中的基本概念
（1）假设检验的步骤
假设检验一般可以分为以下几个步骤：根据实际问题提出一对假设（零假设和备择假设）；构造某个适当的检验统计量，并确定其在零假设成立时的分布；根据观测的样本计算检验统计量的值；根据指定的显著性水平确定检验统计量的临界值并进而给出拒绝域；根据决策规则得出拒绝或不能拒绝零假设的结论。
注意这里“不能拒绝零假设”不同于“接受零假设”。“不能拒绝零假设”只是说明根据现有样本不能认为零假设有问题，但重新抽取一个样本就有可能推翻零假设。如果说成“接受零假设”，就等于说你认为零假设在任何条件下都是成立的。
（2）零假设和备择假设的选择
在假设检验中零假设和备择假设是互斥的，等号必须出现在零假设中；最常用的有三种情况：双侧检验、左侧检验和右侧检验。以单个总体均值的假设检验为例，零假设和备择假设的三种情况如表 4-6。

表 4-6　零假设和备择假设的三种情况

	双侧检验	左侧检验	右侧检验
H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$

单侧检验一般根据以下几个原则设置零假设和备择假设：把研究者要证明的假设作为备择假设；将所作出的声明作为零假设；把现状（Status Quo）作为零假设；把不能轻易否定的假设作为零假设。
在单侧检验中，零假设和备择假设设置不同，有可能得出不同的结论。对于一个实际问题，究竟将哪个假设作为零假设不是可以随意选择的，而要遵循一定的原则：零假设通常是指观测结果表现出的差异仅是由随机因素引起，不是由于某种结构性原因引起，备择假设则相反，认为观测到的差异是由于某种原因造成的，不仅仅是随机差异；很多问题往往将研究者要证明的结论放在备择假设，这样如果通过假设检验作出拒绝零假设选择备择假设的判断，会使得我们要证明的结论更具有说服力。
例如，某种汽车原来平均每加仑汽油可以行驶 24 英里。研究小组提出了一种新工艺来提高每加仑汽油的行驶里程。为了检验新的工艺是否有效，需要生产一些产品进行测试。这里要证明的结论是μ>24，因此零假设和备择假设的选择为： H_0 ：μ≤24 H_1：μ>24。某种减肥产品的广告中声称使用其产品平均每周可减轻体重 8 公斤以上。要检验这种声明是否正确你会如何设定零假设和备择假设？由于该产品的现状为 μ≥8，因此零假设和备择假设应设为：H_0：μ≥8，H_1：μ<8。
（3）检验统计量和拒绝域
我们决策时所依据的样本统计量称为检验统计量。检验统计量所有可能的取值集合可分成两个不相交的部分，可以拒绝零假设的检验统计量取值的集合称为拒绝域；不能拒绝零假设的检验统计量取值的集合称为接受域；划分拒绝域和接受域的数值称为临界值。一旦有了样本，计算检验统计量的观测值，就可以根据规则决定是否拒绝零假设。
（4）假设检验中的两类错误
在假设检验中可能出现两类错误，如表 4-7。在零假设正确的情况下我们作出拒绝零假设的判断，这时发生的错误称为第一类错误（拒真错误）；相反地，在零假设错误的情况下我们作出接受零假设的判断，这种错误称为第二类错误（取伪错误）。

表 4-7　假设检验中的两类错误

决策	实际情况	
	H_0 为真	H_0 为假

接受 H_0	正确	第二类错误
拒绝 H_0	第一类错误	正确

在假设检验中这两类错误是不可避免的。通常来说要减小其中的一种错误，只能通过增加另一种错误的方法实现。在假设检验中通常的做法是控制犯第一类错误的概率不超过某个水平α，在满足该条件的前提下使犯第二类错误的概率尽量小。

允许犯第一类错误的概率α称为显著性水平，通常α取为 0.01，0.05，0.1。根据α可以确定检验统计量的临界值，从而进一步得出检验结论。

二、假设检验中的 p 值

在计算机和统计软件得到广泛应用之前，假设检验的决策规则一般是根据检验统计量的临界值给出的：当根据样本计算出的检验统计量的观测值落到拒绝域时拒绝零假设。传统方法的不足之处是，检验统计量的临界值一般需要根据给定的显著性水平查表得出，在实际应用中不够方便，如果显著性水平有所变化，还需要重新查表得到新的临界值。

在统计软件中最常用的假设检验方法是根据检验统计量的观测值计算p值，然后将p值与α比较得出检验结论，当p值小于α时拒绝零假设。简单的说，p值是在零假设成立的条件下，出现检验统计量的样本观测结果或更极端结果的概率，也称为观测到的显著性水平，是能拒绝H_0的α的最小值。p值几乎不可能通过手工方式来计算，但在统计软件中它的计算非常容易。根据p值进行假设检验与传统的根据临界值的检验方法完全等价，但由于使用p值更方便，因而在实际应用中这种方法已经取代了根据临界值进行检验的方法。

根据p值进行假设检验时，决策规则是不变的：p值小于α时拒绝H_0，但p值的计算方法与检验的种类（双侧检验、左侧检验还是右侧检验）以及分布的类型有关。对于t分布，用t_{obs}表示t统计量的观测值，双侧检验时p值=P($|t|\geq|t_{obs}|$)；右侧检验时p值=P($t\geq t_{obs}$)；左侧检验时p值=P($t\leq t_{obs}$)。正态分布时p值的计算与t分布类似，只是将t统计量换成z统计量。

p值的含义可以用以下三个图形来说明。假设t统计量的样本观测值等于2，在双侧检验时的p值等于右侧阴影面积的两倍，如图4-8所示；右侧检验时的p值等于右侧的阴影面积，如图4-9所示；左侧检验时的p值等于观测值左侧的阴影面积，如图4-10所示。

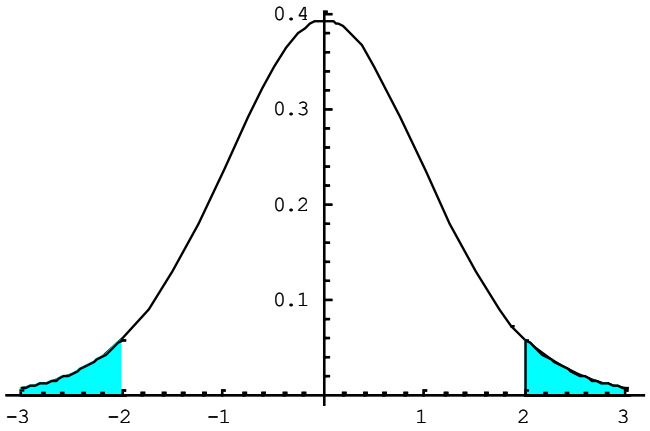


图4-8　t_{obs}=2，双侧检验时的p值等于阴影部分的面积

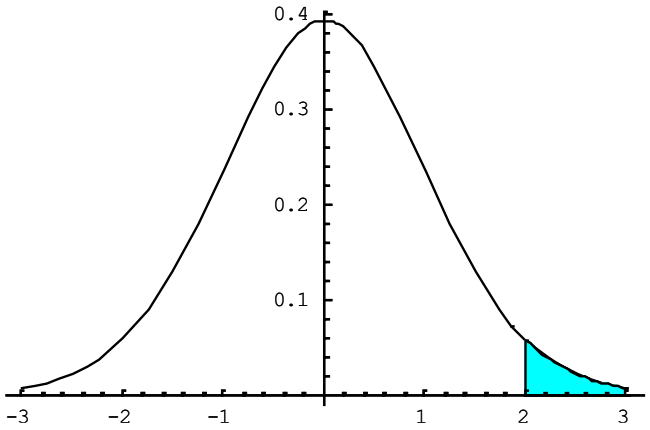


图4-9　t_{obs}=2，右侧检验时的p值等于阴影部分的面积

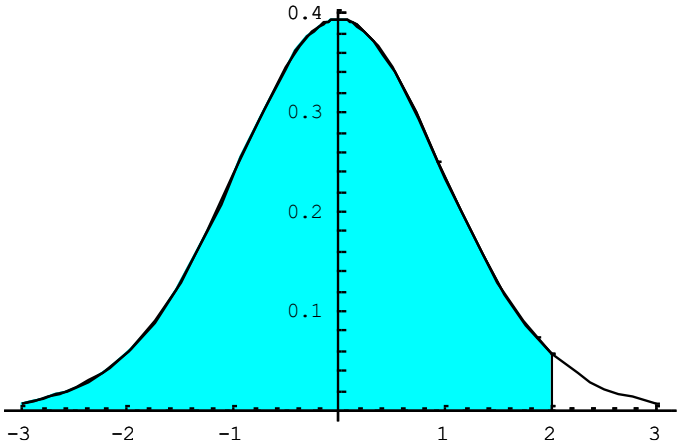


图4-10　t_{obs}=2，左侧检验时的p值等于阴影部分的面积

三、几种常用假设检验方法的软件实现和结果分析

接下来我们来看三种常用的假设检验的软件实现和结果分析：单个总体均值的 t 检验，基于两个独立样本的 t 检验，以及基于两个匹配样本的 t 检验。

1. 单样本t检验

根据单个样对总体均值进行假设检验，要求是大样本或者总体服从正态分布，如果不满足条件则需要使用非参数检验方法。由于应用中一般需要根据样本方差估计总体方差，因此实际中都是根据t统计量进行检验。

【例4.5】使用数据文件“学生调查.sav”，以5%的显著性水平检验能否认为总体中学生的平均体重等于60公斤。

解：在SPSS中打开相应的数据文件，选择“分析”→“比较均值”→“单样本t检验”，在弹出的对话框中将体重变量作为检验变量，检验值框中填入60，其余使用系统默认值，输出结果如表4-8。

表4-8 单样本t检验

	检验值 = 60					
					差分的 95% 置信区间	
	t	df	Sig.(双侧)	均值差值	下限	上限

	检验值 = 60					
					差分的 95% 置信区间	
	t	df	Sig.(双侧)	均值差值	下限	上限
体重	-1.872	34	.070	-3.057	-6.38	.26

根据题目的要求，这里应采用双侧检验，零假设和备择假设为：$H_0:\mu=60 \leftrightarrow H_1:\mu\neq 60$。表4-8中给出的Sig.（双侧）就是双侧检验时的p值。这里p值=0.07，大于0.05，因此检验的结论是不能拒绝总体均值等于60的零假设。图4-11显示了本题中p值和α的大小关系，图中深色阴影部分的面积为α，浅色阴影的面积为p值。

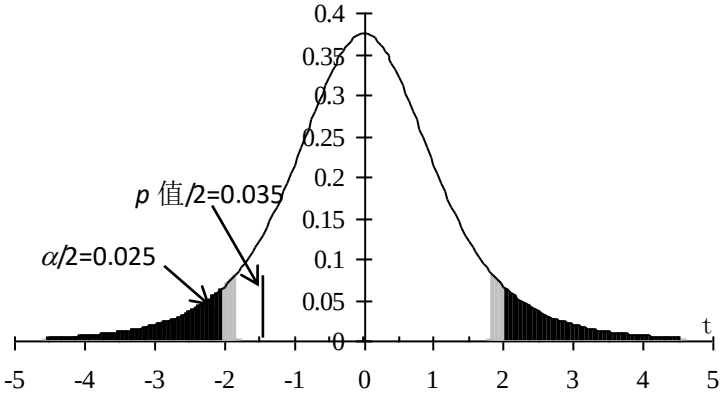


图4-11　双侧检验时p值和α的大小关系

在这个题目中如果希望证明的是总体均值大于60，则是一个右侧检验的问题，零假设和备择假设为：$H_0:\mu\leq 60 \leftrightarrow H_1:\mu>60$。这时SPSS软件的输出结果保持不变，但p值和$\alpha$的大小关系如图4-12所示，由于p值大于$\alpha$，因此检验的结论是不能拒绝零假设。

而如果我们希望证明的是总体均值小于60，则是一个左侧检验的问题，零假设和备择假设为：$H_0:\mu\geq 60 \leftrightarrow H_1:\mu<60$。这时SPSS软件的输出结果仍然保持不变，但p值和α的大小关系如图4-13所示，由于p值小于α，因此检验的结论是拒绝零假设。

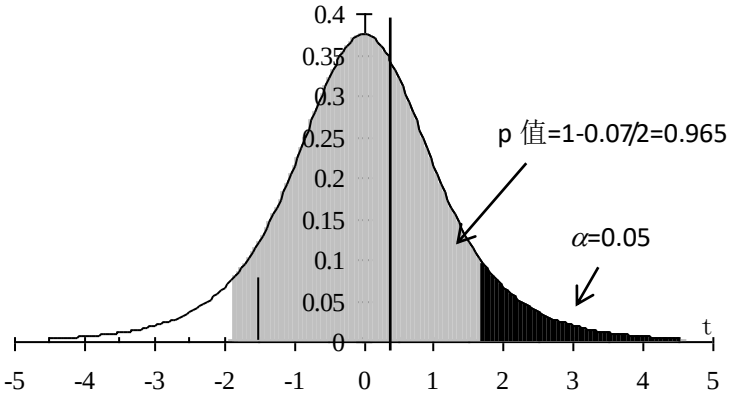


图4-12　右侧检验时p值和α的大小关系

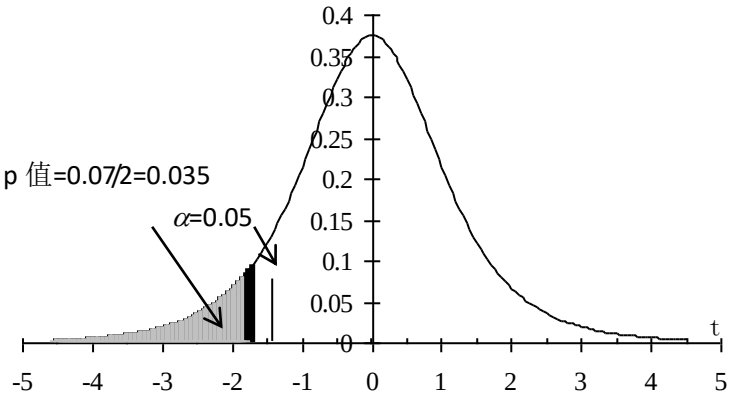


图4-13　左侧检验时p值和α的大小关系

2. 两个独立样本的t检验

两个独立样本的 t 检验需要首先判断两个样本数据是否是大样本或者来自正态总体，如果不满足条件则需要使用非参数检验方法。然后，在正态分布或大样本时，需要检验两个总体的方差是否相等（F 检验）；方差相等时采用等方差的 t 检验，方差不相等时采用不等方差的 t 检验。

与一个总体的情况类似，两个总体均值假设检验中也可以进行单侧或双侧检验。两个总体均值假设检验中的备择假设也可以有以下三种情况：$H_1:\mu_1 \neq \mu_2$；$H_1:\mu_1 > \mu_2$；$H_1:\mu_1 < \mu_2$。下面我们来看使用 SPSS 进行独立样本 t 检验的过程。

【例4.6】使用数据文件“学生调查.sav”，用SPSS检验在5%的显著性水平下男女生的平均身高是否相等，假设身高服从正态分布。

选择“分析”→“比较均值”→“独立样本t检验”，弹出的对话框如图4-14。把身高变量作为检验变量，性别作为分组变量。然后单击“定义组”按钮来设置分组规则，这里在两个矩形框中分别输入0和1，单击“继续”返回主对话框，单击“确定”就可以了。输出结果如表4-9。

图4-14　独立样本t检验的相关设置

表4-9　SPSS两个独立样本均值t检验的输出结果

	方差是否相等的 Levene 检验	均值是否相等的t检验

									差分95%置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
身高	假设方差相等	4.195	.049	7.845	33	.000	10.207	1.301	7.560	12.854
	假设方差不相等			7.449	21.459	.000	10.207	1.370	7.361	13.053

注意表4-9中包含了等方差的检验、等方差时的t检验以及异方差时的t检验结果。读这个表时先看等方差的检验的结果（数据的前两栏），这里SPSS使用的是Levene检验，这种检验不需要正态性的假设条件，比F检验更稳健。如果Levene检验的*p*值小于显著性水平（如5%），则认为总体是异方差的，接下来在t检验中要使用方差不等时的检验结果；如果方差检验中认为两个总体是等方差的，则在t检验中使用方差相等时的检验结果。在这个例子中，方差检验的p值等于0.049<0.05，因此拒绝方差相等的原假设，认为男女生身高的方差不等。接下来的t检验中使用方差不等的检验结果（下面的一行），检验的*p*值保留三位小数时等于0.000，小于常用的α值，因此认为总体中男女生的身高均值不相等。

3. 匹配样本的t检验

如果两个样本中的数据有一一对应的关系，也就是说属于匹配样本（paired-sample)，由此对两个总体的均值进行假设检验时，需要先将两个样本的对应数据相减，得到一个新的变量 *d*，然后对新的变量 *d* 应用单个样本的 t 检验。匹配样本的 t 检验中需要变量 *d* 服从正态分布，或者是大样本，否则需要使用非参数的检验方法。

【例4.7】随机选择了 8 名肥胖儿童试验一种减肥方案，减肥前后的体重如表4-10。根据实验结果，在5%的显著性水平下能否认为减肥方案有效？

表4-10　一种减肥方案的试验数据

减肥前	45	55	54	48	56	53	62	49
减肥后	43	48	50	47	50	47	59	46

解：把数据输入SPSS数据表，选择“分析”→“比较均值”→“配对样本t检验”，在弹出的对话框中将两个变量作为一组数据选为分析变量，相关设置见图4-15，输出结果如表4-11。

图4-15　匹配样本t检验的相关设置

表4-11　成对样本t检验的输出结果

	成对差分					t	df	Sig.(双侧)
				差分的 95% 置信区间				
	均值	标准差	均值的标准误	下限	上限			

	成对差分					t	df	Sig.(双侧)
	均值	标准差	均值的标准误	差分的 95% 置信区间 下限	差分的 95% 置信区间 上限			
对 1 减肥前 - 减肥后	4.00000	2.13809	.75593	2.21251	5.78749	5.292	7	.001

根据本题的题意，将假设检验的零假设设为 $\mu_1 - \mu_2 \le 0$，备择假设设为 $\mu_1 - \mu_2 > 0$。如果拒绝零假设则说明减肥方案有效。

根据表4-11，检验的t统计量等于5.292，双侧检验的p值为0.01，因此右侧检验的p值为0.01/2=0.005。在5%的显著性水平下显然应拒绝零假设，结论是减肥方案有效。

小 结

参数估计和假设检验是根据随机样本的信息对总体进行统计推断的两类基本方法，有着广泛的应用。由于相关知识在先修课程中已经学习过，本章主要在回顾相关知识的基础上，补充讲解必要样本容量的计算、p值、参数估计和假设检验方法的软件操作和结果分析等内容。

抽样分布是指统计量的概率分布，是区间估计和假设检验的基本依据。抽样分布则一般利用概率统计的理论推导得出，在应用中也是不能直接观测的，其形状和参数可能完全不同于总体或样本数据的分布。

参数估计是指利用样本信息对总体数字特征作出的估计。在SPSS软件中可以直接给出简单随机抽样、有放回条件下按照t统计量计算的置信区间。在SPSS软件中进行假设检验时也会输出相应的置信区间。

样本容量的确定受到多方面因素的影响，具体的确定方法会因为抽样方法的不同而不同，本章给出了简单随机抽样情况下的样本量的计算方法。不重复抽样时的必要样本容量要小于重复抽样的情况。

假设检验是确定根据样本观测数据能否拒绝关于总体的某种陈述的统计方法。假设检验的基本原理是：小概率事件在一次试验中不可能发生。

在双侧检验和单侧检验中假设检验的零假设和备择假设的设置方法有所不同。具体问题不

同，假设检验中的检验统计量也可能不同。主要有Z统计量，t统计量，χ^2统计量和F统计量等。

在统计软件中一般根据检验统计量的样本观测值计算出一个概率值提供给用户，由用户根据实际情况（单侧还是双侧检验等）确定p值的大小，然后根据p值和α的大小关系得出检验结论。p值是零假设成立时，出现检验统计量的样本观测结果或更极端结果的概率。p值越小，说明结果越不可能发生，拒绝零假设的证据就越强。在大部分情况下SPSS软件提供的sig.值就是需要的p值，但在单侧检验时的p值需要根据sig.值进行推算。

本章介绍了三种常用假设检验的软件实现和结果分析：单个总体均值的t检验，基于两个独立样本的t检验，以及基于两个匹配样本的t检验。

思考与练习

1. 怎样确定假设检验问题的零假设和备择假设？
2. 什么是抽样分布？常用的抽样分布有哪些？
3. 假设检验有哪些步骤？
4. 单侧和双侧假设检验问题的拒绝域有何区别？
5. 怎样理解假设检验问题的p值？如何根据p值和显著性水平的关系得出检验结论？
6. 根据表 4-3 对 100 名儿童随机调查的结果（数据文件：看电视时间.sav），能否认为
（1）该地区儿童平均看电视的时间等于 25.5 小时？
（2）该地区儿童平均看电视的时间小于 25.5 小时？
（3）该地区儿童平均看电视的时间大于 25.5 小时？
7. 一培训机构拟在会计学和管理学之间选择一门专业进行培训，在确定培训专业之前在当地进行了一项薪水调查。分别从两个专业的大学毕生中抽取 10 个人调查他们的年薪，结果得到数据如表 4-12。假设总体服从正态分布，把数据整理成两个变量（专业、收入）的形式，使用统计软件检验在 5%的显著性水平下检验总体中两个专业的平均年薪有无显著差异。

表 4-12　两个专业大学毕业生的年薪（单位：万元）

会计	4.6	3.8	4.2	5.0	5.5	4.8	8.0	6.2	6.7	3.2
管理	4.5	6.9	4.0	9.0	3.6	6.3	5.0	7.9	3.5	2.4

8. 某市场研究公司调查了 10 个人在广告播出前后的购买潜力等级分值，分数越高说明购买潜力越高。试检验广告是否有明显效果？假设两个总体服从正态分布，显著性水平=0.05。

表 4-13　10 个人在广告播出前后的购买潜力等级分值

个体	1	2	3	4	5	6	7	8	9	10
广告后	6	6	7	4	3	9	7	6	5	6
广告前	5	4	7	3	5	8	5	6	4	6

第五章　方差分析

毕业生的起薪是衡量大学生就业质量的重要指标之一。假设为了比较四个专业的起薪，我们从某高校四个专业的毕业生中分别随机选择6人调查他们的起薪。我们的任务是根据样本数据对不同专业毕业生的平均起薪进行比较，判断它们之间是否有显著差异。

用什么方法进行检验呢？我们之前学过的两个独立样本的t检验可以检验两个总体的均值是否相等。在上面的例子中对四个样本均值进行两两的比较（共需进行6次t检验），如果在所有的两两比较中平均起薪都没有明显差异，则说明专业对起薪没有显著影响；如果至少在一次t检验中均值的差异是显著的，就说明专业对起薪有显著影响。

采用t检验对多个总体进行分析的主要问题是犯第一类错误的概率显著增大了：假设我们在每次t检验中犯第一类错误的概率都等于5%，那么在整体检验中不犯第一类错误就要求每次检验中都不犯第一类错误，这一概率等于$(1-0.05)^4=0.7351$，从而在整体检验中犯第一类错误的概率为1-0.7351=0.2649。

在需要比较多个总体均值的情况下，方差分析（Analysis of variance，ANOVA）是更适当的方法。方差分析的优点是计算量较小，并且犯第一类错误的概率保持不变，比t检验更通用。在上面的例子中，如果我们还需要同时考虑其他因素对起薪的影响（例如性别），t检验就无能为力了，而方差分析则能够轻易地处理这种情况。

第一节　方差分析中的基本概念和假设

一、方差分析中的基本概念

方差分析一般用来对多个总体的均值进行推断，检验多个总体均值之间差异的显著性。在只比较两个均值时这种方法与两个独立样本的t检验是等价的。这一方法之所以被称为“方差”分析，是因为这种方法中对均值差异性的检验是通过对方差的分解进行的。方差分析的思想在20世纪20年代由英国统计学家费希尔（R.A. Fisher）最早提出，开始应用于生物和农业田间试验，以后在许多学科领域得到了广泛应用。

方差分析一般用来分析一个定量变量与一个或多个定性变量的关系，例如大学生毕业生的起薪与学生的性别、专业、毕业院校之间的关系，企业销售额与广告媒体（报纸和电视）以及定价策略（低、中、高）的关系，某种疾病的治疗效果与治疗方案之间的关系，产品性能与生产工艺之间的关系等等。在方差分析中，作为结果的变量称为因变量（dependent variable)，例如毕业生的起薪、企业的销售额等；作为原因的、把观测结果分成几个组以进行比较的变量称为自变量（independent variable)，例如所学专业、广告策略、生产工艺等。

我们前面已经指出，统计研究中的数据主要有两种来源：观察和实验。方差分析所针对的数据一般是经过专门设计而收集的实验数据，通过科学的设计可以保证数据符合方差分析的要求，并且提高数据的利用效率，根据较少的数据得出检验结论。因此，方差分析与“实验设计”这一统计领域有着非常密切的联系。当然，这并不是说方差分析不能用于观察数据，只要满足方差分析所需要的假设条件即可。

由于与实验研究有密切联系，方差分析中也使用了一些与实验有关的概念。在方差分析中，自变量被称为因素(factor)；因素的不同表现，也就是每个自变量的不同取值称为因素的水平(level)。只有一个自变量的方差分析称为单因素方差分析（one-factor ANOVA）；如果要同时研究多个因素对因变量的影响，则称为多因素方差分析（multi-factor ANOVA），其中最简单的情况是双因素方差分析（two-factor ANOVA）。

方差分析模型可以分为固定效应模型与随机效应模型。在固定效应模型中，因素的所有水平都是由实验者审慎安排而不是随机选择的，而在随机效应模型中因素的水平是从多个可能的水平中随机选择的。例如为了研究所学专业对学生就业质量的影响，从所有专业中随机选择10个进行研究，如果重复进行一次实验被选中的专业很可能是不同的，在这种情况下“专业”这一因素的效应就是随机的。相反，如果研究目的是比较10个特定专业就业质量之间的差异，则在每次实验中专业这一因素都只能有这10个固定的水平，这时专业这一因素的效应即为固定效应。固定效应模型和随机效应模型的侧重点有所不同，本章介绍的都是固定效应模型。

二、方差分析中的基本假设与检验

1. 方差分析中的基本假设

我们学习任何一种统计方法时都要注意这种方法的适用条件，注意检验这种方法是否适用于你的数据。方差分析是对多个总体的比较，比较中需要以下三个假设条件：

（1）在各个总体中因变量都服从正态分布；

（2）在各个总体中因变量的方差都相等；

（3）各个观测值之间是相互独立的。

2. 方差分析中假设条件的检验方法

（1）正态性检验

正态性假设可以通过观察各组数据的直方图、Q-Q图等来判断，也有一些统计检验方法，例如K-S检验等。需要特别注意的是，正态性检验不是对数据整体分布的检验，而是对按因素水平分组后各组数据的检验。此外，在很多实验中实际得到的数据数量非常少，例如每组中只有两三个数据，在这种情况下没有检验正态性的很好方法，这时这一假设是否成立需要根据所研究的现象本身的性质加以判断。

（2）方差齐性检验

对各总体方差是否相等的检验称为方差齐性检验。检验各组之间的方差是否相等的一个经验方法是，计算各组数据的标准差，如果最大值与最小值的比例小于2：1，则可以认为是数据同方差的。如果用各组的方差进行比较，则要求最大值与最小值的比例小于4：1。Levene检验是一种更为正式的检验方法。

（3）关于基本假设的进一步说明

方差分析对前两个假设条件是稳健的。一般来说，方差分析中可以允许数据的分布对正态分布一定程度的偏离。如果样本容量很大，方差分析也可以应用于非正态的情况，因为这时中心极限定理可以保证样本均值的抽样分布为正态分布。在各个组别中的样本量比较接近的情况下，也可以粗略的认为数据是等方差的。

独立性的假设条件一般可以通过对数据搜集过程的控制来保证，在方差分析中很少对这一假设进行直接的统计检验。

如果数据确实严重偏离了前两个假设条件，则使用方差分析时需要先对数据进行数学变换，例如取对数、开方等，也可以使用非参数的方法（例如Kruskal - Wallis检验）来比较各组的均值。

接下来我们将分别介绍单因素和双因素方差分析的基本原理和应用。

第二节　单因素方差分析

一、单因素方差分析的数据结构和模型

单因素方差分析的所要分析的问题是：根据分别来自r个等方差正态总体的数据检验这些总体的均值是否相等。为了表述方便，我们假设在单因素方差分析中所研究的因素为因素A，共有r个水平，每个水平的样本容量为m，共有n=rm个观察值。根据这些条件，单因素方差分析的数据结构见图5-1。

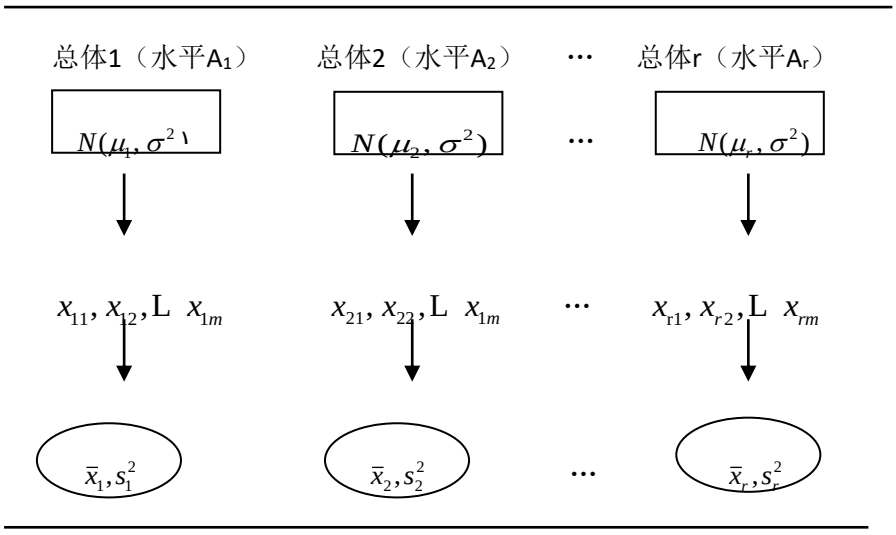


图5-1　单因素方差分析的数据结构

在单因素方差分析模型中，任何一个样本数据都包含了三部分因素的影响：总体平均水平的影响；因素水平的影响；以及随机因素的影响。单因素方差分析模型可以写成公式（5-1）：

$$X_{ij}=\mu_i+\varepsilon_{ij}\ =\mu+\alpha_i+\varepsilon_{ij} \qquad (5\text{-}1)$$

其中

i=1到r，代表因素的不同水平。

j=1到m，代表在同一因素水平下的不同观测值。

μ = 根据所有数据计算的总均值。

μ_i = 第i组的均值。

α_i = 第i组的均值与总均值的差。

ε_{ij} = 随机误差项，$\varepsilon_{ij}\sim N(0,\sigma^2)$。

二、方差分析的基本原理

为了说明方差分析的基本原理，我们先来看总离差平方和SST（Sum of squares for total）的分解问题。SST也称为总变异（Total Variation），用$\bar{\bar{x}}$表示所有数据的总均值，其计算公式是：

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{m}(x_{ij}-\bar{\bar{x}})^2 \tag{5-2}$$

总离差平方和可以分解成两个组成部分。组间离差平方和SSA（Sum of squares for factor A），也称为解释的变异，用 $\bar{x}_i$ 表示各组的组均值，计算公式为：

$$SSA=\sum_{i=1}^{r}m(\bar{x}_i-\bar{\bar{x}})^2 \tag{5-3}$$

组内离差平方和SSE（Sum of squares for error），是与自变量无关的由不可控因素（例如不可控制的个体差异，随机因素，测量误差等）引起的变异，计算公式为：

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)^2 \tag{5-4}$$

可以证明，SST、SSA和SSE之间有以下关系：

$$SST=SSA+SSE \tag{5-5}$$

SST、SSA和SSE具有不同的自由度。对于SST来说，因为它只有一个约束条件，即 $\sum\sum(x_{ij}-\bar{\bar{x}})=0$，在n个$x_{ij}$中有n-1个可以自由取值，因此它的自由度为n-1；对于SSA来说，其约束条件为 $\sum m(\bar{x}_i-\bar{\bar{x}})=0$，因而在 $\bar{x}_1$，$\bar{x}_2$，…，$\bar{x}_r$ 这r个变量中只有r-1个是可以自由取值的，SSA的自由度为r-1；对SSE来说，由于对每一个水平i都要求 $\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)=0$，因此它共有r个约束条件，SSE的自由度为n-r。SST、SSA、SSE的自由度有以下关系：n-1=(r-1)+(n-r)。

SSA、SSE分别除以它们的自由度就可以得到组间和组内均方（Mean Square），MSA和MSE。

$$MSA=\frac{SSA}{r-1} \tag{5-6}$$

$$MSE=\frac{SSE}{n-r} \tag{5-7}$$

可以证明，$E(MSE)=\sigma^2$，而 $E(MSA)=\sigma^2+\frac{m}{r-1}\sum_{i=1}^{r}(\mu_i-\mu)^2$。因此，当 μ_i 都相等（原假设成立）时MSA与MSE都是对模型中随机误差项方差的无偏估计。μ_i 之间的差异越大，MSA的期望值MSE的期望值的比值就越大。

那么，MSA和MSE相差多大时可以认为这种差异是显著的呢？理论分析表明，在零假设成立时MSA和MSE的比值服从自由度为r-1和n-r的F分布。因此我们可以设定一个显著性水平α，通过对这个检验统计量的分析做出接受或拒绝原假设的决策。上述计算过程一般用方差分析表来表示（表5-1）。

$$F=\frac{MSA}{MSE} \qquad (5\text{-}8)$$

表5-1　单因素方差分析表

变异来源	离差平方和 SS	自由度 df	均方 MS	F值
组　间	SSA	r-1	MSA	MSA/MSE
组　内	SSE	n-r	MSE	
总变异	SST	n-1		

三、方差分析的步骤

方差分析过程包括以下基本步骤：

（1）检验数据是否符合方差分析的假设条件。

（2）提出零假设和备择假设。不管所研究问题的背景如何，单因素方差分析中的零假设总是相同的：各总体的均值之间没有显著差异，即$H_0:\mu_1=\mu_2=\Lambda=\mu_r$；备择假设也相同：至少有两个均值不相等，即$H_1$：$\mu_1$，$\mu_2$，…，$\mu_r$不全相等。

（3）根据样本计算F统计量的值和p值。

（4）根据决策规则得出检验结论。决策规则可以用两种方式来表述。一是根据事先确定的显著性水平α和自由度计算F检验的临界值，当实际值大于临界值时拒绝零假设（图5-2）。二是根据样本统计量计算p值，当α>p值时拒绝零假设。

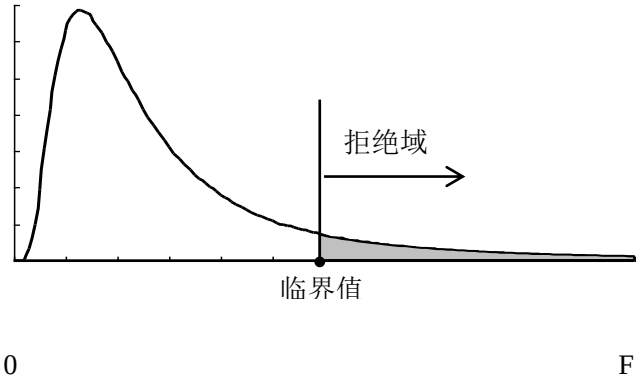


图5-2　F检验的临界值和拒绝域

【例5.1】　从某高校四个专业的毕业生中分别随机选择6人调查他们的起薪，数据见表5-1（数据文件起薪1.xls），表中用1，2，3，4表示四个专业。试分析四个专业毕业生的起薪是否有显著差异。

我们使用SPSS的“单因素方差分析”模块做方差分析，输出各组的描述性统计指标（表5-3）、方差齐性检验的结果（表5-4）。注意这里我们对软件输出结果的格式进行了修改，并删除了不需要的内容。

（1）关于方差分析基本假设的检验。由于每一组中只有6个观察值，我们很难分析数据

分布分布的正态性。为了使用方差分析的方法，假设各组数据来自正态分布总体。在表5-3中，标准差的最大值和最小值的比值为1.58，小于2，因此可以认为是等方差的；根据表5-4的Levene检验的结果，由于表中的p值等于0.7060，是个非常大的值，因此也不能拒绝等方差的原假设。

（2）方差分析的结果分析（表5-5）。检验中的零假设和备择假设为：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$，H_1：μ_1、μ_2、μ_3、μ_4不全相等。表5-5给出的p值等于0.002，小于我们通常要求的α值，因此我们应拒绝零假设，从而得出专业对起薪有显著影响的结论，也就是说不能认为四个专业的起薪都相等。

表5-2大学毕业生的专业和起薪

序号	*专业*	*起薪（元）*	*序号*	*专业*	**起薪（元）**
1	1	3000	13	3	2000
2	1	3100	14	3	2600
3	1	3300	15	3	2500
4	1	4000	16	3	3500
5	1	3700	17	3	3000
6	1	3500	18	3	2800
7	2	4000	19	4	2200
8	2	3000	20	4	2400
9	2	2500	21	4	2000
10	2	3500	22	4	3000
11	2	4000	23	4	2000
12	2	3700	24	4	2800

表5-3 例5.1的描述统计指标

	N	*均值*	*标准差*
1	6	3433	378
2	6	3450	596
3	6	2733	505
4	6	2400	420

表 5-4 例 5.1 的方差齐性检验

Levene 统计量	*df1*	*df2*	*显著性*
0.4708	3	20	0.7060

表5-5 例5.1的方差分析表

	平方和	df	均方	F	p 值
组间	4927916.667	3	1642638.889	7.078	0.002
组内	4641666.667	20	232083.333		
总数	9569583.333	23			

在前面的分析中为了表述方便我们一直假定各个水平下的样本容量都是相等的。各水平下的样本容量不同时单因素方差分析的方法也完全适用，只是公式的形式稍有不同，在使用

软件进行分析时几乎看不出这种差别。下面我们看一个样本容量不同的例子。

例【5.2】 一份研究伐木业对热带雨林影响的统计研究报告中指出，“环保主义者对于林木采伐、开垦和焚烧导致的热带雨林的破坏几近绝望”。这项研究比较了类似地块上树木的数量，这些地块有的从未采伐过，有的1年前采伐过，有的8年前采伐过。根据表5-6中的数据（数据文件林木采伐.xls），采伐对树木数量有显著影响吗？

表5-6 不同地块的树木数量

采伐状况	*树木数量*	*采伐状况*	*树木数量*
1	27	2	18
1	22	2	17
1	29	2	14
1	21	2	14
1	19	2	2
1	33	2	17
1	16	2	19
1	20	3	18
1	24	3	4
1	27	3	22
1	28	3	15
1	19	3	18
2	12	3	19
2	12	3	22
2	15	3	12
2	9	3	12
2	20		

采伐状况：1表示从未采伐过；2表示1年前采伐过；3表示8年前采伐过。

在SPSS菜单中选择分析→比较均值→单因素方差分析，经过相应的设定后可以输出各组的描述性统计指标（表5-7）、方差齐性检验的结果和方差分析表（表5-8）。

（1）关于正态性的分析。使用SPSS软件得出的分组的直方图如图5-3。图5-3表明，在各个水平下林木数量都呈对称分布，没有极端值出现，因此可以认为不违背正态性假设。

（2）方差齐性检验。表5-7表明，各组的标准差差异不大，最大值与最小值之比等于1.16，明显小于2，因此可以认为是等方差的；方差齐性检验中Levene统计量的值为0.259，对应的p值为0.774，也支持上述结论。

（3）综上所述，可以用方差分析对热带雨林采伐的影响进行比较。检验的零假设是雨林采伐对林木数量没有显著影响；备择假设是有显著影响。根据表5-8的方差分析表，p值保留三位小数时为0.000，远远小于0.05，因此检验的结论是采伐对林木数量有显著影响。

表5-7 不同采伐状况地块林木数量的描述统计

	N	均值	标准差
1	12	23.75	5.065
2	12	14.08	4.981
3	9	15.78	5.761

表5-8　雨林采伐研究的方差分析表

	平方和	df	均方	F	p 值
组间	625.157	2	312.578	11.426	.000
组内	820.722	30	27.357		
总数	1445.879	32			

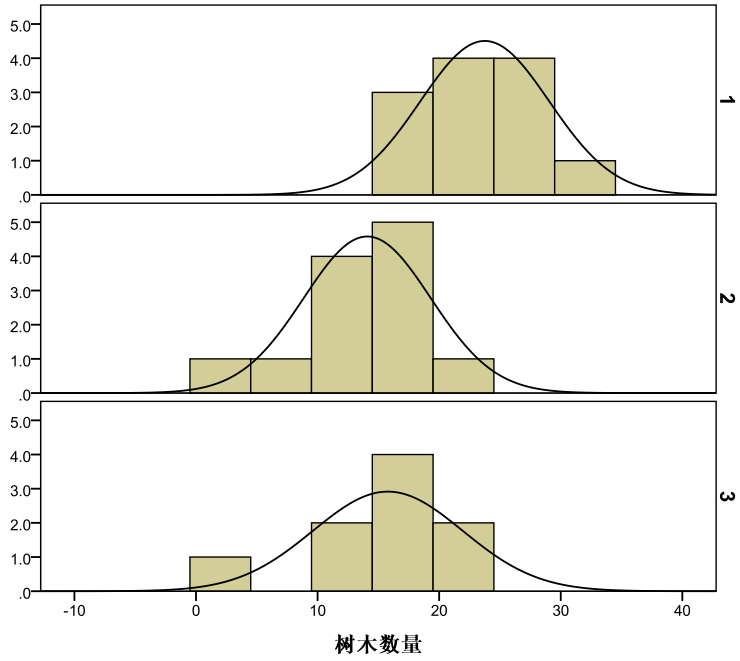


图5-3　不同采伐状况的地块树木数量的直方图

四、方差分析中的多重比较

在方差分析中，当零假设被拒绝时我们可以确定至少有两个总体的均值有显著差异。但要进一步检验哪些均值之间有显著差异还需要采用多重比较的方法进行分析，这在方差分析中称为事后检验(Post Hoc test)。当然，如果在F检验中不能拒绝原假设就不需要做事后检验了。

多重比较是对各个总体均值进行的两两比较，有很多种方法，如Fisher最小显著差异（Least Significant Difference，LSD）方法，Tukey的诚实显著差异（Honestly Significant Difference，HSD）方法等，这里我们只介绍Fisher的最小显著差异方法。

LSD方法与我们前面学过的两个总体均值的t检验非常类似，其检验步骤如下：

（1）写出检验的零假设和备择假设检验：H_0：$\mu_i = \mu_j$，H_1：$\mu_i \neq \mu_j$。

（2）计算检验统计量 $t = \dfrac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE(\dfrac{1}{n_i} + \dfrac{1}{n_j})}}$

（5-9）

这个公式与两个总体均值t检验的公式非常类似，只是将原来的S_p^2（根据两个总体估计的总体方差）换成了MSE（根据多个总体估计的总体方差）。

（3）做出决策。如果$t < -t_{\alpha/2}$或$t > t_{\alpha/2}$则拒绝H_0；也可以根据p值和显著性水平α的大小关系得出结论，p值< α时拒绝H_0；还可以计算$\bar{x}_i - \bar{x}_j$的置信区间$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2}\sqrt{MSE(\frac{1}{n_i}+\frac{1}{n_j})}$，如果0包含在置信区间内则不能拒绝$H_0$。

其中t检验的临界值$t_{\alpha/2}$是根据自由度n-r和显著性水平α确定的。n是全部样本单位数，r是因素A的水平数。

表5-9是SPSS对【例5.2】进行多重比较的输出结果。结果表明，从未采伐过的地块与1年前采伐过的地块林木数量均值之差的置信区间为5.31~14.03，区间不包括0，因此差异是显著的。而1年前采伐过的地块与8年前采伐过的均值之差的置信区间为-6.04~3.02，区间包括了0，因此二者的差异不显著。

表5-9 雨林采伐研究中多重比较的SPSS输出结果

(I) 采伐类型	(J) 采伐类型	均值差 (I-J)	标准误	p 值	95% 置信区间	
					下限	上限
从未采伐过	1 年前采伐过	9.67	2.14	0.0001	5.31	14.03
	8 年前采伐过	5.97	2.31	0.0017	3.26	12.68
1 年前采伐过	从未采伐过	-9.67	2.14	0.0001	-14.03	-5.31
	8 年前采伐过	-1.69	2.31	0.4682	-6.40	3.02
8 年前采伐过	从未采伐过	-5.97	2.31	0.0017	-12.68	-3.26
	1 年前采伐过	1.69	2.31	0.4682	-3.02	6.40

第三节　双因素方差分析

在实际工作中我们遇到的问题大部分都是复杂的，会涉及多个自变量。通过一个变量就能够完全解释一种现象的情况并不多见。例如在毕业生起薪的例子中，要分析专业对起薪的影响，我们还需要考虑到所有可能影响起薪的因素，如毕业生的性别、行业、毕业院校等等。方差分析可以同时分析多个因素的影响。

与t检验相比方差分析的另一个优势是可以分析因素之间的交互作用（interaction effect）。当一个因素对因变量的影响程度受另一个因素的影响时，我们就说两个因素之间存在交互作用。下面我们通过一个例子来说明。假设我们需要向大学在校学生和企业在职员工讲授统计学知识，可以采用两种不同的教学方式：一是传统的课堂讲授方法，二是一种新的交互式教学方法。把两类学员随机分成两组，对两组学员分别采用两种不同的教学方法教

学，假设最后各类学员的平均考试成绩如表5-10。对表中的结果我们能认为课堂讲授比交互式教学效果好或者在校学生比在职学生成绩好吗？都不能。我们从表5-10能够得出的最恰当的结论是：课堂讲授的方式更适合于在校学生，而交互式教学方式更适合于在职学生。在这种情况下我们说模型中的两个因素（教学方式和学生类型）之间存在着交互作用。

表5-10　平均考试成绩

	课堂讲授	交互式教学
在校学生	90	75
在职员工	75	90

多因素方差分析的比较复杂，这里我们只分析两因素的情况。双因素方差分析有两种类型：一种是无交互作用的双因素方差分析，它假定因素A和因素B的效应之间是相互独立的，不存在交互关系；另一种在存在交互作用的方差分析，我们可以分析两个因素之间的交互作用是否显著。

一、无交互作用的双因素方差分析模型

设双因素方差分析中要分析的两个因素为A、B，A因素有r个不同水平A_1，A_2，…，A_r；B因素有s个不同水平B_1，B_2，…，B_s，并假设每组试验条件的试验重复了m次，总数据量n=rsm，数据结构见表5-11。

A因素的r个水平和B因素的s个水平的组合可以形成r×s个总体。在双因素方差分析中的基本假设是，这r×s个总体中的都服从正态分布，有相同的方差，并且各个观测值之间相互独立。

表5-11 双因素方差分析的数据结构

		因素B			
		B_1	B_2	…	B_s
因素A	A_1	X_{111}，…，X_{11m}	X_{121}，…，X_{12m}	…	X_{1s1}，…，X_{1sm}
	A_2	X_{211}，…，X_{21m}	X_{221}，…，X_{22m}	…	X_{2s1}，…，X_{2sm}
	⋮	⋮	⋮	…	⋮
	A_r	X_{r11}，…，X_{r1m}	X_{r21}，…，X_{r2m}	…	X_{rs1}，…，X_{rsm}

在无交互作用的双因素方差分析模型中，因变量的取值受四个因素的影响：总体的平均值；因素A导致的差异；因素B导致的差异；以及误差项。写成模型的形式就是：

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad (5\text{-}10)$$

其中

i=1到r，代表因素A的不同水平。

j =1到s，代表因素B的不同水平。

k =1到m，代表在同一实验条件下的不同观测值。

μ = 根据所有数据计算的总均值。

α_i =因素A的第i个水平对因变量的效应。

β_j=因素B的第j个水平对因变量的效应。

ε_{ijk} = 随机误差项，$\varepsilon_{ijk} \sim N(0,\sigma^2)$。

相应的，总变异（离差平方和）可以分解为3个来源：因素A、因素B和误差因素导致的变异。根据表5-11的数据结构，定义

$$\overline{\overline{X}} = \frac{1}{rsm}\sum\sum\sum X_{ijk} \qquad (5\text{-}11)$$

$$\overline{X}_{ij} = \frac{1}{m}\sum_{k=1}^{m} X_{ijk} \qquad (i=1，2，\cdots r，j=1，2，\cdots s) \qquad (5\text{-}12)$$

$$\overline{X}_{i\bullet} = \frac{1}{r}\sum_{j=1}^{r} \overline{X}_{ij}\ ， \qquad （i=1，2，\cdots，r） \qquad (5\text{-}13)$$

$$\overline{X}_{\bullet j} = \frac{1}{s}\sum_{i=1}^{s} \overline{X}_{ij}\ ， \qquad （j=1，2，\cdots，s） \qquad (5\text{-}14)$$

则离差平方和可以进行如下分解：

$$\begin{aligned} SST &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{m}(X_{ijk}-\overline{\overline{X}})^2 \\ &= sm\sum_{i=1}^{r}(\overline{X}_{i\bullet}-\overline{\overline{X}})^2 + rm\sum_{j=1}^{s}(\overline{X}_{\bullet j}-\overline{\overline{X}})^2 \\ &\quad + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{m}(X_{ijk}-\overline{X}_{i\bullet}-\overline{X}_{\bullet j}+\overline{\overline{X}})^2 \\ &= SSA+SSB+SSE \end{aligned} \qquad (5\text{-}15)$$

其中SSA、SSB分别表示因素A和因素B不同水平间的离差平方和，SSE表示由随机因素导致的离差平方和。注意以上分解中允许m=1。

SSA、SSB、SSE的自由度分别为：r-1、s-1、n-r-s+1，三者之和等于SST的自由度n-1。相应的离差平方和除以其自由度就可以得到均方MSA、MSB和MSE，从而可以进一步利用F分布进行假设检验了。以上计算过程可用方差分析表来表示（表5-12）。

表5-12　无交互作用双因素方差分析表

变异来源	离差平方和 SS	自由度 df	均方 MS	F值
A因素	SSA	r-1	MSA=SSA/(r-1)	F_A=MSA/MSE
B因素	SSB	s-1	MSB=SSB/(s-1)	F_B=MSB/MSE
误　差	SSE	n-r-s+1	MSE=SSE/(n-r-s+1)	
合　计	SST	n-1		

二、有交互作用的双因素方差分析模型

如果因素A和因素B之间有交互作用，则因变量的取值会受五个因素的影响：总体的平均值；因素A导致的差异；因素B导致的差异；由因素A和因素B的交互作用导致的差异；误差项。写成模型的形式就是：

$$X_{ij} = \mu + \alpha_i + \beta_i + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{5-16}$$

其中$(\alpha\beta)_{ij}$=因素A的第i个水平和因素B的第j个水平对因变量的交互效应，其他符号的含义与无交互作用的情况相同。

相应的，总差异（离差平方和）可以分解为4个来源：因素A、因素B、二者的交互作用以及误差因素导致的变异：

$$\begin{aligned} SST &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{m}(X_{ijk} - \overline{\overline{X}})^2 \\ &= sm\sum_{i=1}^{r}(\overline{X}_{i\bullet} - \overline{\overline{X}})^2 + rm\sum_{j=1}^{s}(\overline{X}_{\bullet j} - \overline{\overline{X}})^2 \\ &\quad + m\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X}_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{\overline{X}})^2 \\ &\quad + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{m}(X_{ijk} - \overline{X}_{ij})^2 \\ &= SSA + SSB + SSAB + SSE \end{aligned} \tag{5-17}$$

SSA、SSB、SSAB、SSE的自由度分别为：r-1、s-1、(r-1)(s-1)、rs(m-1)。四者之和等于SST的自由度n-1。由离差平方和及其自由度可以计算出均方MSA、MSB、MSAB和MSE，从而计算出F检验值。注意要分析两个因素的交互作用，m的取值必须大于等于2。

有交互作用的双因素方差分析表如表5-13。

表5-13　有交互作用双因素方差分析表

变异来源	离差平方和 SS	自由度 df	均方 MS	F值
A因素	SSA	r-1	MSA=SSA/(r-1)	F_A=MSA/MSE
B因素	SSB	s-1	MSB=SSB/(s-1)	F_B=MSB/MSE
AB交互作用	SSAB	(r-1)(s-1)	MSAB=SSAB/(r-1)(s-1)	F_{AB}=MSAB/MSE
误　差	SSE	rs(m-1)	MSE=SSE/rs(m-1)	
合　计	SST	n-1		

三、双因素方差分析的步骤

双因素方差分析的步骤与单因素分析类似，主要包括以下步骤：

1、分析所研究数据能否满足方差分析要求的假设条件，需要的话进行必要的检验。如果假设条件不满足需要先对数据进行变换。

2、提出零假设和备择假设。双因素方差分析可以同时检验两组或三组零假设和备择假设。

（1）要说明因素A有无显著影响，就是检验如下假设：

H_0：$\alpha_1=\alpha_2=\Lambda=\alpha_r=0$；$H_1$：$\alpha_1$，$\alpha_2$,$\Lambda$，$\alpha_r$不全为0。

（2）要说明因素B有无显著影响，就是检验如下假设：

H_0：$\beta_1=\beta_2=\Lambda=\beta_s=0$；$H_1$：$\beta_1$，$\beta_2$,$\Lambda$，$\beta_s$不全为0。

（3）在有交互作用的双因素方差中，要说明因素A与第二个因素B的交互作用是否显著还需要同时检验第三组零假设和备择假设：

H_0：$(\alpha\beta)_{11}=\alpha\beta_{12}=\Lambda=(\alpha\beta)_{ss}=0$；$H_1$：$(\alpha\beta)_{11}$,$(\alpha\beta)_{12}$,$\Lambda$,$(\alpha\beta)_{rs}$不全为0。

3、计算F检验值，并根据实际值与临界值的比较，或者p值与α的比较得出检验结论。与单因素方差分析的情况类似，对F_A、F_B和F_{AB}，当F的计算值大于临界值F_α时拒绝零假设H_0。临界值F_α可以根据显著性水平α以及相应的自由度计算。

【例5.3】 在5.1的例子中，我们在研究专业对起薪影响时没有考虑其他因素的影响。有分析人员认为性别对起薪有重要影响。如果在随机调查中有些专业调查的男生多，有的专业调查的女生多，就有可能影响到不同专业的平均起薪，从而影响我们对问题的分析。为此，我们重新进行一次的调查，每个专业随机调查3名男生、3名女生。调查数据见表5-14。如果性别和专业两个因素之间没有交互作用，根据新的数据专业对起薪的影响显著吗？

表5-14 大学毕业生的专业、性别和起薪

序号	*专业*	*性别*	*起薪（元）*	*序号*	*专业*	*性别*	*起薪（元）*
1	1	0	3000	13	3	0	2000
2	1	0	3100	14	3	0	2600
3	1	0	3300	15	3	0	2500
4	1	1	4000	16	3	1	3500
5	1	1	3700	17	3	1	3000
6	1	1	3500	18	3	1	2800
7	2	0	3500	19	4	0	2200
8	2	0	3000	20	4	0	2400
9	2	0	2500	21	4	0	2000
10	2	1	4000	22	4	1	3000
11	2	1	4000	23	4	1	2000
12	2	1	3700	24	4	1	2800

同时考虑专业和性别因素时，每种实验条件下的数据只有3个，不适合直接进行正态性和等方差性检验。但根据经验可以认为起薪服从正态分布；各种实验条件下的样本容量都相等，也可以粗略的认为是等方差的。

根据题目的背景，需要采用无交互作用的双因素方差分析模型进行分析。在SPSS菜单中选择“分析”→“一般线性模型”→“单变量”，经过相应的设定后输出的方差分析表如表5-15。注意5-15的内容比标准的方差分析表（表5-12）多一些，表5-12对应的内容我们用黑体标出了。

表5-15 例5.3中SPSS无交互作用的双因素方差分析表

源	*III 型平方和*	*df*	*均方*	*F*	**Sig.**

校正模型	7528333	4	1882083.33	17.52	0.0000
截距	216600417	1	216600416.67	2016.12	0.0000
专业	4927917	3	1642638. 89	15. 29	0. 0000
性别	2600417	1	2600416. 67	24. 20	0. 0001
误 差	2041250	19	107434. 21		
总计	226170000	24			
校正的总计	9569583	23			

根据表5-15的结果，由于专业变量对应的p值（Sig.一栏）为0.0000，说明在考虑了性别因素以后各专业之间的平均起薪差异仍然是显著的。从性别对起薪的影响看，该变量对应的p值为0.0001，小于通常使用的α值，说明平均起薪的性别差异也是显著的。

【例5.4】 如果失业者在失业期间可以领取失业保险金，那么相对来说他们找工作的积极性可能会低一些，对工作岗位也会更加挑剔。为了减小失业保险制度的对就业的这种负面影响，一些学者提出了为失业者提供再就业奖励的政策建议：如果失业者可以在限定的时间内重新就业，他将可以获得一定数额的奖金。为了检验这类政策的有效性，1988年美国劳工部在华盛顿州进行了一次共有12000多名失业者参加的社会实验。在实验中奖金水平被分为无奖金（对照组）和低、中、高共四种情况。实验结果表明，提供再就业奖励的确会缩短失业时间，奖金数额越高失业时间的降低越明显。

假设我们得到了不同奖金水平的四类失业人员失业时间和年龄的数据（表5-16）。同时考虑这两个因素，根据数据分析奖金水平对失业时间的影响是否显著？年龄对失业时间有无显著影响？奖金水平和年龄因素有无交互作用？

表5-16　失业保险实验中的数据

奖金	年龄	失业时间（天）	奖金	年龄	失业时间（天）
1	1	92	3	1	96
1	1	100	3	1	92
1	1	85	3	1	90
1	2	88	3	2	77
1	2	89	3	2	79
1	2	90	3	2	71
1	3	94	3	3	82
1	3	80	3	3	75
1	3	78	3	3	81
2	1	86	4	1	78
2	1	108	4	1	75
2	1	93	4	1	76
2	2	88	4	2	87
2	2	89	4	2	73
2	2	75	4	2	83
2	3	78	4	3	82
2	3	72	4	3	68
2	3	79	4	3	72

注：奖金的取值为，1=无奖金；2=低奖金；3=中奖金；4=高奖金。

年龄的取值为，1=青年；2=中年；3=老年。

根据题目的背景，需要采用有交互作用的双因素方差分析模型。假设各组数据来自等方差的正态总体，用SPSS软件进行方差分析。在SPSS菜单中选择“分析”→“一般线性模型”→“单变量”，经过相应的设定后输出的方差分析表如表5-17。注意5-17的内容比标准的方差分析表（表5-13）多一些，表5-13对应的内容我们在表中用黑体标出了。根据表中的结果，奖金因素对应的p值等于0.0067，年龄因素对应的p值等于0.0012，都小于通常使用的α值，说明奖金水平和年龄对失业时间都有显著影响；二者交互作用的F检验值为2.12，p值等于0.0887大于0.05，说明在5%的显著性水平下奖金水平与年龄的交互作用对失业时间的影响不显著。

表5-17　例5.4中SPSS有交互作用的双因素方差分析表

源	*III 型平方和*	*df*	*均方*	*F*	*Sig.*
校正模型	1856	11	168.69	4.20	0.0016
截距	250167	1	250166.69	6223.91	0.0000
奖金	625	3	208. 32	5. 18	0. 0067
年龄	720	2	360. 11	8. 96	0. 0012
奖金 * 年龄	510	6	85. 07	2. 12	0. 0887
误差	965	24	40. 19		
总计	252987	36			
校正的总计	2820	35			

第四节　SPSS 方差分析的相关操作

一、对数据格式的要求

用 SPSS 软件做方差分析时，数据要以“列表格式”（List format）存储，而不能用其它格式。“列表格式”可用表 5-18 来说明，表 5-19 则不是“列表格式”。可能只有 Excel 中的方差分析仍然需要使用表 5-19 的格式。在表 5-18 中，每一行称为一个观测，每一列称为一个变量。

表 5-18 以列表格式的存储的数据

序号	性别	学历	工资
1	1	1	2600
2	1	1	2700
3	1	2	4100
4	1	2	4000
5	2	1	3200
6	2	1	2500
6	2	2	5300
8	2	2	5500

表 5-19 不是以列表格式的存储的数据

性别 \ 学历	1（本科）	2（研究生）
1（女）	1600	4100
	1700	4000
2（男）	3200	5300
	2500	5500

二、用 SPSS 检验数据分布的正态性

我们使用例5.2的林木采伐数据进行说明用SPSS检验数据分布的正态性的相关操作。

在SPSS中打开Excel数据文件，选择“分析”→“描述统计”→“探索”，在“探索”对话框中将树木数量作为因变量，兴趣作为因子；单击“绘制”按钮，选中“直方图”复选框和“带检验的正态图”，单击“继续”按钮，在单击主对话框中的“确定”，可以得到分类别的描述统计信息（图5-4）。从数据的茎叶图、直方图和箱线图都可以对数据分布的正态性做出判断。

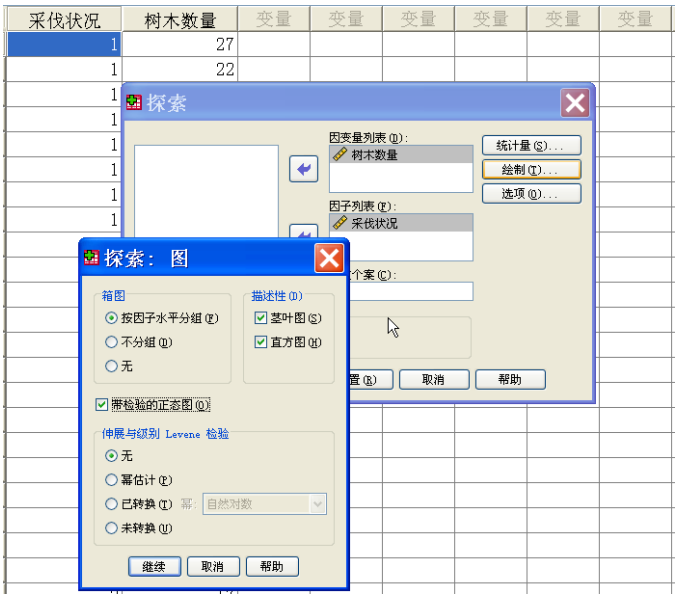


图5-4 用“探索”过程进行正态性检验

输出结果中的Q-Q图是观察数据分布正态性的一种常用图形。这类图形大致是这样绘制的：计算数据在样本中对应的经验分布函数值（类似于累积分布的函数值，取值在0-1之间）；然后计算标准正态分布（或者均值、方差相同的正态分布）对应于经验分布函数值的分位数。以实际值为横坐标，正态分布的分位数为纵坐标作散点图，如果图形中的点大致在一条直线上则说明数据服从正态分布。图5-5是从未采伐过的林地上树木数量的Q-Q图，从图中可以判断数据并没有严重背离正态分布。

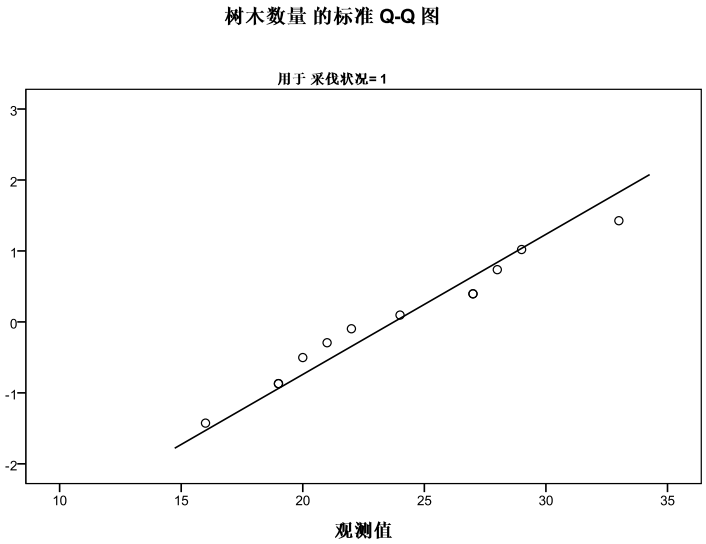


图5-5 Q-Q图

表7-8是对数据进行正态性检验的结果。SPSS中采用的是Kolmogorov-Smirnov检验和Shapiro-Wilk检验。这两种检验方法都属于非参数统计的内容，统计量的计算方法可以参考有关书籍。我们可以根据软件给出的p值对数据是否服从正态分布进行检验：由于表5-20中的p-值都大于0.05，因而我们不能拒绝零假设，也就是说没有证据表明各组的数据不服从正态分布（检验中的零假设是数据服从正态分布）。

表5-20　正态性检验的结果

	采伐状况	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
树木数量	1	.156	12	.200*	.961	12	.799
	2	.171	12	.200*	.902	12	.168
	3	.206	9	.200*	.906	9	.289

a. Lilliefors 显著水平修正

*. 这是真实显著水平的下限。

三、用 SPSS 进行单因素方差分析和多重比较

SPSS的单因素ANOVA过程可以进行单因素方差分析和均值的多重比较。

在例5.2中，单击“分析”→“比较均值”→“单因素ANOVA”，在对话框中将变量“树木数量”选入因变量列表，将变量“采伐情况“移入因子栏。单击对话框中的“选项”按钮，在弹出的对话框中选中“描述统计”、“方差同质性检验”复选框（图5-6）。单击主对话框中的“两两比较”，选中“LSD”复选框。单击主对话框中的“确定”按钮，就可以得到相应的分析结果了。

相关输出结果的分析见第二节的分析。

图5-6 单因素方差分析的对话框和选项设定

四、用 SPSS 进行双因素方差分析

1. 无交互作用的双因素方差分析。

SPSS的“一般线性模型”中的“单变量”过程可以用来进行单因素或多因素方差分析，检验不同因素以及因素之间的交互作用对均值的影响是否显著。

我们以例5-3的相关操作为例说明用SPSS进行无交互作用双因素方差分析的过程。在SPSS中读入Excel数据表，选择“分析”→“一般线性模型”→“单变量”，在主对话框中将“起薪”放入因变量矩形框，将“专业”和“性别”放入固定因子矩形框中（图5-7）。

图5-7 单变量一般线性模型主对话框

在主对话框中点击“模型”按钮进入“模型”对话框，选择“设定”，把“专业”和“性别”变量选入右边的模型框中，单击“继续”返回主对话框（图5-8）。其它选项采用默认值，单击主对话框中的“确定”按钮，可以得到无交互作用的双因素方差分析结果，相关结

果的分析见第三节。

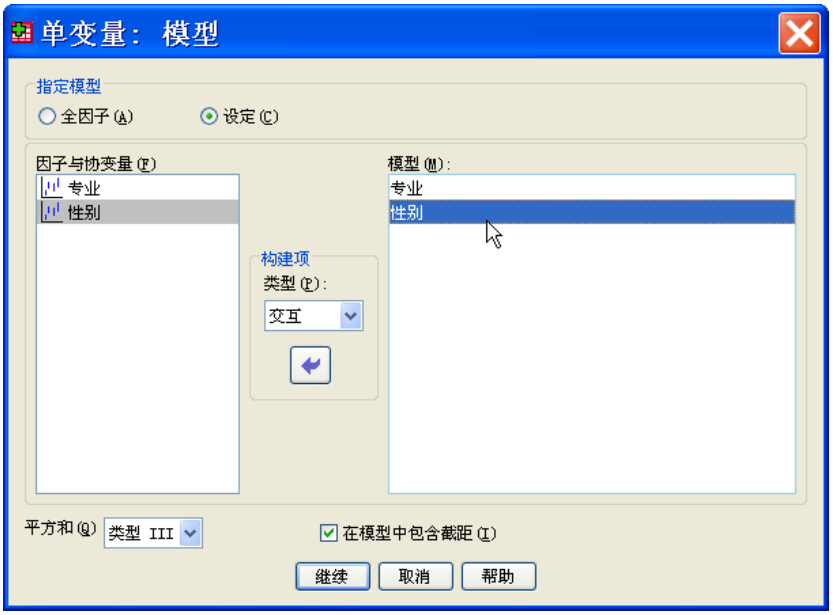


图5-8　单变量一般线性模型的模型定义对话框（不考虑交互作用）

2. 有交互作用的双因素方差分析。

SPSS的“一般线性模型”中的“单变量”过程可以用来进行单因素或多因素方差分析，检验不同因素以及因素之间的交互作用对均值的影响是否显著。

我们以例5.4的相关操作为例说明用SPSS进行有交互作用双因素方差分析的过程。在SPSS中读入Excel数据表，选择“分析”→“一般线性模型”→“单变量”，在主对话框中将“失业时间”放入因变量矩形框，将“奖金”和“年龄”放入固定因子矩形框中（图5-9）。单击“确定”，即可得到需要的方差分析表。单变量一般线性模型设定中默认的模型就是包含交互作用的。关于输出结果的分析见第三节。

图5-9　单变量一般线性模型的模型定义对话框（考虑交互作用）

在双因素方差分析中也可以输出基本的描述统计指标、绘制相关图形、进行方差齐性检验、多重比较等，具体操作都可通过相应的选项实现，这里不再展开。

小 结

方差分析（ANOVA），一般用来分析一个定量因变量与一个或几个定性自变量（因素）之间的关系，它可以对多个总体的均值是否相等进行整体检验。根据研究所涉及的因素的多少，方差分析可分为单因素方差分析和多因素方差分析（包括双因素方差分析）。单因素方差分析可以检验一个因素的作用是否显著，双因素方差分析则可以同时检验两个因素以及这两个因素的交互作用是否显著。

方差分析中的基本假设是，来自各个总体的数据都服从正态分布，相互独立，且有相同的方差。在进行方差分析之前，需要检查方差分析的假设条件是否成立。

方差分析的基本思想是，将观察值之间的总离差平方和分解为由所研究的因素引起的总离差平方和和由随机误差项引起的总离差平方和，然后通过对这两类总离差平方和的比较，做出接受或拒绝原假设的判断的。方差分析的主要步骤包括：建立假设；计算F检验值；根据实际值与临界值的比较做出决策。F检验值的计算过程一般通过方差分析表来描述，实际的计算过程可通过计算机完成的。

在方差分析中，当拒绝H_0时表示至少有两个均值有显著差异。但要知道哪些均值之间有显著差异还需要借助于多重比较的方法，例如LSD方法。

思考与练习

1. 试述方差分析的基本思想。

2. 方差分析有哪些基本假设条件？如何检验这些假设条件？

3. 对三个不同专业的学生的统计学成绩进行比较研究，每个专业随机抽取6人。根据数据得到的方差分析表的部分内容如表5-21。请完成该表格。如果显著性水平$\alpha=0.05$，能认为三个专业的考试成绩有显著差异吗？

表5-21 不同专业考试成绩的方差分析表

差异源	SS	df	MS	F

组间	193.0	_____	_____	_____
组内	819.5	_____	_____	
总计	1012.5	_____		

根据以下背景资料和数据回答4-7题。

为测试A、B、C、D、E五种节食方案，一位营养学家选择了50名志愿者随机分成五组，每组采用一种方案测量两个月后每个人的降低的体重，得到的实验数据如表5-22。

表5-22　不同节食方案的降低的体重（公斤）

序号	方案A	方案B	方案C	方案D	方案E
1	6.5	2.9	8	5.1	11.5
2	11.6	5.5	11.9	2.5	13.2
3	7.7	4.3	8.5	1.5	11
4	8.7	3.6	8.9	2.2	13.1
5	8.4	3.9	9.1	1.4	13.8
6	4.1	6.7	11.4	3.1	12.8
7	8.7	4.5	12.6	5.4	12
8	6.6	1.7	12.4	1.9	11.5
9	7.1	6.5	9.4	4.1	14.6
10	8.9	5.4	10.6	3.6	13.7

4. 不同节食方案的实验效果的描述统计资料如表5-23。根据这些结果可以认为各组数据是等方差的吗？

5-23　不同节食方案的实验效果的描述统计

组	计数	平均	方差
A	10	7.83	3.88
B	10	4.50	2.47
C	10	10.28	2.92
D	10	3.08	2.07
E	10	12.72	1.39

5. 利用表5-22的数据进行单因素方差分析，则何为H_0与H_1？

6. 用SPSS软件对表5-22的数据进行方差分析，设显著性水平α=0.05，检验的结论如何？提示：你需要把数据调整为SPSS需要的形式。

7. 在上题中我们拒绝了零假设。试在同样的显著性水平α=0.05的条件下，用SPSS软件中的LSD法检验方案B与其他4种方案差异的显著性。

8. 五个地区每天发生的交通事故次数如表5-24。由于是随机样本，有些地区的样本容量较大，而有些地区的样本容量较小。试以α＝0.05的显著性水平检验这五个地区平均每天的交通事故次数是否相等。

表5-24 五个地区的交通事故次数

序号	东部	北部	中部	南部	西部
1	15	12	20	14	13
2	17	10	24	9	12
3	14	13	13	7	9
4	11	17	15	10	14
5		14	12	8	10
6				7	9

9. 在α＝0.01的显著性水平下求解回答第8题中的问题。

10.在一次无重复双因素方差分析实验中，因素A有4个水平，因素B有6个水平。得到的方差分析表的部分内容如表5-25。完成该表格。在5%的显著性水平下因素A的影响显著吗？因素B呢？

表5-25 无重复双因素方差分析的方差分析表

差异源	SS	df	MS	F
因素 A	77.50	_____	_____	_____
因素 B	123.83	_____	_____	_____
误差	92.50	_____	_____	
总计	293.83	_____		

11.为了比较三个不同专业毕业生的收入设计了以下实验：从三个专业毕业两年的毕业生中，按毕业时的平均学习成绩（分为6个等级）各选择一名学生进行调查，调查结果如表5-26。显著性水平α=0.01。试分析不同专业毕业生的收入有显著差异吗？学习成绩对收入有显著影响吗？

表5-26不同专业毕业生的月收入（百元）

平均成绩	会计	营销	经济
A+	51	45	41
A	45	38	36
B+	31	33	27
B	35	29	32
C+	32	31	26
C	27	25	23

12.在一次双因素方差分析实验中，因素A和因素B各有3个水平，在每种实验条件下重复进行了2次实验。得到的方差分析表的部分内容如表5-27。完成该表格。在5%的显著性水平下因素A的影响显著吗？因素B呢？二者的交互作用是否显著？

表5-27 不同专业考试成绩的方差分析表

差异源	SS	df	MS	F
因素 A	6100	_____	_____	_____
因素 B	45300	_____	_____	_____
AB 的交互作用	11200	_____	_____	_____
误差	19850	_____	_____	

总计	82450	_____

13. 人们通常认为，一个人的学历越高其工资也会越高。同时，性别也可能是影响工资水平的一个因素。为了对两个因素进行研究，一名研究人员从毕业两年的学校毕业生中根据学历和性别随机选择了30人，他们的工资水平、学历和性别资料如表5-28。假设数据是正态的和等方差的。在α=0.05的显著性水平下，试回答：（1）学历对工资的影响显著吗？（2）性别对工资的影响显著吗？（3）学历和性别的交互作用显著吗？

表5-28　30名学校毕业生的月工资（元）

序号	性别	学历	工资	序号	性别	学历	工资
1	1	1	1600	16	2	1	1200
2	1	1	1700	17	2	1	1300
3	1	1	1700	18	2	1	1400
4	1	1	2100	19	2	1	2000
5	1	1	2200	20	2	1	2200
6	1	2	2800	21	2	2	2600
7	1	2	3700	22	2	2	2700
8	1	2	3900	23	2	2	2900
9	1	2	4000	24	2	2	3100
10	1	2	4100	25	2	2	3500
11	1	3	4000	26	2	3	3200
12	1	3	4300	27	2	3	3500
13	1	3	4700	28	2	3	3600
14	1	3	5300	29	2	3	3800
15	1	3	5500	30	2	3	3900

性别：1=男性，2=女性

学历：1=中专 2=本科　3=研究生

第六章 非参数检验

在前面的章节中我们介绍了多种假设检验的方法，例如单个总体的t检验、基于两个独立样本的t检验、基于两个匹配样本的t检验、方差分析等。在这些检验都需要对总体的分布特征作出某些假设（例如在t检验和方差分析中都需要假设总体服从正态分布），然后根据检验统计量的抽样分布对总体参数（如均值、比率等）进行检验。这类检验方法称为参数检验。我们前面强调过，在需要的假设条件不满足的情况下，特别是小样本的情况下，t检验、F检验都是不适用的。

那么，如何检验数据是否来自正态分布或者其他分布？在参数检验假设条件不满足的情况下如何对相应的问题进行分析？非参数检验方法可以帮助我们回答这类问题。在这一章中，我们将首先简要说明非参数检验的概念和优缺点，然后介绍几种常见的非参数检验方法及其在SPSS中的实现方法。

第一节 非参数检验概述

非参数检验（nonparametric tests）也称为与总体分布无关的检验（distribution free tests），与参数检验相比，在非参数检验中不需要对总体分布的具体形式作出严格假设，或者只需要很弱的假设。大部分非参数检验都是针对总体的分布进行的检验，但也可以对总体的某些参数进行检验。

与参数检验相比，非参数检验主要有以下几个方面的特点：

（1）非参数检验不需要严格假设条件，因而比参数检验有更广泛的适用面。

（2）非参数检验几乎可以处理包括定类数据和定序数据在内的所有类型的数据，而参数检验通常只能用于定量数据的分析。

（3）虽然对于满足参数检验的假设条件的数据也可以采用非参数检验法进行分析，但在参数检验和非参数检验都可以使用的情况下，由于非参数检验没有充分利用样本内所有的数量信息，因此其检验的功效（power）要低于参数检验方法。也就是说，在备择假设为真的情况下，采用参数检验方法拒绝原假设的概率要高于非参数检验的方法，从而更容易发现显著的差异。

在假设检验中，犯取伪错误的概率记为β，则1-β越大，意味着当备择假设为真时，拒绝原假设的概率越大，检验的判别能力就越好；1-β越小，意味着当备择假设为真时，拒绝原假设的概率越小，检验的判别能力就越差。可见1-β是反映统计检验判别能力大小的重要指标，我们称之为检验功效或检验力。

根据非参数检验的以上特点，在以下情况下应当首选非参数方法进行统计推断：

（1）参数检验中的假设条件不满足，从而无法应用。例如总体分布为偏态或分布形式未知，且样本为小样本时。

（2）检验中涉及的数据为定类或定序数据。

（3）所涉及的问题中并不包含参数，如判断某样本是否为随机样本，判断某样本是否来自正态分布等。

（4）对各种资料的初步分析。

非参数检验的方法很多，本章的介绍的非参数检验方法包括：用于单个样本的χ^2拟合优度检验、K-S拟合优度检验、中位数的符号检验，用于两个匹配样本的符号检验、Wilcoxon

符号秩检验，用于两个独立样本的Wlicoxon秩和检验，以及用于多个独立样本的Kruskal-Wallis检验。

第二节　单样本的非参数检验方法

在推断统计中，我们往往对样本所属总体的分布形态感兴趣，希望判断样本是否来自于已知的理论分布。例如，在 t 检验和方差分析中，我们往往需要判断总体的正态性。虽然我们可以通过一些人们可以通过绘制直方图、计算相应的描述统计指标来作粗略判断，但如果希望得到比较准确的结论，则需要用非参数检验。下面介绍的χ^2检验和 *K-S* 检验都可以帮助我们检验能否认为样本数据来自某种概率分布。本节还介绍了单样本中位数的符号检验方法。

一、χ^2 拟合优度检验

我们通过一个例子来说明 χ^2 拟合优度检验的基本原理。

【例 6.1】一种饮料的容器材料可以选择玻璃、塑料或者金属。为了料较消费者对包装材料的偏好，对 120 名消费者进行了抽样调查，发现最喜欢玻璃、塑料和金属容器的分别有 55、25 和 40 人。根据调查结果，能否认为消费者对 3 种材料的偏好程度是无差异的？

在这个例子中，如果消费者对 3 种材料的偏好程度是无差异的，也就是说消费者对材料的偏好服从均匀分布，则理论上来说，调查 120 名消费者，偏好每种材料的人数应该是相等的，也就是 40 人。各组观测到的人数与理论人数（期望值）之间的差异应该都是由于抽样的随机性造成的，因此不应该太大。如果二者之间的差异特别大，则说明我们所作的假设（消费者对 3 种材料的偏好程度是无差异的）很可能不成立。

在进行检验时需要按照式（6-1）构造的χ^2统计量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (6\text{-}1)$$

式中：k 是样本分类的个数，O_i 表示实际观察到的频数，E_i 表示理论频数。观察频数与期望频数越接近，则χ^2值越小。根据皮尔逊定理，当 n 充分大时，χ^2统计量渐近服从于 k-1 个自由度的χ^2分布。我们可以计算出χ^2统计量的值，然后根据其与显著性水平α下的临界值的大小关系得出检验结论。在统计软件中一般根据χ^2统计量的值给出 p 值，从而可以根据 p 值与α的大小关系得出检验结论。

由于奠定检验基础的皮尔逊定理要求样本充分大，所以在搜集资料时要求是大样本，同时每个单元中的期望频数不能太小，如果有的组中频数小于 5，则需要将它与相邻的组进行合并。

下面我们用 SPSS 软件来分析例 6.1 的数据。

在 SPSS 软件中按照图 6-1 的格式录入数据，然后选择“数据”→“加权个案”，在弹出的对话框中按照图 6-1 把频数变量指定为权数。如果分析的是未加整理原始数据，则不需要指定权数变量。

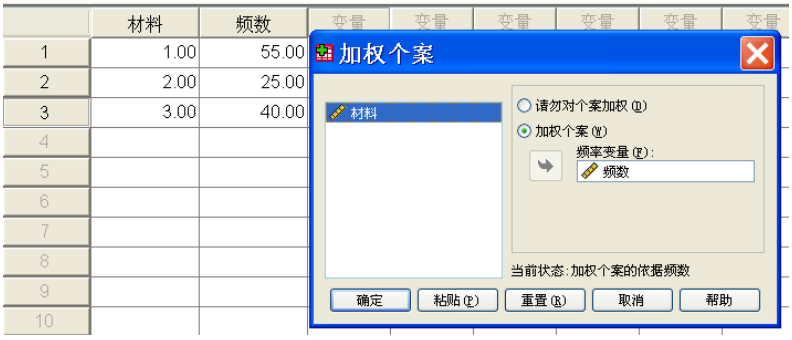


图 6-1 例 6.1 中的数据录入和权数设定

然后，选择“分析”→“非参数检验”→“卡方”，在弹出的对话框中将“材料”设定为检验变量；单击对话框中的“精确…”，选中弹出对话框中的“精确”，单击“继续”、“确定”，得到的分析结果见表 6-1、6-2。

图 6-2 卡方检验的相关设定

表6-1 例6.1中各组的频数和期望频数

	观察数	期望数	残差
1.00	55	40.0	15.0
2.00	25	40.0	-15.0
3.00	40	40.0	.0
总数	120		

表 6-1 中是各组的频数、期望频数以及二者的差。表 6-2 中是统计量的计算结果和相应的 p 值。根据表 6-2，计算出的χ^2统计量的值为 11.250，自由度为 2，相应的 p 值（渐近显著性）为 0.004，小于通常使用的α值。所以检验的结论是拒绝总体中消费者对 3 种材料的偏好程度无差异的零假设。

在这个例子中我们要求软件输出了精确检验的 p 值（精确显著性），结果为 0.003。这个结果是根据统计量的真实分布或者采用重抽样方法得到的，而不是根据渐进分布得到的，因而在小样本时也可以使用。在χ^2检验中，如果样本量比较小，或者有的组频数小于 5，应该按照精确方法计算得到的 p 值得出结论。在 SAS 等软件中也可以计算精确检验的 p 值。采用精确检验的方法可以得到更可靠的 p 值，但不足之处是比较耗费计算机资源，特别是样本

量很大时。

表6-2　例6.1的计算结果和相应的p值

	材料
卡方	11.250
df	2
渐近显著性	.004
精确显著性	.003
点概率	.000

χ^2检验也可以按照同样的思想对正态分布或者任何想象的其他分布进行检验，但主要用于对定性变量的检验。

二、单样本 K-S 检验

单样本 K-S 检验是以两位苏联数学家 Kolmogorov 和 Smirnov 命名的。K-S 检验通过对两个分布差异的分析确定能否认为样本的观察值来自所设定的理论分布总体。

设 $F_n(x)$ 是一个样本量为 n 的随机样本的累积概率分布函数，即经验分布函数； $F(x)$ 是一个特定的累积概率分布函数，即理论分布函数。定义 $D=\left|F_n(x)-F\ (x)\right|$，显然若对每一个 x 值来说，如果 $F_n(x)$ 与 $F(x)$ 十分接近，则表明经验分布函数与特定分布函数的拟合程度很高，有理由认为样本数据来自具有该理论分布的总体。K-S 检验中的检验统计量如式（6-2）所示：

$$D_{\max}=\max\left|F_n(x)-F(x)\right| \qquad (6\text{-}2)$$

根据检验统计量 $D_{\max}$ 的精确分布和渐进分布，我们可以计算出假设检验的 p 值，从而得出检验的结论。

下面通过一个例子来看一下 K-S 检验在 SPSS 软件中的操作和结果分析。

【例 6.2】例 4.1 中有 100 名儿童每周看电视时间的数据（数据文件：看电视时间.sav）。检验能否认为总体中儿童每周看电视的时间服从正态分布。

这里 K-S 检验的零假设和备择假设为：

H_0：总体中儿童每周看电视的时间服从正态分布。

H_1：总体中儿童每周看电视的时间不服从正态分布。

在 SPSS 软件中打开数据文件，选择“分析”→“非参数检验”→“1 样本 K-S”，在弹出的对话框中将“时间”设定为检验变量；检验分布为默认的“常规”（正态分布）。单击 “确定”（图 6-3），得到的分析结果见表 6-3。

图 6-2 K-S 检验的相关设定

表6-3　单样本 K-S检验的计算结果和相应的p值

		时间
N		100
正态参数[a,,b]	均值	27.191
	标准差	8.3728
最极端差别	绝对值	.096
	正	.096
	负	-.039
Kolmogorov-Smirnov Z		.960
渐近显著性(双侧)		.315

a. 检验分布为正态分布。

b. 根据数据计算得到。

根据表 6-3，计算出的 D_{max} 统计量的值为 0.960（表中的 Kolmogorov-Smirnov Z），相应的 p 值（渐近显著性）为 0.315。由于 0.315 大于 0.05，所以在 5%的显著性水平下不能拒绝原假设，也就是说根据样本数据不能认为总体数据是非正态的。

注意这里并不能得出总体服从正态分布的严格结论。总体服从正态分布的结论可能犯第二类错误（取伪错误），这个概率是未知的，在有些情况下可能会很大。

在 K-S 检验中如果使用的是小样本，则根据渐进分布计算 p 值的误差会增大。这时应该通过相应的设定要求软件输出精确检验的 p 值，根据精确检验的 p 值得出检验结论。

三、单样本中位数的符号检验

在前面我们讲过的单样本t检验是对总体均值的假设检验。在数据呈偏态分布的情况下，我们可能对总体的中位数更感兴趣，希望对总体的中位数作出推断，这时可以使用符号检验（sign test）的方法。在非正态总体小样本的情况下，如果要对总体分布的位置进行推断，由于 t 检验不适用，也可使用符号检验的方法。

我们通过一个例子说明符号检验的基本思想。

【例6.3】在某地区随机调查了60个家庭的月收入（数据文件：家庭月收入.sav）。能否认为总体的家庭月收入的中位数等于5000元？

检验的零假设和备择假设如下：$H_0: M_e = 5000 \leftrightarrow H_1: M_e \neq 5000$。

样本中的每个数据都减去零假设中的中位数 5000，记录其差值的符号。用 S_+ 和 S_- 别是正、负符号的个数（差值为 0 的不计算在任何一个中），当原假设为真是时，S_+ 和 S_- 应该很接近；若两者相差太远，就有理由拒绝原假设。定义 $S_+ + S_- = n'$。定义检验统计量 $S = \min(S_+, S_-)$，则在原假设成立的条件下，检验统计量 S 服从二项分布 $B(n', 0.5)$。用 s 表示 S 统计量的样本观测值，则检验中的 p 值可以按照以下公式计算：p 值 $= 2P(S \leq s)$。这个概率可以根据二项分布 $B(n', 0.5)$ 计算得到，从而得出检验的结论。

当 $n' > 25$ 时，$Z = \dfrac{(S + 0.5) - n'/2}{\sqrt{n'}/2}$ 近似服从标准正态分布，可以按照正态分布进行近似计算。

下面我们来看在 SPSS 中对中位数进行检验的相关操作和结果分析。

在 SPSS 中打开数据文件。为了对中位数进行检验，先在 SPSS 中生成一个显得变量 Median，取值为 5000：单击“转换”→“计算变量”，在弹出的对话框中按照图 6-3 进行设置，单击确定。

图 6-3　在 SPSS 中计算新的变量

然后，选择“分析”→“非参数检验”→“2 个相关样本”，在弹出的对话框中将“Median”和“家庭月收入”设定检验的一对变量；选中“符号检验”，取消选择“Wilcoxon”，单击 “确定”（图 6-4），得到的分析结果见表 6-4、6-5。

表 6-4 是各个数值与中位数比较的结果，有 29 个值小于 5000，31 个值大于 5000。表 6-5 表明，用正态分布进行近似计算时，Z 统计量的值为-0.129，双侧检验的 p 值为 0.897。显然，这里我们的结论是不能拒绝原假设。

在对中位数的符号检验中也可以进行单侧检验。根据表 6-4，样本中位数要略大于 5000。因此样本数据不可能支持样本中位数小于 5000 的备择假设。因此，对于零假设和备择假设 $H_0: M_e \geq 5000 \leftrightarrow H_1: M_e < 5000$，检验的 p 值=1-双侧检验 p 值/2=1-0.4485=0.5515。由于样本数据有可能支持样本中位数大于 5000 的备择假设，因此对于零假设和备择假设 $H_0: M_e \leq 5000 \leftrightarrow H_1: M_e > 5000$，假设检验的 p 值=双侧检验 p 值/2=0.4485。根据以上计算的 p 值，样本数据不支持样本中位数大于 5000 的备择假设，或者小于 5000 的备择假设。

图 6-4　在 SPSS 中进行符号检验的相关设置

表6-4　各个数值与中位数比较结果的汇总表

		N
家庭月收入 - Median	负差分[a]	29
	正差分[b]	31
	结[c]	0
	总数	60

a. 家庭月收入 < Median

b. 家庭月收入 > Median

c. 家庭月收入 = Median

表6-5 例6.3中检验统计量的值和p值

	家庭月收入 - Median
Z	-.129
渐近显著性(双侧)	.897

在这个例子中，如果使用 t 检验对总体的均值做右侧检验，$H_0:\mu\le 5000 \leftrightarrow H_1:\mu>5000$，SPSS 计算的 t 统计量为 2.52，双侧检验的 p 值为 0.014，从而右侧检验的 p 值为 0.007，在1%的显著性水平下可以拒绝原假设。可见，总体的均值要大于 5000，而没有证据表明总体的中位数大于 5000。从中位数和均值的关系看，总体的分布属于典型的右偏分布。

与前面提到的非参数检验方法类似，在使用符号检验时，如果样本量较小，则需要使用软件输出的精确检验的 p 值进行推断，使用正态分布得到的结果不可靠。在小样本时，如果要求进行精确检验，SPSS 会自动按照二项分布进行概率计算，请读者自行验证。

第三节　两个样本和多个样本的非参数检验

在参数检验部分我们介绍过两个匹配样本的 t 检验、两个独立样本的 t 检验和单因素方差分析等方法。这些方法在所需的假设条件不满足时不适用的。在这种情况下可以采用相应的非参数检验方法来对总体进行统计推断。需要注意的是参数检验和非参数检验中检验的零假设和备择假设是不同的，使用中应注意这种区别。

一、匹配样本的非参数检验

对于匹配样本，假设得到的样本数据为$(x_1,y_1),(x_2,y_2),\Lambda,(x_n,y_n)$，用对应的数据相减得到新的序列$z_i=x_i-y_i$（应用中哪个序列在前不影响检验结论），然后根据序列z_i进行检验。如果序列z_i服从正态分布就是一个典型的匹配样本的t检验问题。

如果t检验的假设条件不满足，t检验就不适用了。符号检验和Wilcoxon符号秩检验都可以用做替代的检验方法。

在匹配样本的非参数检验中，双侧检验时的零假设和备择假设为如下：

H_0：差值总体的中位数=0；H_1：差值总体的中位数≠0。

当然，如果需要也可以进行单侧检验。

1. 符号检验

前面讲过的符号检验也可以用于匹配样本的情况。对于序列z_i，计算其中正数的个数S_+和负数的个数S_-，然后就可以按照符号检验的方法进行假设检验了。

2. Wilcoxon符号秩检验

假设序列z_i为连续的对称分布，Wilcoxon符号秩检验的步骤如下：

（1）设定零假设和备择假设。

（2）计算序列$z_i=x_i-y_i$，并计算差值绝对值的秩。

将$|z_i|$从小到大排序，其位次就是$|z_i|$的秩（rank），等于 0 值不参与排序。在排序时，如果数据中有相同的数值则称为结（tie）。结中数字的秩为排序后它们所占位置的平均值。

（3）分别计算出差值序列中正数的秩和W_+以及负数的秩和W_-。

（4）根据W_+和W_-建立检验统计量，计算 p 值并得出检验的结论。

在双侧检验中检验统计量可以取为$W=\min(W_+,W_-)$。显然，如果零假设成立，W_+与 W_-应该比较接近。如果二者过大或过小，则说明零假设不成立。

有了检验统计量 W，我们就可根据其统计分布计算 p 值了，双侧检验的 p 值等于 $2\,\mathrm{P}(W\le w)$，式中 w 为检验统计量的样本观测值。W 的统计分布比较复杂，小样本时应使用精确检验的结果或者查专门的表格，大样本（n>50）时可以按照正态分布进行近似计算。

左侧检验时（备择假设为差值总体的中位数<0）检验统计量可以取为$W=W_+$，右侧检验时（备择假设为差值总体的中位数>0）检验统计量可以取为$W=W_-$。这时检验中的 p 值等于 $\mathrm{P}(W\le w)$。

由于 W_+与 W_-的对称性（$\mathrm{W}_+ + \mathrm{W}_- = n(n+1)/2$），检验统计量的取法可以不同于以上设置。

符号检验在匹配数据分析应用中只用到差值的符号，而对差值数值的大小未能考虑，因而失去了部分信息。Wilcoxon符号秩检验既考虑差值的符号，又考虑差值的大小，因此在所需的假设条件满足时其功效比符号检验高。

显然，Wilcoxon符号秩检验也可以用于单样本中位数的非参数检验，这时只需要将第二个样本的值设为零假设中的数值即可。

下面我们通过一个例子来说明符号检验和Wilcoxon符号秩检验的软件操作和思想。

【例6.4】适时管理（JIT，Just In Time）是为适应消费需要变得多样化、个性化而建立的一种生产体系及为此生产体系服务的物流体系。从实施适时管理（JIT）的企业中随机抽取10家进行效益分析，得到它们在实施JIT前后三年的平均资产报酬率如表6-6（数据文件：JIT管理.sav）。在5％的显著性水平下企业在实施JIT前后的资产报酬率是否有显著差异？

表6-6　10家企业实施JIT前后的资产报酬率（%）和秩

x_i（实施 JIT 前）	*15.8*	*14.9*	*15.2*	*15.8*	*15.0*	*14.6*	*14*	*14.9*	*15.1*	**15.4**
y_i（实施 JIT 后）	14.6	15.5	15.5	14.7	15.2	14.8	14.8	14.6	15.3	15.5
$z_i = x_i - y_i$	+1.2	-0.6	-0.3	+1.1	-0.2	-0.2	-0.8	+0.3	-0.2	-0.1
$\|z_i\|$的秩	10	7	5.5	9	3	3	8	5.5	3	1

如果序列 z_i 服从正态分布，这个问题可以用两个匹配样本的 t 检验来进行分析。个例子

我们用表 6-6 中来说明秩的计算方法。按照$|z_i|$，最小的值为 0.1，因此其秩为 1，最大值为 1.2，其秩为 10。有 3 个$|z_i|$为 0.2，在排序时占据的位置为 2、3、4，因此这三个数的秩都等于（2+3+4）/3=3。容易计算这里 W_+=24.5，W_-=30.5。

下面我们用SPSS软件来分析这一问题。根据题目内容，检验的零假设和备择假设设置如下：

H_0：差值总体的中位数=0；　H_1：差值总体的中位数≠0。

在 SPSS 软件中打开数据文件，选择“分析”→“非参数检验”→“2 个相关样本”，在弹出的对话框中将“JIT 后”和“JIT 前”设定检验的一对变量；选中“Wilcoxon”和“符号检验”。由于这里样本量很小，我们要求进行精确检验：单击对话框中的“精确…”，选中弹出对话框中的“精确”，单击“继续”、“确定”（相关操作参见图 6-5），得到的分析结果见表 6-7、6-8、6-9 和 6-10。

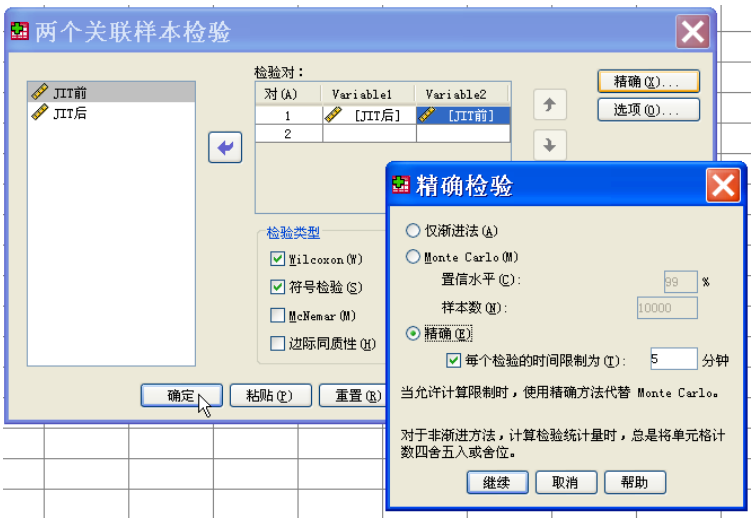


图 6-5　在 SPSS 中进行 Wilcoxon 符号秩检验和符号检验的相关设置

表6-7是Wilcoxon符号秩检验中关于秩和的计算结果，与我们前面手工计算的结果一致。表6-8中是采用渐进分布（正态分布）的Z值（-0.307）、p值（0.759），以及精确检验的p值（0.787）。由于这里样本量较小，应该采用精确检验的结果。由于0.787远远大于0.05，显然不能拒绝原假设，也就是说没有证据表明小于企业在实施JIT前后的资产报酬率有显著变化。

表6-9和6-10是符号检验的结果。表6-9表明有差值序列中有7个负数，3个正数；表6-10表明采用精确检验（二项分布）计算的双侧检验的p值为0.344，也不能拒绝原假设。

Wilcoxon符号秩检验的p值（0.787）比符号检验p值（0.344）高出不少，其主要原因在于：尽管 $x_i - y_i (i = 1, \Lambda, n)$ 中正数的个数与负数的个数相差较为悬殊，但由于正数的符号秩较大，在一定程度上冲减了两者的差异。

表6-7　Wilcoxon符号秩检验中秩和的计算结果

		N	秩均值	秩和
JIT前 - JIT后	负秩	7[a]	4.36	30.50
	正秩	3[b]	8.17	24.50
	结	0[c]		
	总数	10		

a. JIT前 < JIT后

b. JIT前 > JIT后

c. JIT前 = JIT后

表6-8　Wilcoxon符号秩检验的p值

	JIT前 - JIT后
Z	-.307[a]
渐近显著性(双侧)	.759
精确显著性（双侧）	.787
精确显著性（单侧）	.394
点概率	.020

a. 基于正秩。

表6-9　差值序列中的正数和负数的个数汇总表

		N
JIT前 - JIT后	负差分[a]	7
	正差分[b]	3
	结[c]	0
	总数	10

a. JIT前 < JIT后

b. JIT前 > JIT后

c. JIT前 = JIT后

表6-10　匹配样本符号检验的检验结果

	JIT前 - JIT后
精确显著性（双侧）	.344[a]
精确显著性（单侧）	.172
点概率	.117

a. 已使用的二项式分布。

二、两个独立样本的 Wlicoxon 秩和检验

在两个独立样本的t检验不适用时，Wlicoxon秩和检验可以作为一种替代的非参数检验方法使用。这一检验可以用来对两个总体的中位数进行检验。

假设两个总体X和Y的中位数为M_x和M_y，(x_1, x_2, Λ, x_m)和(y_1, y_2, Λ, y_n)是来自两个总体的随机样本。如果两个总体具有相似的分布形状，并且中位数相同，那么由m个x_i、n个y_i组成的m十n＝N个观察值可以被看作来自同一总体的一个随机样本。将全部x_i和y_i从小到大排序确定每个数值的秩，然后计算m个x_i的秩的和W_x、n个y_i的秩的和W_y。由于抽样的随机性，x_i、y_i应较均匀地分布在混合排列的样本中。因此，如果零假设成立，在样本量相同的情况下W_x和W_y应该比较接近，样本量不同的情况下x_i平均秩和y_i的平均秩应该比较接近。否则就说明两个总体的中位数是不相等的。

由于W_x和W_y的对称性，二者都可以用作Wilcoxon秩和检验的检验统计量W。SPSS软件中使用的是平均秩较小的一组的秩和。

确定了检验统计量，就可以根据统计量的分布计算p值了。统计量W的统计分布可以精确计算出来，并且在样本量较大时（m和n都不小于10）可以用正态分布来进行近似。得到p值之后再通过比较p值和α的大小得出结论。

由于Wilcoxon秩和检验与Mann和Whitney提出的U检验完全等价，因此这种方法也被称为Wlicoxon-Mann-Whitney检验，或者Mann-Whitney U检验。

【例6.5】 在对某企业职工的收入调查中20名本科毕业生和15名研究生的月收入（元）如下（数据文件：本科研究生收入.sav）：

本科生：1410，920，1160，860，1370，970，1350，670，1400，1180，690，1890，1260，730，1110，2150，1700，900，660，1290；

研究生：2550，960，1630，1310，2520，1640，1130，1760，2010，3250，1660，1490，1380，1990，1200。

试比较本科生和研究生的收入水平。

由于收入一般是右偏分布，因此不适合用t检验进行分析。我们用Wilcoxon秩和检验来比较两个总体的中位数。检验的零假设和备择假设如下：

H_0：本科和研究生月收入的中位数相等；H_1：本科和研究生月收入的中位数不相等。

在 SPSS 软件中打开数据文件，选择“分析”→“非参数检验”→“2 个独立样本”，在弹出的对话框中将“月收入”设定为检验变量，“学历”设定为分组变量，然后单击“定义组”，按照“学历”的取值进行设定，然后单击“继续”，检验类型使用默认“Mann-Whitney U”，单击 “确定”（相关操作参见图 6-6），得到的分析结果见表 6-11、6-12。

图 6-5　在 SPSS 中进行 Wilcoxon 秩和检验的相关设置

表6-11　Wilcoxon秩和检验中秩和的计算结果

	学历	N	秩均值	秩和
月收入	本科	20	13.55	271.00
	研究生	15	23.93	359.00
	总数	35		

表6-12　Wilcoxon秩和检验的检验统计量和p值

	月收入
Mann-Whitney U	61.000
Wilcoxon W	271.000
Z	-2.967
渐近显著性(双侧)	.003
精确显著性[2*（单侧显著性）]	.002[a]

a. 没有对结进行修正。

根据表 6-11，本科生收入的平均秩为 13.55，研究生为 23.93，说明从样本看研究生的收入的中位数要高于本科生。从表 6-12 看，Wilcoxon W 统计量为 271，用正态分布近似计算时的 Z 值为-2.967。表中显示用正态分布计算时的 p 值（双侧检验）为 0.003，与精确计算的 p 值 0.002 相差不大。根据精确检验的 p 值，在显著性水平大于 0.002 时我们应该拒绝原假设，结论是本科与研究生的收入的中位数不相等。

那么，如果进行单侧检验，相应的p值等于多少呢？根据样本数据研究生的收入的中位数要高于本科生，因此这个样本不可能支持本科月收入的中位数大于研究生的备择假设。对于 H_0：本科月收入的中位数≤研究生以及 H_1：本科月收入的中位数>研究生，相应的p值等于1-双侧检验的p值/2=0.999。对于 H_0：本科月收入的中位数≥研究生以及 H_1：本科月收入的中位数<研究生，相应的p值应该等于双侧检验的p值/2=0.001。因此，样本数据支持本科生收入的一般水平低于研究生的备择假设。

关于 Wilcoxon 秩和检验需要做以下说明：（1）在样本量较小时，应当使用精确检验的结果，根据正态分布进行近似会有较大的误差。（2）严格来说用 Wilcoxon 秩和检验对中位数进行假设检验，需要假两个总体分布有类似的形状才能得出可靠的结论。在实际应用中这一点往往会被忽略。

三、多个独立样本的 Kruskal-Wallis 检验

Kruskal-Wallis 检验是 Wlicoxon 秩和检验的推广，用来对多个总体的中位数进行比较。在单因素方差分析模型不适用于所研究的问题时，Kruskal-Wallis 往往是一种可以替代的非参数检验方法。

对于 k 组样本，每组样本的中位数为 $M_i(i=1,2,\Lambda\ ,k)$，每组样本的容量为 $n_i(i=1,2,\Lambda\ ,k)$，总的样本容量为 $n=\sum_{i=1}^{k} n_i$。Kruskal-Wallis 检验的假设为：

$H_0: M_1=M_2=\Lambda\ =M_k \leftrightarrow H_1: M_i(i=1,2,\Lambda\ ,k)$ 不全相等。

Kruskal-Wallis 检验也是根据秩和来构造检验统计量的。将所有样本的数据合在一起，从小到大排序得到每个数值的秩，然后计算各样本的秩和以及平均秩。如果各组没有显著性差异，则各组的平均秩应该趋于相等；如果各组的平均秩相差较大，则说明各组中位数有显著性差异的可能性较大。

Kruskal-Wallis 检验中的检验统计量为 H 统计量（式 6-3）：

$$H=\frac{12}{n(n+1)}\sum_{i=1}^{k}n_i(\bar{R}_i-\bar{R})^2=\frac{12}{n(n+1)}\sum_{i=1}^{k}R_i^2/n_i-3(n+1) \qquad (6\text{-}3)$$

其中，R_i 为每组的秩和；$\bar{R}_i=R_i/n_i$，$\bar{R}=\sum_{i=1}^{k}R_i/n$。

从式 6-3 看，如果零假设成立，则各组的平均秩与所有数据的平均秩应该相差不大，从而 H 统计量的值不会太大。如果 H 统计量的值过大，则说明，零假设不大可能成立。

当样本组数 k，每组样本样本容量 n_i 不是很小时，检验统计量 H 的抽样分布近似服从自由度为 k-1 的χ^2分布，可以根据χ^2分布计算假设检验中的 p 值。在 k=3，$n_i \le 5$ 时，用χ^2分布分布的误差较大，应该使用统计软件中的精确检验方法计算 p 值，或者根据 Kruskal-Wallis 统计量分布表进行统计决策。得到 p 值以后，如果 p 值小于显著性水平α，则拒绝零假设，说明 k 个总体中位数之间有显著差异。

【例 6.6】例 5.1 用单因素方差分析的方法对 4 个专业的平均起薪进行了比较分析（数据文件起薪 1.xls）。由于不确定总体是否服从正态分布，请使用 Kruskal-Wallis 检验比较四个专业毕业生的起薪是否有显著差异。

检验中的零假设和备择假设如下：

H_0: 四个专业起薪的中位数都相等；H_1: 四个专业起薪的中位数不全相等。

相关软件操作如下：在 SPSS 软件中打开数据文件，选择“分析”→“非参数检验”→“k 个独立样本”，在弹出的对话框中将“起薪”设定为检验变量，“专业”设定为分组变量，然后单击“定义组”，按照“专业”的取值进行设定，然后单击“继续”，检验类型使用默认“Kruskal-Wallis H”，单击 “确定”（相关操作参见图 6-7），得到的分析结果见表 6-13、6-14。

由表 6-13，各组的平均秩处于 5.83-17.50 之间。表 6-14 表明， Kruskal-Wallis 检验中使用χ^2分布进行近似计算时的χ^2统计量为 12.31，自由度为 3，相应的 p 值为 0.006。由于 p 值很小，所以有理由拒绝原假设，即认为这四个专业起薪的中位数不全相等。

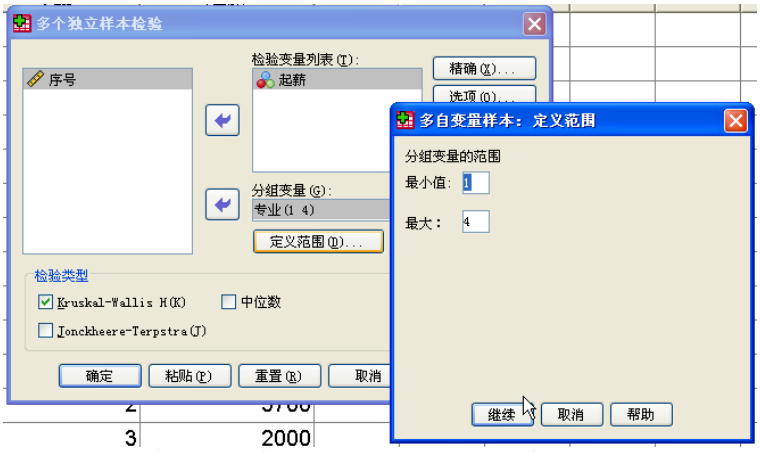


图 6-7　在 SPSS 中进行 Kruskal-Wallis 检验的相关设置

表6-13　Kruskal-Wallis检验中计算的各组的平均秩

	专业	N	秩均值
起薪	1	6	17.50
	2	6	17.25
	3	6	9.42
	4	6	5.83
	总数	24	

表6-14　Kruskal-Wallis检验的检验统计量和p值

	起薪
卡方	12.316
df	3
渐近显著性	.006

关于 Kruskal-Wallis 检验也需要做以下说明：（1）在样本量较小时，应当使用精确检验的结果，根据χ^2分布进行近似会有较大的误差。（2）严格来说用 Kruskal-Wallis 检验对多个总体的中位数进行假设检验，需要假多个总体分布有类似的形状才能得出可靠的结论。在实际应用中这一点往往会被忽略。

本章小结

1. 非参数检验是与总体分布无关的检验，检验中不需要对总体分布的具体形式作出严格假设，或者只需要很弱的假设。

2. 与参数检验相比，非参数检验的特点包括：不需要严格假设条件，适用面广，但在参数方法和发参数方法都适用的情况下非参数检验的功效要小于参数检验。

3. 非参数检验主要是适用于以下场合：参数检验不适用的场合；涉及定性数据时；检验对象不涉及参数时；对各种资料的初步分析。

4. χ^2检验和 K-S 检验都可以帮助我们检验能否认为样本数据来自某种概率分布。前者一般用于定性数据，后者用于定量数据。

5. 符号检验是利用正号和负号的个数对假设进行判定的检验方法；Wilcoxon 符号秩检验是根据正数和负数的秩和构建统计量进行检验的非参数方法。这两种方法都可以用于单样本中位数检验和两个匹配样本的检验，与正态分布时单样本的 t 检验和匹配样本的 t 检验相对应。由于利用了更多的样本信息，在数据呈对称分布时 Wilcoxon 符号秩检的功效要高于符号检验。

6. Wlicoxon 秩和检验可以用来检验两个独立样本的中位数的差异，与正态分布时独立样本的 t 检验相对应。

7. Kruskal-Wallis 检验是与单因素方差分析相对应的非参数检验方法，可以用来检验多个总体中位数的差异。

8. 在大部分非参数检验方法中，在大样本时一般可以根据一些渐进分布计算假设检验中的 p 值。但小样本时，按照渐进方法的计算结果误差会比较大。这时应该使用精确检验的方法计算 p 值，或者查专门的统计表进行统计检验。在一些非参数检验中需要有一些关于总体分布的假设，如连续对称分布等，在使用中要注意判断。

思考与练习

1. 与参数检验相比，非参数检验有哪些优缺点？主要适用于那些场合？

2. 使用“学生调查.sav”文件中的数据检验：

（1）能否认为总体中学生的学习兴趣呈均匀分布？

（2）能否认为总体中学生的身高服从正态分布？

3. 某企业生产一种钢管，规定长度的中位数是 l0 米。现随机地从正在生产的生产线上选取 10 根进行测量，结果为：9.8，10.1，9.7，9.9，9.8，10.0，9.7，10.0，9.9，9.8。问该企业的生产过程是否需要调整。

4. 从上海证券交易所的上市公司随机抽取 10 家，观察其 2008 年年终财务报告公布前后三日的平均股价（如表 6-15），试用参数和非参数方法检验：我国上市公司年报对股价是否有显著性影响？

表 6-15　10 家公司年终财务报告公布前后三日的平均股价

序号	1	2	3	4	5	6	7	8	9	10
年报公布前	15	21	18	13	35	10	17	23	14	25
年报公布后	17	18	25	16	40	8	21	31	22	25

5. 对两种燃料添加剂进行检验以确定它们对汽油行驶里程的影响。对于添加剂 1，我们检测了 7 辆汽车，得到单位汽油的行驶里程数为 82，64，53，61，59，83，76；对于添加剂 2，我们检验了 9 辆汽车，得到单位汽油的行驶里程数为：80，60，65，91，86，84，77，93，75。在显著性水平 $\alpha=0.05$ 下，试用参数和非参数方法检验两种添加剂对汽油行驶里程的影响是否显著。二者的结论是否一致？

6.某公司分别从 A、B、C 三所大学招聘了 7 名、6 名和 7 名毕业生任职于公司同一业务部门。一年后，该公司的人力资源部对他们的工作业绩评定分数如表 6-16，希望分析这 3 所大学毕业生的工作业绩评分是否存在差异。

试分析：该问题是否适合用方差分析的方法进行分析？参数检验与非参数检验的结论是否一致？

表 6-16　三所大学毕业生的工作业绩评定分数

大学 A	25	70	60	85	95	90	80
大学 B	60	20	30	15	40	35	
大学 C	50	70	60	80	90	70	75

第七章 相关与回归分析

从本质来说，所有模型都是错的，但一些很有用。
——美国统计学家 George E. P. Box

一个国家香烟的消费量与癌症的发病率有关系吗？父母的身高是否影响其子女的身高？公司股票的市盈率与老总的薪酬有关联吗？接受高学历教育的人是否比低学历的人有更高的薪水？……现实世界中存在着大量诸如此类的问题，用统计语言来概况，就是两个或者更多个变量之间，是否存在相互关联？进而，存在相关关系的变量间又是如何相互影响的？
相关分析（correlation analysis）和回归分析（regression analysis）可以用来回答这类问题，它们是研究现象之间相互关系的两种基本方法。相关分析研究变量之间相关的方向和相关的程度，但是相关分析不能指出变量间相互关系的具体形式，也无法从一个变量的变化来推测另一个变量的变化情况。回归分析则是研究变量之间相互关系的具体形式，它对具有相关关系的变量之间的数量联系进行测定，确定一个回归方程，根据这个回归方程可以从已知量来推测未知量，从而为估算和预测提供了一个重要的方法。

第一节 相关分析

一、函数关系与相关关系

变量之间的数量关系存在着两种不同的类型：一种是函数关系，另一种是相关关系。
当一个变量取一定数值时，另一个变量有确定值与之相对应，这种关系称为确定性的函数关系。例如，某种商品的销售收入 Y 与该商品的销售量 X、商品价格 P 之间的关系可用 Y＝PX 表示，这就是一种函数关系。
当一个变量取一定数值时，与之相对应的另一变量的数值虽然不确定，但它仍按某种规律在一定的范围内变化。这种关系称为不确定性的相关关系。例如，商品的需求量与商品价格之间的关系、商品的需求量与居民收入之间的关系。

二、相关关系的类型

1．按相关方向划分
按相关的方向可分为正相关和负相关。当两个变量的变化方向相同时，称为正相关；当两个变量的变化方向相反时，称为负相关。
2．按相关形式划分
按相关的形式可分为线性相关和非线性相关。当两个变量的关系大致呈现线性关系时，称之为线性相关。如果两个变量之间的关系呈现于某种曲线方程，称之为非线性相关。
3．按变量多少划分
按所研究的变量多少可分为单相关和复相关。两个变量的相关，即一个变量对另一个变量的相关关系，称为单相关；一个变量对两个以上变量的相关关系，称为复相关。

三、相关关系的测度

1．散点图
散点图又称相关图，是研究相关关系的直观工具。它是以直角坐标系的横轴代表变量 X，纵轴代表变量 Y，将两个变量间相对应的变量值用坐标点的形式描绘出来，用来反映两变量之间相关关系的图形。一般在进行详细的定量分析之前，可以先利用它对现象之间存在的相关

关系的方向、形式和密切程度作出大致的判断。图 7-1 就是不同形态的散点图。

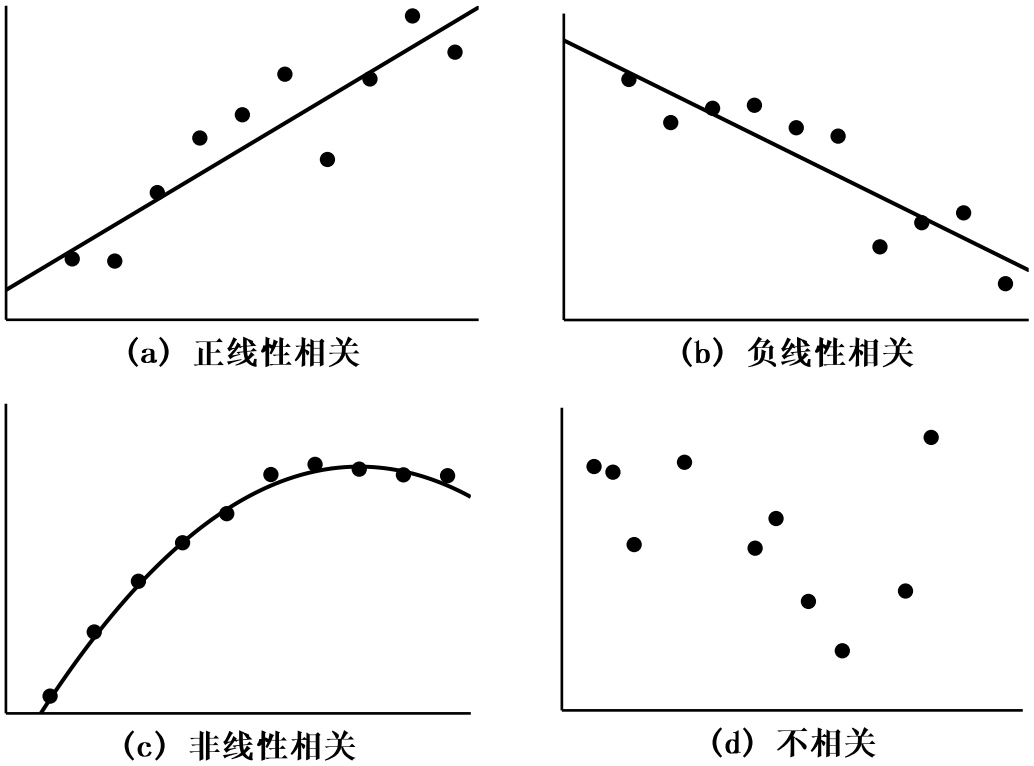


图 7-1　不同形态的散点图

2．单相关系数

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（简称相关系数）。通常以 ρ 表示总体的相关系数，以 r 表示样本的相关系数。

总体相关系数的定义式是：

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{7-1}$$

式中，Cov（X,Y）是变量 X 和 Y 的协方差；Var（X）和 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：

$$r = \frac{\sum(X_t - \overline{X})(Y_t - \overline{Y})}{\sqrt{\sum(X_t - \overline{X})\sum(Y_t - \overline{Y})^2}} \tag{7-2}$$

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。样本相关系数是总体相关系数的一致估计量，它有以下特点：

（1）r 的取值介于－1 与 1 之间。

（2）当 r=0 时，只是表明两个变量之间不存在线性关系，它并不意味着 X 与 Y 之间不存在其他类型的相关关系。因此这时不能轻易得出两个变量之间不存在相关关系的结论，而应结合散点图做出合理的解释。

（3）如果|r|＝1，则表明 X 与 Y 完全线性相关，当 r＝1 时，称为完全正相关，而 r＝－1 时，称为完全负相关。

（4）在大多数情况下，0＜|r|＜1，即 X 与 Y 的样本观测值之间存在着一定的线性关系，当 r＞0 时，Y 值随着 X 增加而增加，此时 X 与 Y 为正相关；当 r＜0 时，Y 随着 X 增加而减少，此时 X 与 Y 为负相关。

3．相关系数的检验

实际中，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，因此需要对相关系数的显著性进行检验。相关系数的显著性检验可分为两类：一是对总体相关系数是否等于 0 进行检验；二是对总体相关系数是否等于某一个给定的不为 0 的数值进行检验。这里只介绍对总体相关系数 ρ 是否等于 0 进行检验。

通过 r 在 X 与 Y 都服从于正态分布条件下，对于 ρ＝0 检验，可以采用 t 检验：

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (7\text{-}3)$$

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 $t_{a/2}$。若$|t|>t_{a/2}$，表明 r 在统计上是显著的。若$|t|\leq t_{a/2}$，表明 r 在统计上是不显著的。

【例 7.1】表 7-1 是 1985-2007 年北京市城镇居民人均年消费性支出（变量 Y）和人均年可支配收入（变量 X）的有关资料，请对 X 和 Y 变量进行相关分析。

表 7-1　北京市城镇居民家庭人均年消费支出与收入情况

年　份	人均可支配收入 X（千元）	北京市城镇居民家庭恩格尔系数 Z（%）	人均消费支出 Y（千元）	X^2	Y^2	XY
1985	0.9077	50.60	0.9233	0.8239	0.8525	0.8381
1986	1.0675	50.90	1.0674	1.1396	1.1393	1.1394
1987	1.1819	52.70	1.1476	1.3968	1.3170	1.3563
1988	1.4370	51.10	1.4556	2.0649	2.1186	2.0916
1989	1.5971	55.30	1.5204	2.5507	2.3116	2.4282
1990	1.7871	54.20	1.6461	3.1937	2.7095	2.9416
1991	2.0404	54.66	1.8602	4.1634	3.4602	3.7955
1992	2.3637	52.80	2.1347	5.5870	4.5567	5.0456
1993	3.2960	47.80	2.9396	10.8639	8.6412	9.6890
1994	4.7312	46.40	4.1341	22.3846	17.0909	19.5595
1995	5.8684	48.50	5.0198	34.4376	25.1980	29.4578
1996	6.8855	46.60	5.7295	47.4098	32.8266	39.4500
1997	7.8131	43.70	6.5318	61.0447	42.6645	51.0338
1998	8.4720	41.10	6.9708	71.7744	48.5925	59.0567
1999	9.1828	39.50	7.4985	84.3238	56.2275	68.8572
2000	10.3497	36.30	8.4935	107.1161	72.1395	87.9051
2001	11.5778	36.20	8.9227	134.0455	79.6146	103.3052
2002	12.4639	33.80	10.2858	155.3488	105.7983	128.2016
2003	13.8826	31.70	11.1238	192.7266	123.7389	154.4273
2004	15.6378	32.20	12.2004	244.5408	148.8498	190.7874
2005	17.6530	31.80	13.2442	311.6284	175.4088	233.7999
2006	19.9780	30.80	14.8250	399.1205	219.7806	296.1739
2007	21.9890	32.20	15.3300	483.5161	235.0089	337.0914
合计	182.1632	—	145.0046	2381.2016	1410.0462	1828.4322

数据来源：《北京统计年鉴（2008）》，北京：中国统计出版社，2008。

首先，绘制变量 X 与 Y 的散点图[3]：

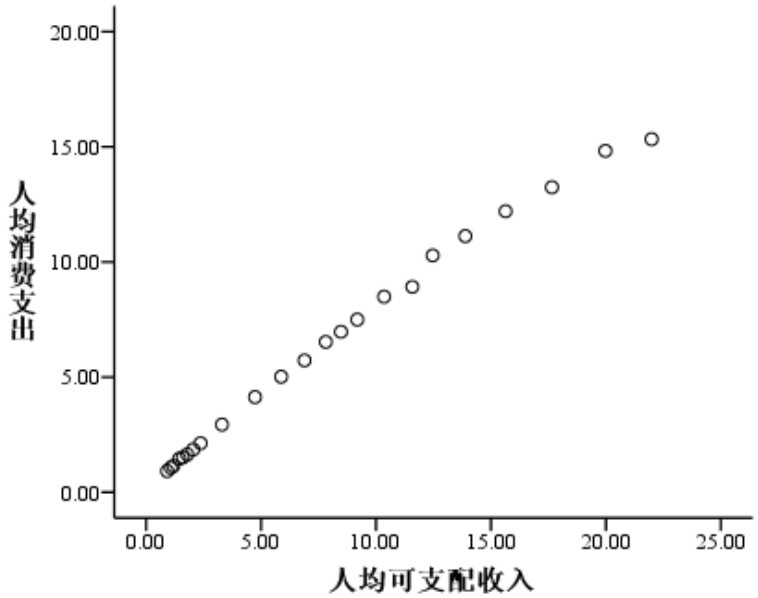


图 7-2　北京市人均可支配收入与人均消费散点图

根据表 7-1 提供的数据，用 SPSS 计算相关系数的具体步骤如下：
首先，将变量 Y 和变量 X 的数据输入到 SPSS“数据视图”窗口中。
第 1 步：选择“分析”下拉菜单；
第 2 步：选择“相关”选项；
第 3 步：选择“双变量”；

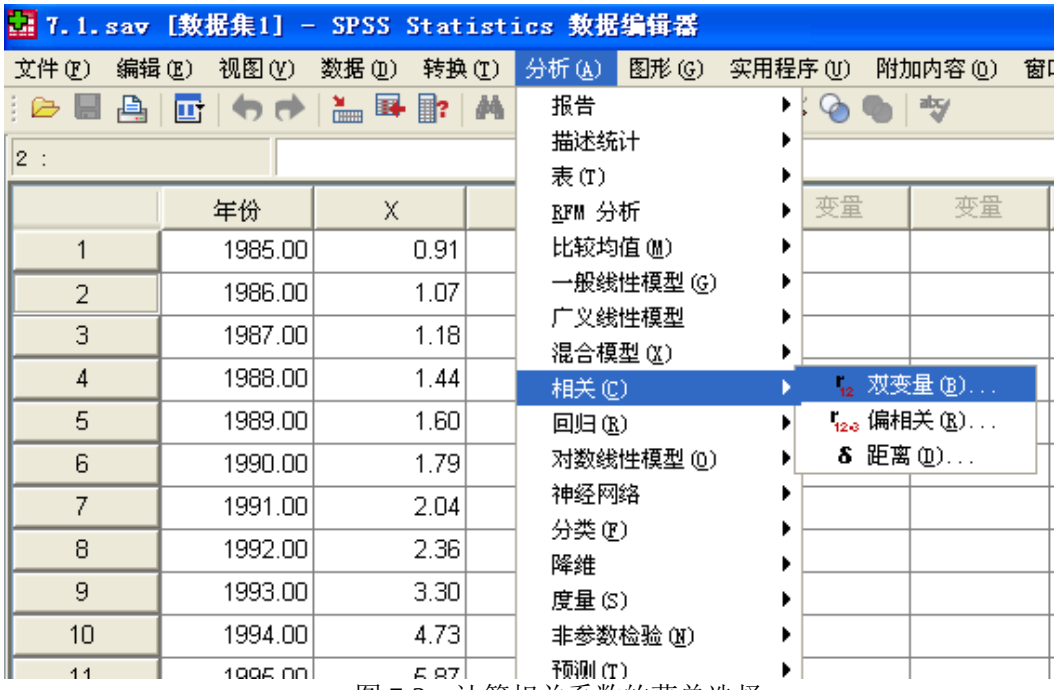


图 7-3　计算相关系数的菜单选择

第 4 步：在“双变量相关”对话框中，将 X、Y 变量放入右侧“变量”窗口，并在“相关系数”中选择“Pearson”相关系数，显著性检验选择“双侧检验”，然后点击“确定”。

[3] 使用 SPSS 17.0 绘制散点图的方法是：依次选择“图形”→“旧对话框”→“散点/点状”，然后在弹出的对话框中选择“简单分布”，再点击“定义”进入“简单分布”对话框，将 X、Y 变量分别选入“X 轴”和“Y 轴”栏中，最后点击“确定”即可。

图 7-4　计算相关系数的对话框和设定

在 SPSS 输出窗口，得到计算结果：

表 7-2　　SPSS 相关系数输出结果

相关性

		人均可支配收入	人均消费支出
人均可支配收入	Pearson 相关性	1	.997**
	显著性（双侧）		.000
	N	23	23
人均消费支出	Pearson 相关性	.997**	1
	显著性（双侧）	.000	
	N	23	23

**. 在0.01 水平（双侧）上显著相关。

结果显示，X 与 Y 变量的相关系数为 0.997，且通过了显著性检验（双侧显著性检验的 P 值为 0.000）。

4．相关关系与因果关系

统计一下世界各国人均电视机拥有量 X 和预期寿命 Y，你会得到很高的正相关系数：拥有较多电视机的国家，其人民的预期寿命也较长。你是否能够做出结论：人均电视机拥有量与预期寿命存在着因果关系？

因果关系的基本含义是：只要改变 X 的值，就可以使 Y 的值改变。我们能不能运一批电视机到博茨瓦纳，来延长当地人民的寿命呢？当然不行。富国人民的预期寿命长是因为他们有较好的营养、干净的饮用水以及较好的医疗服务。电视机与寿命长短之间并没有因果关系。人均电视机拥有量 X 和预期寿命 Y 之间的相关属于“虚假相关”，相关是事实，虚假的是“改变其中一个变量会导致另一个变量改变”的结论。此外，我们还可以举出很多“虚假相关”例子：

统计分析表明，庆祝生日的次数与人的年龄是正相关的，因此，多庆祝几次生日是否有利于长寿？

对小学各年级学生的抽样调查表明，学生的识字水平与他们鞋子的尺寸高度正相关。因此，

让学生穿更大尺码的鞋，他的识字水平就会水涨船高吗？
上述错误的因果推断显然是很荒谬的。实际上，变量之间的相关关系，可以有不同的来源：
（1）直接关系。变量之间存在直接的因果关系，即 X 导致 Y，或 Y 导致 X。
（2）共同反应。X 和 Y 都会随另一个变量 Z 的改变而改变。这种共同反应会制造出一种相关关系，即使 X 和 Y 之间根本不存在直接的因果关系。人均电视机拥有量 X 和预期寿命 Y 即同时受到人均 GDP 的影响。
（3）交叉关系。解释变量 X 和潜在变量 Z 可能同时影响因变量 Y。因为变量 X 和 Z 有关系，我们无法把 X 对 Y 的影响和 Z 对 Y 的影响分离，因此无法界定 X 对 Y 的直接影响有多大。

第二节　一元线性回归分析

统计可以根据目前所拥有的信息（数据）来建立我们所关心的变量和其他有关变量的关系。这种关系一般称为模型（model）。如果用 Y 表示感兴趣的变量，用 X 表示其他可能与 Y 有关的变量（X 也可能是由若干变量组成的向量），则所需要的是建立一个函数关系 Y=f(X)。这里 Y 称为因变量，又称被解释变量或响应变量（dependent variable, response variable），而 X 称为自变量，也称为解释变量或协变量（independent variable, explanatory variable, covariate）。建立这种关系的过程就叫做回归（regression）[4]。

一、总体回归函数与样本回归函数

1．总体回归函数
在回归分析中，最简单的模型是只有一个因变量和一个自变量的线性回归模型，即一元线性回归模型，又称简单线性回归模型。该模型假定因变量 Y 主要受自变量 X 的影响，它们之间存在着近似的线性函数关系，即有；

$$Y_t = \beta_0 + \beta_1 X_t + u_t \qquad (7\text{-}4)$$

公式 7-4 被称为总体回归函数。式中的 Y_t 和 X_t 分别是 Y 和 X 的第 t 次观测值（t=1,…,N），β_0 和 β_1 是未知的参数，又叫回归系数。u_t 是随机误差项，又称随机干扰项，它是一个特殊的随机变量，反映未列入方程式的其他各种因素对 Y 的影响。
2．样本回归函数
总体回归函数事实上是未知的，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。一元线性回归模型的样本回归直线可表示为：

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t \qquad (7\text{-}5)$$

式中的 X_t 是 X 的第 t 次观测值（t=1,…,n）。$\hat{Y}_t$ 是样本回归直线上与 X_t 相对应的 Y 值，可视为 E（Y_t）的估计。$\hat{\beta}_0$ 是样本回归函数的截距系数，$\hat{\beta}_1$ 是样本回归函数的斜率系数，它们是对总体回归系数 β_0 和 β_1 的估计。

实际观测到的因变量 Y_t 值，并不完全等于 $\hat{Y}_t$，如果用 e_t 表示二者之差（$e_t=Y_t-\hat{Y}_t$），则有：

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_t + e_t \qquad (7\text{-}6)$$

上式称为样本回归函数，式中 e_t 称为残差。

---

[4] 回归这个术语是由英国著名统计学家 Francis Galton 在 19 世纪末期研究孩子及他们的父母的身高时提出来的。Galton 发现身材高的父母，他们的孩子也高。但这些孩子平均起来并不像他们的父母那样高。对于比较矮的父母情形也类似：他们的孩子比较矮，但这些孩子的平均身高要比他们的父母的平均身高要高一些。 Galton 把这种孩子的身高向中间值靠近的趋势称之为一种回归效应，而他发展的研究两个数值变量的方法称为回归分析。

3．样本回归函数与总体回归函数的区别

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，总体回归函数中的 β_0 和 β_1 是未知的参数，表现为常数；而样本回归直线中的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 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。

因此，样本回归函数是对总体回归函数的近似反映。回归分析的主要任务就是采用适当的方法，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。

二、一元线性回归模型的估计

1．一元线性回归模型的统计假设

回归分析的主要任务就是要建立能够近似反映真实总体回归函数的样本回归函数。估计回归参数最常用的方法是普通最小二乘法（ordinary least squares, OLS），这是由德国数学家高斯最先提出的。为了使该方法得到模型估计结果保证优良的统计特性，同时使回归模型具有更强的客观现实代表性，一元线性回归模型需满足如下统计假设：

（1）随机误差项的期望值为 0，即对所有的 t 总有：$E(u_t)=0$

（2）随机误差项的方差为常数，即：$\mathrm{var}(u_t)=\sigma^2$

（3）随机误差项之间不存在序列相关关系，协方差为零，即当 $t \neq s$ 时，有：$\mathrm{cov}(u_t,u_s)=0$

（4）自变量 X_t 是非随机变量，也即其取值是确定的，与随机误差项线性无关。

（5）随机误差项服从正态分布。

2．普通最小二乘法

在根据样本数据确定样本回归方程时，一般总是希望 Y 的估计值 $\hat{Y}$ 从整体来看尽可能地接近其实际观测值。这就是说，残差 e_t 的总量越小越好。可是，由于 e_t 有正有负，简单的代数和会相互抵消，因此为了数学上便于处理，通常采用残差平方和作为衡量总偏差的尺度。普通最小二乘法就是根据这一思路，通过使残差平方和最小来估计回归系数的一种方法[5]。设：

$$\begin{aligned} Q &= \sum e_t^2 = \sum (Y_t - \hat{Y}_t)^2 \\ &= \sum (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t)^2 \end{aligned} \qquad (7\text{-}7)$$

很明显，残差平方和 Q 的大小将依赖于 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的取值。根据微积分中求极小值的原理，可知欲使 Q 达到最小，Q 对 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的偏导数必须等于零。

将 Q 对 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 求偏导数，并令其等于零，可得：

$$-2\sum (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) = 0$$

$$-2\sum X_t (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) = 0$$

经整理可得：

$$\hat{\beta}_1 = \frac{n\sum X_t Y_t - \sum X_t \sum Y_t}{n\sum X_t^2 - (\sum X_t)^2} \qquad (7\text{-}8)$$

$$\hat{\beta}_0 = \sum Y_t / n - \hat{\beta}_1 \sum X_t / n = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (7\text{-}9)$$

公式 7-9 中，$\bar{X}$ 和 $\bar{Y}$ 分别是 X 和 Y 的样本均值。

3．高斯-马尔可夫定理

我们为何要使用普通最小二乘法来估计一元线性模型呢？高斯-马尔可夫定理为应用这种方

[5] “二乘”来自于古汉语，就是平方的意思；“普通”意指对各残差项 e_t 求平方和时，采用等权处理，与普通最小二乘法相对应的，还有“加权最小二乘法”等方法，有兴趣的读者可参阅计量经济学教材。

法提供了理论依据。可以证明，在统计假设条件得到满足的条件下，回归系数的最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的期望值等于回归系数真值 β_0 和 β_1，且 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是因变量 Y_t 的线性函数，因此，最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是总体回归系数的线性无偏估计量。数学上还可以进一步证明，在所有的线性无偏估计量中，回归系数的最小二乘估计量的方差最小；同时随着样本容量的增加，其方差会不断缩小。也就是说，此时回归系数的最小二乘估计量是最优线性无偏估计量（Best Linear Unbiased Estimator, BLUE）和一致估计量。

回归系数的最小二乘估计量所具有的上述性质首先是由数学家高斯和马尔可夫提出并证明的，因此被称为高斯-马尔可夫定理。通俗的讲，该定理表明，在满足上述统计假定条件时，普通最小二乘法是一种最佳的估计方法。

但是应当明确，这并不意味着根据这种方法计算的每一个具体的估计值都比根据其他方法计算出的具体估计值更接近回归系数真值，而只是表明如果反复多次进行估计值计算或是扩大样本容量进行估计值计算，按普通最小二乘法计算出的估计值接近真值的可能性（概率）最大。

三、一元线性回归模型的评价和检验

得到回归模型的参数估计，还必须对其进行检验。回归模型的检验包括理论意义检验、一级检验和二级检验。理论意义检验主要涉及参数估计值的符号和取值区间。例如，在对恩格尔系数进行估计时，其取值区间应在 0 至 1 之间。如果根据样本数据估计的恩格尔系数大于 1 或小于 0，则不能通过经济意义检验。一级检验又称统计学检验，它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，具体包括拟合程度评价和显著性检验。一级检验是对所有现象进行回归分析时都必须通过的检验。二级检验又称经济计量学检验，它是对标准线性回归模型的假定条件能否得到满足进行检验，具体包括序列相关检验、异方差性检验等。本节介绍一元线性回归模型的一级检验。

1．拟合优度的评价

一元线性回归方程在一定程度上描述了变量 X 与 Y 之间的数量关系，根据这一方程，可根据自变量 X 的取值来估计或预测因变量 Y 的取值，但估计或预测的精度将取决于回归直线对观测数据的拟合程度。如果各观测数据都落在这一直线上，那么这条直线就是对观测数据的完全拟合，此时用 X 来估计 Y 是没有误差的。各观测数据越是紧密围绕回归直线，说明回归直线对观测数据的拟合程度越好。我们把回归直线与各观测数据的接近程度称为回归直线的拟合优度（goodness of fit）。度量回归直线的拟合优度最常用的指标是判定系数（又称可决系数）该指标是建立在对总离差平方和进行分解的基础之上的。

因变量的实际观测值与其样本均值的离差即总离差（$Y_t - \overline{Y}$）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（$\hat{Y}_t - \overline{Y}$），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（$Y_t - \hat{Y}_t$），它是不能由回归直线加以解释的残差 e_t。

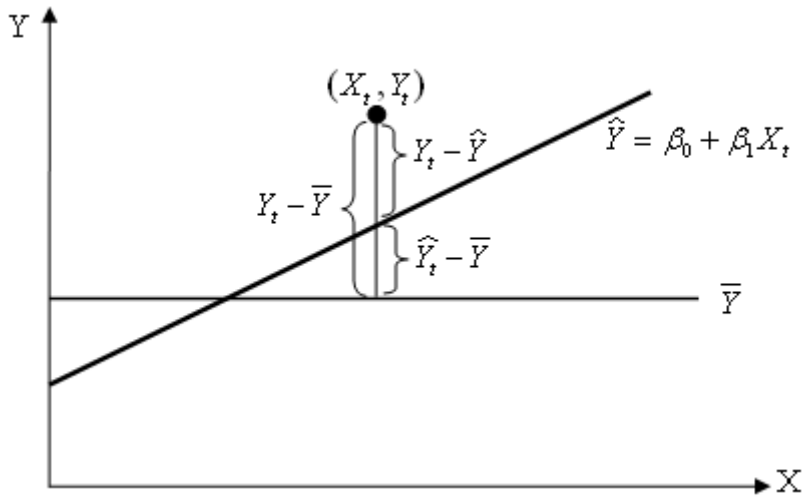


图 7-5　离差分解图

对任一实际观测值 Y_t 总有：

$$Y_t - \bar{Y} = (Y_t - \hat{Y}) + (\hat{Y} - \bar{Y})$$

对上式两边平方并求和，得到：

$$\sum (Y_t - \bar{Y})^2 = \sum (Y_t - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2 + 2\sum (Y_t - \hat{Y})(\hat{Y} - \bar{Y})$$

可以证明：

$$\sum (Y_t - \hat{Y})(\hat{Y} - \bar{Y}) = 0$$

因此：

$$\sum (Y_t - \bar{Y})^2 = \sum (Y_t - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2$$

即

$$SST = SSR + SSE \qquad (7\text{-}10)$$

其中，SST 是总离差平方和，SSR 称为回归平方和，它是由回归直线可以解释的离差平方和。SSE 称为残差平方和，是用回归直线无法解释的离差平方和。公式 7-10 的两边同时除以 SST，得：

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

显而易见，各个样本观测点与样本回归直线靠得越紧，SSR 在 SST 中所占的比例就越大。因此，可定义该比例为判定系数，即：

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y_t - \bar{Y})^2} = 1 - \frac{\sum (Y_t - \hat{Y})^2}{\sum (Y_t - \bar{Y})^2} \qquad (7\text{-}11)$$

判定系数是对回归模型拟合程度的综合度量。如果所有观测点都落在直线上，残差平方和 SSE＝0，R^2＝1，拟合是完全的；如果 Y 的变化与 X 无关，X 完全无助于解释 Y 的离差，则 R^2＝0，可见 R^2 的取值范围是[0，1]。R^2 越接近于 1，表明回归平方和占总离差平方和的比例越大，回归直线与各观测点越接近，回归直线的拟合程度就越好；反之，R^2 越接近于 0，回归直线的拟合程度就越差。

在一元线性回归中，相关系数 r 的平方就等于判定系数；而相关系数 r 与回归系数 $\hat{\beta}_1$ 的正负号是相同的。这一结论不仅可以使我们能由相关系数直接计算判定系数 R^2，也可以使我们进一步理解相关系数的意义。实际上，相关系数 r 也从另一个角度说明了回归直线的拟合优度。|r|越接近 1，表明回归直线对观测数据的拟合程度就越高。

2．估计标准误

估计标准误（standard error of estimate）是对各观测数据在回归直线周围分散程度的一个度量值，它是对随机误差项 u_t 的标准差 σ 的估计。其计算公式为：

$$s_y=\sqrt{\frac{\sum(Y_t-\hat{Y}_t)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE} \qquad (7\text{-}12)$$

估计标准误 S_y 可以看作是在排除了 X 对 Y 的线性影响后，Y 随机波动大小的一个估计量。从估计标准误的实际意义看，它反映了用估计的回归方程预测因变量 Y 时预测误差的大小。若各观测数据越靠近回归直线，S_y 越小，回归直线对各观测数据的代表性就越好，根据估计的回归方程进行预测也就越准确。由此可见，S_y 也从另一个角度说明了回归直线的拟合优度。

3．显著性检验

回归分析中的显著性检验包括两方面的内容：一是对“各回归系数”的显著性检验；二是对“整个回归方程”的显著性检验。前者通常采用 t 检验，后者则是在方差分析的基础上采用 F 检验。

（1）t 检验。回归系数的显著性检验就是检验回归系数是否等于 0，具体步骤如下：

第 1 步：提出假设

H_0：$\beta_1=0$　　H_1：$\beta_1\neq 0$

第 2 步：计算检验的统计量

$$t=\frac{\hat{\beta}_1-\beta_1}{S_{\hat{\beta}_1}} \qquad (7\text{-}13)$$

这个统计量服从自由度为 n－2 的 t 分布。上式中的 $S_{\hat{\beta}_1}$ 是回归系数 $\hat{\beta}_1$ 估计的标准误，统计软件都会给出该值的计算结果：

$$S_{\hat{\beta}_1}=\frac{s_y}{\sqrt{\sum X_i^2-\frac{1}{n}(\sum X_i)^2}} \qquad (7\text{-}14)$$

第 3 步：做出决策。确定显著性水平 a，并根据自由度 df＝n－2 查 t 分布表，找到相应的临界值 $t_{a/2}$。若$|t|>t_{a/2}$，拒绝 H_0，回归系数等于 0 的可能性小于　，表明自变量 X 对因变量 Y 的影响是显著的，即两个变量之间存在着显著的线性关系。若$|t|<t_{a/2}$，则不能拒绝 H_0，表明自变量 X 对因变量 y 的影响是不显著的，二者之间不存在显著的线性关系。

（2）F 检验。为检验模型中的全部自变量作为一个整体对因变量影响的显著性，也就是说检验“整个回归方程”的显著性，可以在方差分析的基础上进行 F 检验。由于一元线性回归模型中只有一个解释变量，因此对 $\beta_1=0$ 的 t 检验与对整个方程的 F 检验是等价的。

将回归平方和 SSR 除以其相应的自由度（自变量的个数 p，一元线性回归中自由度为 1）后的结果称为回归均方，记为 MSR；将残差平方和 SSE 除以其相应的自由度（n-p-1），一元线性回归中自由度为（n-2）后的结果称为残差均方，记为 MSE。一元线性回归模型下，如果原假设成立（H_0：$\beta_1=0$，两个变量之间的线性关系不显著），则比值 MSR/MSE 的抽样分布服从分子自由度为 1、分母自由度为 n-2 的 F 分布，即

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}\sim F(1,n-2) \qquad (7\text{-}15)$$

所以，当原假设（H_0：$\beta_1=0$）成立时，F 值应接近 1；但如果原假设不成立，F 值将变得无穷大。因此，较大的 F 值将导致拒绝原假设，我们就可以断定变量 X 与 Y 之间存在着显著的线性关系。

线性关系检验的具体步骤如下：

第 1 步：提出假设

H_0：$\beta_1=0$，即两个变量之间的线性关系不显著

第 2 步：计算检验统计量 F：

$$F=\frac{SSR/1}{SSE(n-2)}=\frac{MSR}{MSE}$$

第 3 步：做出决策。确定显著性水平α，并根据分子自由度 df_1＝1 和分母自由度 df_2＝（n-2）查 F 分布表，找出相应的临界值F_α。若 F＞F_α，拒绝 H_0，表明两个变量之间的线性关系是显著的；若 F＜F_α，不能拒绝 H_0，表明两个变量之间的线性关系不显著。

4．利用回归方程进行预测

回归方程通过检验后，给定一个自变量，就可以根据回归方程来预测因变量的数值。但值得注意的是，如果自变量超出了现有数据的范围，进行的预测往往是靠不住的。假设你手上有 3－8 岁孩子的身高资料，根据散点图，很容易发现年龄 X 和身高 Y 之间存在很强的线性相关，利用统计软件，你很容易得到一条回归直线。但是，如果你利用该回归直线预测一个人 22 岁时的身高，预测结果会让你目瞪口呆。当然，对人的身高进行预测时，没人会犯这种错误。但是几乎所有的经济预测都在试图告诉我们未来将会发生什么事情，难怪经济预测常常出错。

【例 7.2】根据表 7-1 的数据，建立北京市城镇居民消费模型，以人均年消费性支出（变量 Y）为因变量，以人均年可支配收入（变量 X）为自变量，建立一元线性回归模型，并对回归方程进行显著性检验。假设 2011 年北京市人均年可支配收入为 2.9 万元，请根据已建立的消费模型预测 2011 年人均消费支出。

已知：n＝23，$\sum X$ ＝182.1632，$\sum Y$ ＝145.0046，$\sum X^2$ ＝2381.2016，$\sum XY$ ＝1828.4322，

根据公式（7.9），计算 $\beta_1 = \dfrac{23\times 1828.4322 - 182.1632\times 145.0046}{23\times 2381.2016 - 182.1632^2} = 0.7246$。

根据公式 7-9，计算 $\hat{\beta}_0 = 145.0046/23 - 0.7246*182.1632/23 = 0.5658$。

所以，一元线性回归方程为：$\hat{Y}_t = 0.5658 + 0.7246X_t$。

上式中，0.7246 是边际消费倾向，表示人均可支配收入每增加 1 千元，人均消费支出会增加 0.7246 千元；0.5628 是自主性消费，即与收入无关的最基本人均消费为 0.5628 千元。

将带入 $X_{2011} = 29$ 代入回归方程，得到 2011 年人均消费支出的预测值 $\hat{Y}_{2011} = 0.5658 + 0.7246\times 29 = 21.5792$。

用 SPSS 进行一元线性回归的具体步骤：

第 1 步：选择“分析”下拉菜单；

第 2 步：选择“回归”选项；

第 3 步：选择“线性”；

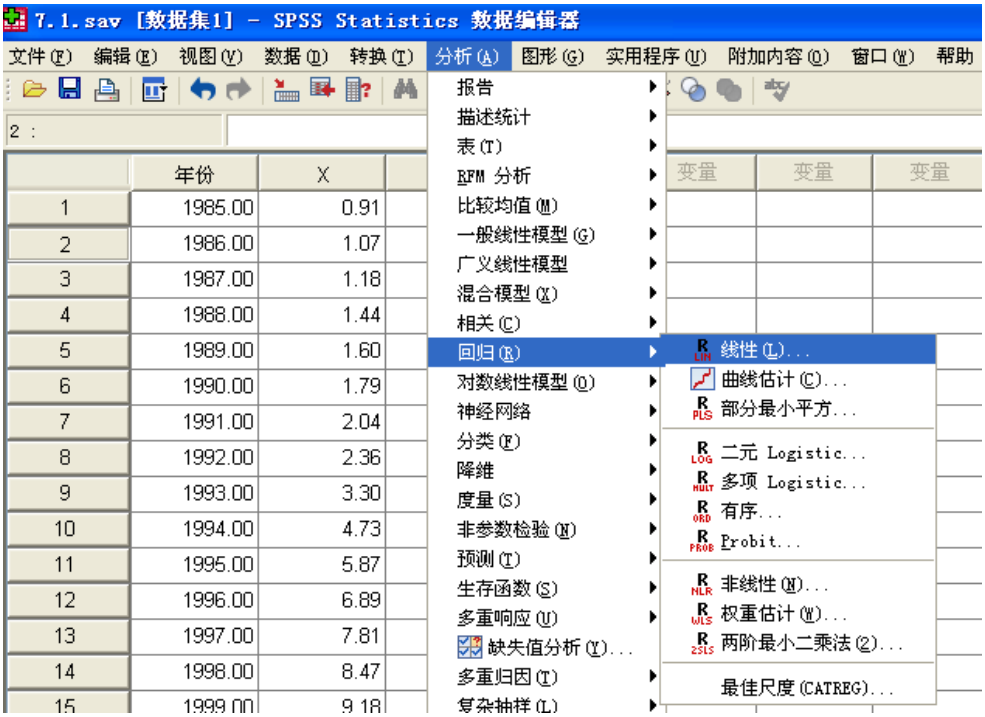


图 7-6　进行一元线性回归的菜单选择

第 4 步：在弹出的“线性回归”对话框中，将 Y 变量选入“因变量”栏中，将 X 变量选入“自变量”栏中，最后点击“确定”。

图 7-7　一元线性回归的对话框和选项设定

现在对 SPSS 输出窗口的计算结果进行分析：

（1）SPSS 输出的回归系数及 t 统计量值：

表 7-3　SPSS 输出的回归系数

系数[a]

模型	非标准化系数		标准系数	t	Sig.
	B	标准　误差	试用版		

1	(常量)	.566	.129		4.390	.000
	人均可支配收入	.725	.013	.997	57.199	.000

a. 因变量: 人均消费支出

①一元线性回归方程：$\hat{Y}_t = 0.566 + 0.725X_t$，与前面的计算相同。

②t 统计量为 57.199，如果显著性水平α为 0.05，自由度为 21，查 t 分布表，找到相应的临界值 $t_{a/2}$=2.08。由于|t|>$t_{a/2}$，则能够拒绝 H_0，表明自变量 X 对因变量 y 的影响是显著的，二者之间存在显著的线性关系。

③表中的“Sig.”即为双侧 t 检验的 P 值，其 0.000 的取值同样说明有相当大的把握拒绝原假设 H_0，表明自变量 X 对因变量 y 的影响是显著的。

（2）F 检验结果

表 7-4　SPSS 输出的 F 检验结果（方差分析表）

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	492.693	1	492.693	3271.723	.000[a]
	残差	3.162	21	.151		
	总计	495.856	22			

b. 因变量: 人均消费支出

①得到 F 统计量为 3271.723，如果显著性水平α =0.05，查 F 分布表（分子自由度为 1、分母自由度为 21），得到临界值F_α =4.32。由于 F>F_α，可以拒绝 H_0，表明方程整体的线性关系显著。

②F 检验的边际概率（P 值）为 0.000，同样表明方程整体线性关系显著。对于一元线性回归方程，其 F 检验和 t 检验是等价的，因此 F 检验与 t 检验获得相同的结论就不足奇了。

③表中给出了回归平方和 SSR=492.693，残差平方和 SSE=3.162，它们除以各自自由度分别得到 MSR=SSR/1=492.693，MSE=SSE/21=0.151，根据公式 7-15，同样可得到相同的 F 统计量值。

（3）拟合优度

表 7-5　SPSS 输出的拟合优度数据

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.997[a]	.994	.993	.38806

a. 预测变量: (常量), 人均可支配收入。

①表 7-5 中的 R 为 R^2 正的平方根。由于$|r| = 0.997$，由于$\hat{\beta}_1$为正数，所以人均消费支出 Y 与人均可支配收入 X 的相关系数为 0.997。

②判定系数 R^2 为 0.994，其统计含义：在人均消费支出的离差中，有 99.7%可以由消费与收入之间的线性回归方程来解释；或者说，在人均消费支出的变动中，有 99.7%是由收入因素决定的，说明该回归方程的拟合程度较好。

③估计标准误s_y为 0.38806 千元，其统计含义：根据收入对消费进行估计时，平均的估计误差为 388.06 元。

⑤根据表 7-4 的数据，回归平方和 SSR=492.693，残差平方和 SSE=3.162，则总离差平方和 SST=492.693+3.162=495.856，判定系数=SSR/SST=492.693/495.856=0.994，与表 7-5 的结果是一致的。

⑥利用公式 7-12，根据表 7-4 给出的 SSE 数据，估计标准误$s_y = \sqrt{3.162/21} = 0.388$，与表 7-5 的结果是一致的。

第三节　多元线性回归分析

在许多实际问题中，影响因变量的因素往往有多个，这种一个因变量同多个自变量的回归问题就是多元回归，当因变量同各自变量之间为线性关系时，称为多元线性回归。多元线性回归分析的原理同一元线性回归基本相同，但计算上要复杂得多，因此需借助于统计软件来完成。

一、多元线性回归模型

设因变量为 Y，p 个自变量分别为 X_1，X_2，…，X_P，多元线性回归模型可定义如下：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + u_t \qquad (7\text{-}16)$$

其中下标 t 表示自变量和因变量的第 t 次观测值（t=1,…,n），β_0，β_1，β_2，…，β_P 是模型的参数，u_t 为误差项。

式 7-16 表明：Y 是 X_1，X_2，…，X_P 的线性函数（$\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt}$）加上误差项 u_t。误差项反映了除 X_1，X_2，…，X_P 之外的随机因素对 Y 的影响，是不能由 X_1，X_2，…，X_P 与 Y 之间的线性关系解释的误差。

与一元线性回归模型相似，为了保证普通最小二乘法估计结果的优良统计特性，同时使回归模型具有更强的客观现实代表性，多元线性回归模型不仅要满足前述一元线性回归模型的五项统计假定，额外还需满足：

（1）要有足够数量的数据来进行估计，即观测值的个数要大于待估计的参数的个数：（p+1）<n。

（2）各自变量（X_1，X_2，…X_P）之间不能存在严格的线性关系。

二、多元线性回归方程的参数估计

利用普通最小二乘法，根据样本数据得到的多元线性回归方程，称为估计的多元线性回归方程。

估计的多元线性回归方程一般形式为：

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \hat{\beta}_2 X_{2t} + \cdots + \hat{\beta}_p X_{pt} \qquad (7\text{-}17)$$

式中 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 是参数 β_0，β_1，β_2，…，β_p 的估计值，$\hat{Y}_t$ 是因变量 Y_t 的估计值。其中，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 称为偏回归系数。$\hat{\beta}_1$ 表示当 X_2，X_3，…，X_P 不变时，X_1 每变动一个单位因变量的平均变动量。

三、回归方程的拟合优度

1．复相关系数和修正的判定系数

在多元线性回归分析中，总离差平方和的分解公式 7-10 仍然成立，因此上一节定义的判定系数的计算公式依然成立：

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y_t - \bar{Y})^2} = 1 - \frac{\sum(Y_t - \hat{Y})^2}{\sum(Y_t - \bar{Y})^2} \qquad (7\text{-}11)$$

多元线性回归方程的判定系数 R^2 的正的平方根称为复相关系数，反映变量 Y 与其他多个变量 X_1，X_2，…，X_k 之间的线性相关程度。需要注意的是，在多个变量的情况下，复相关系数只取正值。这是因为，在多个变量时，偏回归系数有两个或两个以上，其符号可能有正有负，不能按正负来区别，所以复相关系数也就只取正值。由此可见，0≤R≤1。

从公式 7-11 可以看出，R^2 的大小取决于残差平方和 $SSE = \sum(Y_t - \hat{Y})^2$ 在总离差平方和

$SST=\sum(Y_t-\bar{Y})^2$ 中所占的比重。在样本容量一定的条件下，总离差平方和 SST 与自变量的个数 p 无关，但当自变量的个数 p 增加时，可能会使预测误差变小，从而减少残差平方和 SSE（至少不会增加），因此 R^2 是自变量个数 p 的非递减函数。在一元线性回归模型中，不同的模型都仅包含一个自变量，如果使用的样本容量也一样，判定系数便可以直接作为评价拟合程度的尺度。然而，在多元线性回归模型中，不同模型所包含的自变量个数未必相同，如果在模型中额外增加一个自变量，即使这个自变量没有经济意义，在统计上也不显著，R^2 仍可能会变大。因此，R^2 不适合作为衡量拟合优劣的尺度。为避免增加自变量而高估 R^2，统计学家提出用样本容量 n 和自变量的个数 p 去修正 R^2，计算修正自由度的可决系数（又称修正的多重判定系数）。

修正自由度的判定系数计算公式为：

$$R_a^2=1-(1-R^2)\times\frac{n-1}{n-p-1} \qquad (7\text{-}18)$$

由于 R_a^2 同时考虑了样本容量 n 和模型中参数的个数 p 的影响，这就使得 R_a^2 的值永远小于 R^2，而且 R_a^2 的值不会由于模型中自变量个数的增加而越来越接近 1。在多元回归分析中，我们通常用修正的多重判定系数来评价回归模型的拟合优度。

2．估计标准误

同一元线性回归一样，多元线性回归中的估计标准误也是对误差项 u_t 的方差 σ^2 的一个估计值，它在衡量多元回归方程的拟合优度方面也起着重要作用。计算公式为：

$$s_y=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-p-1}}=\sqrt{\frac{SSE}{n-p-1}}=\sqrt{MSE} \qquad (7\text{-}19)$$

多元线性回归中对 S_y 的解释与一元线性回归类似。由于 S_y 所估计的是预测误差的标准差，其含义是：根据自变量 X_1，X_2，…，X_P 来预测因变量 y 时的平均预测误差。

四、显著性检验

在一元线性回归中，整个方程线性关系的检验（F 检验）与回归系数的检验（t 检验）是等价的。但在多元线性回归中，这两种检验不再等价。F 检验主要是检验因变量同多个自变量的线性关系是否显著，在 p 个自变量中，只要有一个自变量同因变量的线性关系不显著，F 检验就不能通过，因此 F 检验是对方程整体线性关系的检验。而 t 检验则是对每个回归系数分别进行单独的检验，它主要用于检验每个自变量对因变量的影响是否都显著。如果某一个自变量没有通过检验，就意味着这个自变量对因变量的影响不显著，也就没有必要将这个自变量放进回归模型中了。

1．线性关系检验（F 检验）

线性关系的检验是检验因变量 y 与 p 个自变量之间的关系是否显著，也称为整体显著性检验或联合检验。检验的具体步骤为：

第 1 步：提出假设

H_0：$\beta_1=\beta_2=\ldots\beta_P=0$

H_1：β_1，β_2，…β_P 至少有一个不等于 0

第 2 步：计算检验的统计量

$$F=\frac{SSR/p}{SSE/n-p-1}\sim F(p,n-p-1) \qquad (7\text{-}20)$$

第 3 步：做出统计决策。给定显著性水平 α，根据分子自由度 p，分母自由度（n-p-1）查 F 分布表得 F_α。若 $F>F_\alpha$，则拒绝原假设；若 $F<F_\alpha$，则不能拒绝原假设。根据计算机输出的结果，可以直接利用 P 值做出决策：若 $P<\alpha$，拒绝原假设；若 $P>\alpha$，则不拒绝原假设。

2．回归系数的检验

回归系数检验的具体步骤如下：

第 1 步：提出假设。对于任意参数 β_i（i=1，2，…，p），

H_0：$\beta_i=0$

H_1：$\beta_i\neq0$

第 2 步：计算检验的统计量

$$t_i=\frac{\hat{\beta}_i-\beta_i}{s_{\hat{\beta}_i}}\sim t(n-p-1)\,, \qquad (7\text{-}21)$$

其中 $s_{\hat{\beta}_i}$ 是回归系数 $\hat{\beta}_i$ 的抽样分布的标准误，即

$$s_{\hat{\beta}_i}=\frac{s_y}{\sqrt{\sum X_i^2-\frac{1}{n}(\sum X_i)^2}} \qquad (7\text{-}22)$$

第 3 步：做出统计决策。给定显著性水平 α，根据自由度（n-p-1）查 t 分布表，得 $t_{a/2}$ 的值。若$|t|>t_{a/2}$，则拒绝原假设；若$|t|<t_{a/2}$，则不能拒绝原假设。

【例 7.3】根据表 7-1 的数据，建立北京市城镇居民消费模型，要求以人均年消费性支出（变量 Y）为因变量，以人均年可支配收入（变量 X）和家庭恩格尔系数（变量 Z）为自变量，建立二元线性回归模型，并对回归方程进行显著性检验。

用 SPSS 进行二元线性回归的具体步骤，与上一节介绍的估计一元线性回归模型非常相似：前 3 步完全相同，只是在第 4 步，在弹出的“线性回归”对话框中，将 Y 变量选入“因变量”栏后，需要将变量 X 和变量 Z 同时选入“自变量”栏，最后点击“确定”。

（1）SPSS 输出的回归系数及 t 统计量值：

表 7-6　SPSS 输出的回归系数及 t 统计量

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	5.755	1.097		5.244	.000
	人均可支配收入	.602	.027	.828	21.972	.000
	恩格尔系数Z	-.097	.020	-.179	-4.744	.000

a. 因变量: 人均消费支出

由表 7-6 可以得到以下结果：

①二元线性回归方程为：$\hat{Y}_t=5.755+0.602X_t-0.097Z_t$

②变量 X 的回归系数为 0.602，其统计含义：在居民家庭恩格尔系数不变的条件下，居民可支配收入每上升 1 个单位（千元），居民消费“平均”上升 0.602 个单位（千元）；变量 Z 的回归系数为 0.097，说明在居民可支配收入不变的条件下，居民恩格尔系数每降低 1 个单位（即降低 1%），居民消费水平就会“平均”上升 0.097 个单位（千元）[6]。

③变量 X 的 t 统计量为 21.972，如果给定显著性水平 $\alpha=0.05$，根据自由度＝23-2-1＝20，查 t 分布表，得 $t_{a/2}=2.086$。由于$|t|>t_{a/2}$，则均可拒绝原假设 $\hat{\beta}_1=0$，说明收入水平对居民消费的影响都是显著的。

当然，t 检验的边际概率（P 值）为 0.000，同样说明有相当大的把握拒绝原假设 H_0，表明自变量 X 对因变量 Y 的影响是显著的。

④变量 Z 的 t 统计量为-4.744，如果给定显著性水平 $\alpha=0.05$，则自由度为 20 的 t 分布临界值为 $t_{a/2}=2.086$。由于$|t|>t_{a/2}$，则拒绝原假设 $\hat{\beta}_2=0$，说明生活质量（用恩格尔系数 Z 表示）对居民消费水平的影响是显著的。同样，其边际概率（P 值）为 0.000，也说明有相当大的把握拒绝原假设 H_0，表明自变量 Z 对因变量 Y 的影响是显著的。

[6] 恩格尔系数是居民用于食物消费所占全部消费的份额。该指标可以反映居民生活质量或生活水平，较高的恩格尔系数意味着较低的生活水平。

（2）F 检验

表 7-7　SPSS 输出的 F 检验结果（方差分析表）

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	494.368	2	247.184	3322.541	.000[a]
	残差	1.488	20	.074		
	总计	495.856	22			

b. 因变量: 人均消费支出

F 统计量为 3322.541，如果显著性水平 α ＝0.05，查 F 分布表（分子自由度为 2、分母自由度为 23-2-1=20），得到临界值 F_α ＝3.49。由于 F> F_α ，应拒绝原假设 H_0，表明销售额与两个自变量之间的整体线性关系是显著的。F 检验的边际概率（P 值）为 0.000，同样说明有相当大的把握拒绝原假设 H_0。

（3）拟合优度

表 7-8　SPSS 输出的拟合优度数据

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.998[a]	.997	.997	.27276

a. 预测变量: (常量), 恩格尔系数Z, 人均可支配收入。

由表 7-8 可以得到以下结果：

①表中的 R 为判定系数 R^2 的正的平方根，即复相关系数。R=0.998 说明城镇居民消费水平与恩格尔系数和人均可支配收入之间存在着很强的相关性。

②修正自由度的判定系数为 0.997，其统计含义是，在消费支出的全部离差中，有 99.7%可以由收入与恩格尔系数的二元线性回归方程所解释。

③回归标准误为 0.27276，其统计含义为，根据该二元线性回归方程对城镇居民消费水平进行预测时，平均的预测误差为 272.76 元。

使用 Excel 也可以进行多元线性回归分析，方法同上一节介绍的一元线性回归分析，只是需要注意的是，作为回归模型自变量的各变量，要以相邻的方式安排在 Excel 数据表中。

第四节　回归分析的其他问题

一、非线性回归分析

在实际工作中，并非所有的社会经济现象都可以用线性关系来解释。例如，尽管我们可以工业总产值为因变量、以固定资产（K）和职工人数（L）为自变量得到线性生产函数（$Y=\beta_0+\beta_1K+\beta_2L$），但该生产函数实际上假定固定资产和劳动投入的边际生产率不变，即固定资产投入每增加一单位，产值总是增加 β_1，劳动投入每增加一单位，产值总是增加 β_2。即资金和劳动之间能够完全替代，即便某一生产要素的投入为 0，只要另一生产要素的投入足够多，产值还会继续增加。显而易见，这是不可能的。而要建立边际生产率递减、生产要素之间可以替代又不能完全替代这样一种更符合客观现实的生产函数，就必须考虑非线性回归模型。

非线性回归分析必须着重解决两个问题：第一，如何确定非线性函数的具体形式。与线性回归分析不同，非线性回归函数有多种多样的具体形式，需要根据所要研究的问题的性质并结合实际的样本观测值做出恰当的选择。第二，如何估计函数中的参数。非线性回归分析最常

用的方法仍然是普通最小二乘法，但需要根据函数的不同类型作适当的处理。

1.非线性函数的主要类型

（1）抛物线函数

抛物线方程的具体形式为：

$$Y=a+bX+cX^2 \qquad (7\text{-}23)$$

式中 a、b 和 c 为待估计参数。

判断某种现象是否适用抛物线函数，可以利用“差分法”。其步骤如下：

首先将样本观测值按 X 的大小顺序排列，然后计算 X 和 Y 的一阶差分$\triangle X_1$、$\triangle Y_t$以及 Y 的二阶差分$\triangle Y_{2t}$。

其中：

$\triangle X_1=X_t-X_{t-1}$

$\triangle Y_t=Y_t-Y_{t-1}$

$\triangle Y_{2t}=\triangle Y_t-\triangle Y_{t-1}$

当$\triangle X_t$接近于一常数，而$\triangle Y_{2t}$的绝对值接近于常数时，Y 与 X 之间的关系可以用抛物线方程近似反映。

（2）双曲线函数

如果 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，则可以选用双曲线来拟合。双曲线的方程式是：

$$Y=a+b(1/X) \qquad (7\text{-}24)$$

（3）幂函数

幂函数方程的一般形式是：

$$Y=aX_1^{b_1}X_2^{b_2}...X_k^{b_k} \qquad (7\text{-}25)$$

幂函数的优点在于：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[7]。因此，幂函数在生产函数分析和需求函数分析中，得到了广泛的应用。

（4）指数函数

指数曲线的函数为：

$$Y=ab^x \qquad (7\text{-}26)$$

式中有两个待估计参数 a 和 b。当 a＞0，b＞1 时，曲线随 X 值的增加而弯曲上升，趋于+∞；当 a＞0，0＜b＜1 时，曲线随 X 值的增长而弯曲下降趋于 0。

指数函数被广泛应用于描述客观现象的变动趋势。例如产值、产量按一定比率增长，就符合第一种形式的曲线；而成本、原材料消耗按一定比率降低，就符合第二种形式的曲线。

（5）对数函数

对数函数的方程形式为：

$$Y=a+b\ln x \qquad (7\text{-}27)$$

对数函数的特点是：随着 x 的增大，X 的单位变动对因变量 Y 的影响效果不断递减。

（6）S 形曲线函数

最常用的 S 形曲线是逻辑曲线，逻辑曲线的方程式如下：

$$Y=\frac{L}{1+ae^{-bx}} \qquad (L，a，b>0) \qquad (7\text{-}28)$$

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。

2.非线性回归模型的参数估计

部分非线性回归函数，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包括：

（1）倒数变换

主要用于双曲线函数的线性变换，令 $X^*=1/X$ 代入原方程式，可有：$Y=a+bX^*$。

[7] 所谓 Y 对于 X_j 的弹性，是指在其他变量不变的条件下，X_j 变动 1%时引起 Y 变动的百分比。

（2）半对数变换

主要用于对数函数的线性变换，令 $X^*=\ln X$ 代入原方程式，可有：$Y=a+bX^*$。

（3）双对数变换

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，对幂函数的两边求对数，可得：

$$\ln Y=\ln a+b_1\ln X_1+b_2\ln X_2+\ldots+b_k\ln X_k$$

令 $Y^*=\ln Y$；$b_0=\ln a$；$X_1^*=\ln X_1,\Lambda\ ,X_k^*=\ln X_k$，代入上式可得：

$$Y^*=b_0+b_1X_1^*+b_2X_2^*+\Lambda\ +b_kX_k^*$$

（4）多种方法的综和

对于一些比较复杂的非线性函数，需要综合利用上述的几种方法。例如，对于逻辑曲线函数，若假定 L=100，则可以采用以下方式进行线性变换。首先，公式 7-28 两边同时取倒数，可得：

$$\frac{1}{Y}=\frac{1+ae^{-bx}}{100}$$

进而又有：

$$100/Y-1=ae^{-bx}$$

上式两边取对数，可得：

$$\mathrm{Ln}(100/Y-1)=\ln a-bX$$

令 $Y^*=\ln(100/Y-1)$；$b_1=\ln a$；$b_2=-b$，代入上式，可得：

$$Y^*=b_1+b_2X$$

在以上变换过程中，综合利用了倒数变换和对数变换。

需要指出的是，并不是所有的非线性函数都可以通过变换得到与原方程完全等价的线性方程。此时，需要利用其他方法如泰勒级数展开法等进行估计。

二、逐步回归

在多元线性回归模型中，是否应该尽量包含全部可能的自变量？在实际问题中，当我们选定一个模型并用数据拟合时，并不一定所有的系数都显著，或者说并不一定所有的系数都有意义。出现这种情况的原因是参加回归方程的 p 个自变量中，有些自变量单独看对因变量 Y 有作用（相关程度密切），但 p 个自变量又可能是相互影响的，在作回归分析时，它们对因变量所起的作用有可能被其他自变量代替，从而使得这些自变量在回归方程中变得无足轻重。这时这些自变量若保留在回归方程中，不但增加计算上的麻烦，而且不能保证有好的回归效果。为了克服这些缺点，产生了逐步回归的方法。

逐步回归（Stepwise Regression）是一种一边回归，一边检验从而对自变量进行合理筛选的方法。常用统计/计量经济学软件都具备逐步回归的功能，它要求回归方程中包含所有对因变量作用显著的自变量，而不包含作用不显著的自变量。

SPSS 软件提供了多种筛选自变量的方法：从只有常数项开始，分别计算各自变量对因变量贡献的大小，按由大到小次序逐个地把显著的自变量加入从而建立最优回归方程，称为“向前引入法（Forward）”；或者从包含全部自变量的模型开始，逐步把不显著的自变量剔除，这称为“向后剔除法（Backward）”；更为常用的方法是“逐步引入-剔除法（Stepwise）”，它是向前引入法与向后剔除法的结合，它首先按向前引入法的原则选择较好的自变量进入模型，然后，转用向后剔除法的方法剔除不合标准的自变量，如此反复，通过不断变换纳入的变量和剔除的变量，直至模型外的变量中已经筛选不出可纳入者，模型内也选不出可删除者为止。

SPSS 软件逐步回归的操作方法，前四步与前文所介绍的线性回归操作方法基本相同：

第 1 步：选择“分析”下拉菜单；

第 2 步：选择“回归”选项；

第 3 步：选择“线性”；

第 4 步：在弹出的“线性回归”对话框中，将模型的因变量选入“因变量”栏中，将待筛选的全部自变量选入“自变量”栏；

第 5 步：点击下拉式菜单“方法”，选择“逐步”，然后点击“确定”。

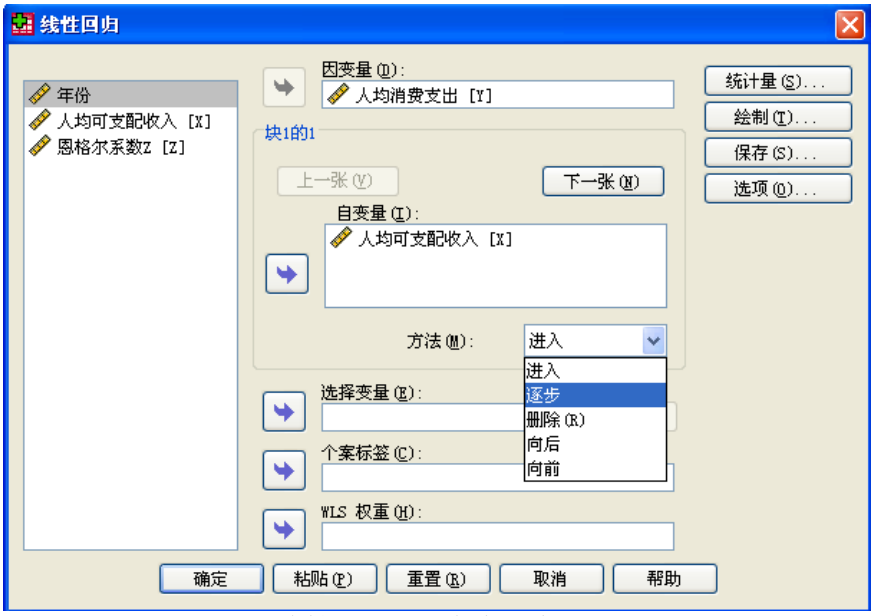


图 7-8　线性回归选项卡中逐步回归的设定

需要注意的是，针对同一个模型和相同的样本数据集，如果选择不同的筛选变量方法，最终得到的模型可能不同。

三、违背回归模型统计假设的后果和补救办法

模型是对丰富多彩的现实世界进行高度简化和概括，因此高度抽象的统计方法与实际建模总是存在差异。在实际建模过程中，由于各种原因，我们常常遇到无法满足第二、三节所介绍的若干项统计假设的情况，如果我们继续用普通最小二乘法估计模型，会出现什么后果？此时高斯-马尔可夫定理可能就不能应用，这意味着普通最小二乘估计量很可能不是“最优线性无偏估计量”，前面介绍过的统计学检验也可能失效，用这样的估计量来推测总体回归模型的系数，显然是不精确的。

对回归模型的假设条件是否得到满足进行检验，并分析由此产生的后果、提出补救方法是计量经济学的重要研究领域，本教材仅对此做简要介绍，有兴趣的同学请查阅相关计量经济学教材。违背回归模型统计假设，主要有如下情况：

1．异方差

当回归模型随机误差项 u_t 的方差不为常数时，即为异方差（Heteroscedasticity）现象：

$$\mathrm{var}(u_t) = \sigma_t^2$$

当异方差出现时，回归模型的估计量不再具有最小方差的性质，因此不再保持有效性；同时，我们此前介绍的 t 检验也失效，无法对回归系数的显著性进行检验。可以采用如下方法对异方差进行假设检验：帕克检验法（Park test）、斯皮尔曼等级相关检验法（Spearman Rank Relation test）、戈德弗尔德—匡特检验法（Goldfeld Quandt test）、格里瑟检验法（Glesjer test）、怀特检验法（White’s General Heteroscedasticity test）以及布鲁奇－帕根检验法（Breusch-Pagan Test）等。

一旦发现异方差的存在，就需要对模型进行变换，使经过变换的模型具有同方差性，然后再使用普通最小二乘法进行估计。通常采用加权最小二乘法（Weighted Least Squres, WLS）对模型进行变换。

2．序列相关

随机误差项之间的协方差不为零时，即存在序列相关（Serial Correlation），又称自相关：当 $t \neq s$ 时，有：

$$\mathrm{cov}(u_t, u_s) \neq 0$$

序列相关的后果与异方差类似，此时尽管普通最小二乘估计量仍为无偏估计量，但不再具有最小方差的性质，即不是“最优线性无偏估计量”；同时，回归系数的显著性检验也失效。

侦测序列相关的检验方法有德宾-沃森检验法（Durbin-Watson d test）、拉格朗日乘数（LM）法等。如果能够确定序列相关的具体形式，就可以采用广义最小二乘法（Generalized Least Squres, GLS）等方法对序列相关进行处理。

3．多重共线性

当模型中两个或多个解释变量高度线性相关时，称为多重共线性（Multi-Collinearity）。解释变量间存在严格线性关系时，称为完全的多重共线性，即存在不全为零的常数 $C_0, C_1, \Lambda\ , C_k$，使：

$$C_0 + C_1 X_1 + \Lambda\ + C_k X_p = 0$$

完全的多重共线性会导致普通最小二乘估计无法应用，通常统计/计量软件会给出出错信息，运算过程中止。如果是较为严重的多重共线性，尽管估计过程不会中断，仍会产生严重的估计问题：各共线解释变量回归系数的估计值会有很大的方差，即估计值的精度降低；解释变量对被解释变量的影响也无法测定；在进行回归系数的显著性检验时，也会增大犯第 II 类错误的概率。

可以根据回归结果来判别多重共线性：如果发现系数估计值的符号明显违背经济常理，或者某些重要的解释变量 t 统计量值较低而同时模型拥有较高的 R^2 值，往往意味着存在多重共线性；此外，一些统计/计量软件（如 SAS、SPSS）也给出了判断多重共线性的指标“方差膨胀因子”（Variance Inflation Factors, VIF），经验法则认为，VIF 指标超过 5，常意味着严重多重共线性的存在。

解决多重共线性的思路是加入额外信息，例如尽量增加样本容量、对模型施加某些约束条件、删除若干个共线解释变量或者将模型适当变形，都可以在一定程度上降低自变量的共线性。此外，也可以采用岭回归（Ridge Regression）等其他估计方法来估计模型系数。

小结

相关分析和回归分析是研究现象之间相关关系的两种基本方法。相关分析研究变量之间相关的方向和相关的程度，但是相关分析不能指出变量间相互关系的具体形式，也无法从一个变量的变化来推测另一个变量的变化情况。

相关系数包括单相关系数（研究两个变量之间的相关程度）、复相关系数（研究一个变量和其它多个变量之间的相关程度）和偏相关系数（研究在其它变量不变的情况下，两个变量之间真实的相关程度）。

回归分析则是研究变量之间相互关系的具体形式，包括线性回归分析（具体包括一元线性回归方程和多元线性回归方程）和非线性回归分析。根据回归方程，可以根据一个变量的变化来预测另一个变量的变化情况。

最小二乘法是通过使残差平方和最小来估计回归系数的一种方法。高斯-马尔可夫定理给为

普通最小二乘法在回归分析中的应用提供了理论基础，因为在统计假设条件得到满足的条件下，回归系数的普通最小二乘估计量是最优线性无偏估计量。

在相关与回归分析中，SPSS的应用主要有以下几个方面：绘制散点图；计算单相关系数、偏相关系数和复相关系数；对一元或多元线性回归模型进行估计，计算模型的拟合优度指标，并对回归系数的估计值进行显著性检验。

思考与练习

1．相关关系与函数关系有何区别？
2．相关分析与回归分析有何区别和联系？
3．什么是单相关系数、复相关系数和偏相关系数？
4．什么是总体回归函数？什么是样本回归函数？它们之间有什么联系和区别？
5．简述最小二乘法的基本原理。
6．解释总离差平方和、回归平方和、残差平方和的含义，并说明它们之间的关系。
7．一元线性回归方程的判定系数的统计含义是什么？
8．一元线性回归方程的估计标准误差的统计含义是什么？
9．对多元线性回归方程的 t 检验和 F 检验有什么区别？
10．为什么要对多元线性回归方程的判定系数进行修正？
11．下面是 7 个地区 2000 年的人均国内生产总值（GDP）和人均消费水平的统计数据。要求：
（1）以人均 GDP 为自变量、人均消费水平为因变量，绘制散点图，并说明二者之间的相关关系。
（2）计算两个变量之间的线性相关系数。
（3）利用最小二乘法求出估计的一元线性回归方程，并解释回归系数的实际意义。
（4）计算判定系数，并解释其意义。
（5）检验回归方程线性关系的显著性（a＝0.05）。
（6）如果某地区的人均 GDP 为 5000 元，预测其人均消费水平。

表 7-9　各地区 2000 年人均国内生产总值与人均消费水平的统计数据

地区	人均 GDP（元）	人均消费水平（元）
北京	22460	7326
辽宁	11226	4490
上海	34547	11546
江西	4851	2396
河南	5444	2208
贵州	2662	1608
陕西	4549	2035

12．下面是 10 个品牌啤酒的广告费用和销售量的数据，要求以广告费支出为自变量，以销售量为因变量，求出估计的回归方程。

表 7-10　各品牌啤酒广告费用与销售量数据

啤酒品牌	广告费（万元）	销售量（万箱）
A	120.0	36.3
B	68.7	20.7
C	100.1	15.9
D	76.6	13.2
E	8.7	8.1
F	1.0	7.1
G	21.5	5.6
H	1.4	4.4
I	5.3	4.4
J	1.7	4.3

13．某汽车生产商欲了解广告费用 X 对销售量 Y 的影响，收集了过去 12 年的有关数据。通过计算得到下面的有关结果。要求：

（1）完成上面的方差分析表。

（2）汽车销售量的变差中有多少是由广告费用的变动引起的？

（3）销售量与广告费用之间的相关系数是多少？

（4）写出估计的回归方程并解释回归系数的实际意义。

（5）检验线性关系的显著性（a＝0.05）。

表 7-11　不同广告费用的方差分析表

方差来源	df	SS	MS	F	Significance F
回归					2.17E-09
残差		40158.07			
总计	11	1642866.67			

表 7-12　广告费用对销售量回归方程的参数估计表

	系数	标准误差	t	Sig.
常数项	363.6891	62.45529	5.823191	0.000168
X	1.420211	0.071091	19.97749	2.17E-09

14．2004 年 10 家航空公司的航班正点率和顾客投诉次数的数据如下。要求：

（1）以航班正点率为自变量，顾客投诉次数为因变量，求出估计的回归方程，并解释回归系数的意义。

（2）检验回归系数的显著性（a＝0.05）。

（3）如果航班正点率为 80%，估计顾客的投诉次数。

表 7-13　航班正点率和顾客投诉次数数据

航空公司编号	航班正点率（%）	投诉次数
1	81.8	21
2	76.6	58
3	76.6	85
4	75.7	68
5	73.8	74
6	72.2	93
7	71.2	72
8	70.8	122
9	91.4	18
10	68.5	125

15．某汽车公司欲了解广告费用（X）对销售量（Y）的影响，收集了最近 10 年广告费用（万元）和销售量（辆）的数据，希望建立二者的一元线性回归方程，并通过广告费用预测汽车的销售量。部分计算结果如下，要求：

（1）写出销售量与广告费用的一元线性回归方程，并解释回归系数的实际意义。

（2）计算判定系数，并解释判定系数的实际意义。

（3）计算汽车销售量与广告费用之间的相关系数。

（4）根据回归方程预测当广告费用为 1000 万元时汽车的销售量。

（5）计算估计标准误差，并解释其实际意义。

表 7-14　某公司广告费用对销售量回归的部分计算结果

回归平方和	755456
残差平方和	37504
回归方程的截距	348.94
回归方程的斜率	14.41

16．下面是随机抽取的 10 家大型超市销售的同类商品的有关数据（单位：元）。要求：

（1）计算 Y 与 X1，Y 与 X2 之间的相关系数，是否有证据表明销售价格与购进价格、销售价格与销售费用之间存在相关关系？

（2）根据上述结果，你认为用购进价格和销售费用来预测销售价格是否有效？

（3）用 Excel 进行回归，并检验模型的线性关系是否显著（a＝0.05）。

（4）解释判定系数 R^2，所得结论与问题（2）中得结论是否一致？

表 7-15　10 家大型超市销售同类商品的有关数据

超市编号	销售价格（Y）	购进价格（X1）	销售费用（X2）
1	1238	966	223
2	1266	894	257
3	1200	440	387
4	1193	664	310
5	1106	791	339
6	1303	852	283
7	1313	804	302
8	1144	905	214
9	1286	771	304
10	1084	511	326

17．一家电器销售公司得管理人员认为，每月的销售额是广告费用的函数，并想通过广告费用对月销售额做出估计。下面是 2004 年 1-8 月的统计数据。要求：

（1）用电视广告费用作自变量、月销售额作因变量，建立估计的回归方程。

（2）用电视广告费用和报纸广告费用作自变量、月销售额作因变量，建立估计的回归方程。

（3）上述问题（1）和问题（2）所建立的回归方程，电视广告费用的回归系数是否相同？对其回归系数分别进行解释。

（4）根据问题（2）所建立的回归方程，在销售收入的总变差中，回归方程所解释的比例是多少？

（5）根据问题（2）所建立的回归方程，检验回归方程的线性关系是否显著（a＝0.05）。

表 7-16　某电器销售公司广告费用与销售收入统计数据

月销售收入 Y（万元）	电视广告费用 X1（万元）	报纸广告费用 X2（万元）
96	5.0	1.5

90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

18．某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下。要求：

（1）确定早稻收获量对春季降雨量和春季温度的二元线性回归方程。

（2）解释回归系数的实际意义。

（3）利用 SPSS 的统计结果，对线性回归方程进行检验。

表 7-17　某农场早稻收获量与春季降雨量及温度的统计数据

收获量（Y）	降雨量（X1）	温度（X2）
2250	25	6
3450	33	8
4500	45	10
6750	105	13
7200	110	14
7500	115	16
8250	120	17

第八章　时间序列分析

时间序列，也称为动态数列或时间数列，从统计意义上讲，就是将某一个指标在不同时间上的不同数值，按照时间的先后顺序排列而成的数列。时间序列可以表明社会经济现象的发展变化过程及其趋势，用以预测现象的发展方向及前景。时间序列从形式上一般由两个基本要素组成，一个是反映客观现象特征的数值（观测值），另一个是观测值所属的时间。比如某企业1990-2002年各月销售额数据（单位：百万元，部分数据见表8-1，数据文件：销售额.sav）。

表 8-1　某企业 1990-2002 年各月销售额数据（部分）

年份	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1990	39.01	44.24	39.50	38.25	38.04	44.15	43.52	46.65	42.30	41.87	39.52	35.18
1991	44.25	43.42	38.90	39.81	39.25	40.96	42.71	45.06	42.49	43.14	43.04	35.31
1992	40.09	46.62	39.65	37.19	39.08	43.70	44.49	49.65	44.46	44.69	42.01	38.17

我们在实际工作中会遇到大量的时间序列，如经济领域中每年的产值、出口、某一商品的销量、某一商品的价格等，社会领域中某一地区的人口数、医院就诊患者人数、铁路客流量等，自然领域的太阳黑子数、月降水量、河流流量、气温等等。每一个序列都是一个客观现象过去的历史记录，包含了产生该序列的系统的历史行为的全部信息。我们所关心的是怎样才能根据这些时间序列，较精确地找出相应系统的内在统计特性和发展规律性，尽可能多地从中提取出我们所需要的信息，并对未来进行预测。用来实现上述目的的方法与过程称为时间序列分析。它是一种根据动态数据揭示系统动态结构和发展变化规律的统计方法，是统计学科的一个分支。

时间序列分析方法如果按其采用的手段不同可概括为数据图法，指标法和模型法三类。数据图法是将时间序列在平面坐标系中绘出坐标图，根据图形直接观察序列的总体趋势和周期变化以及异常点，升降转折点等。这种方法简单、直观、易懂易用，但获取的信息少且肤浅和粗略，分析结果的主观性较大，实践中常与其他方法结合使用。指标法是指通过计算一系列核心指标来反映所研究系统的动态特征。如反映变化率的发展速度和增长速度，反映均衡性和节奏性的动态平均指标和变异指标等等。虽然指标法较数据图法客观，但它所提取的信息仍是肤浅的、有限的。模型法是对给定的时间序列，根据统计理论和数学方法，建立描述该序列的适应统计模型，并进行预测或控制。模型法是本章要介绍的主要内容。

对表8-1的销售额数据序列，采用数据图法可以得到图8-1。从这个数据图可以看出，销售额总的变化趋势是增长的，但增长并不是单调上升的，而是有涨有落，交替起伏。这种起伏有着某种规律，似乎和季节或月份的周期有关系。除了增长的趋势和季节周期之外，还有些无规律的随机因素的作用。采用指标法可以计算年平均增长速度、月同比（平均）增长速度等。如果要揭示序列具体的动态变化规律则需要建立该序列的适应统计模型。

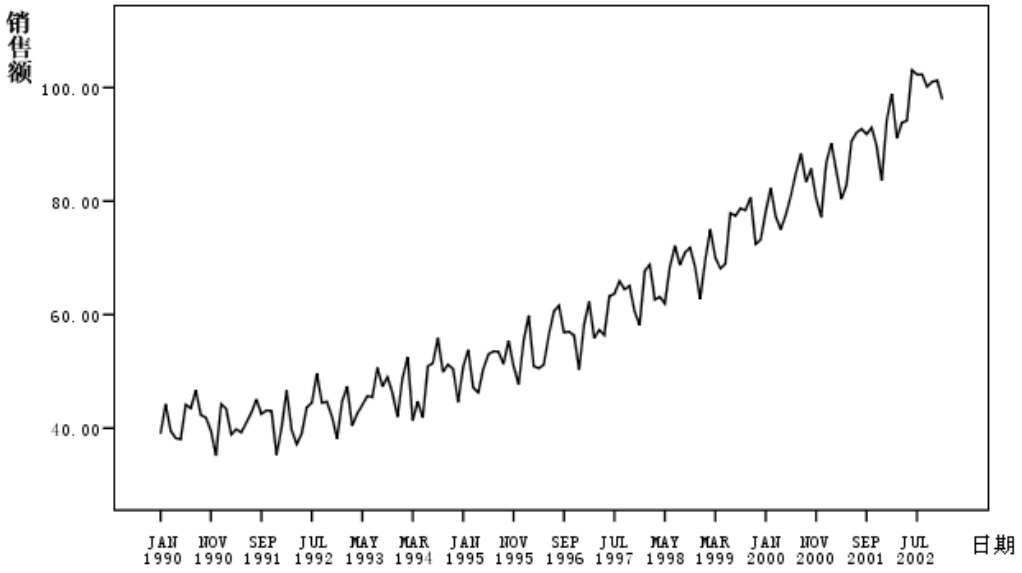


图 8-1　某企业销售额数据图

第一节　时间序列的分解

时间序列的变化是多种因素综合作用的结果。根据影响因素对时间序列进行分解，可以分析各个因素的作用大小，更好的把握时间序列的变化规律，进而预测现象未来的变化。

一、时间序列的构成成分

根据时间序列的影响因素，通常将时间序列的构成成分区分为长期趋势、季节变动、循环变动和不规则变动四种。一个时间序列往往是以下几类变化形式的叠加或耦合。

1. 长期趋势（T_t）。长期趋势是指时间序列朝着一定的方向持续上升或下降，或停留在某一水平上的倾向，它反映了客观事物的主要变化趋势。例如，随着资金和劳动力的增加，技术的不断进步以及劳动生产率的不断提高，我国的经济不断增长，呈向上发展趋势。

2. 季节变动（S_t）。季节变动指一年或更短的时间之内，由于受某种固定周期性因素的影响而呈现出有规律的周期性波动。比如，多数商品的销售量随季节交替出现的周期性波动。

3. 循环变动（C_t）。通常是指周期为一年以上的有规律的波动。比如经济发展周期等。循环变动与季节变动不同，它的波动周期较长，周期的长短不一，变动的规则性和稳定性较差。

4. 不规则变动（I_t）。不规则变动是指由于偶然事件引起的变动，如气候变化、自然灾害、战争、政治事件、国际形势、消费心理、社会舆论、经济政策调整等原因影响经济的变动。通常把不规则变动分为突然变动和随机变动。突然变动是指战争、自然灾害或是其它意外事件引起的变动。随机变动是指由于大量的随机因素导致的变动。根据中心极限定理，通常认为随机变动近似服从正态分布。

二、时间序列分解模型

时间序列可用多种模型进行分解，常见的有加法模型、乘法模型和加乘混合模型。以下用Y_t表示时间序列。

1. 加法模型

$$Y_t = T_t + S_t + C_t + I_t \tag{8-1}$$

加法模型假设时间序列中每一个指标数值都是长期趋势、季节变动、循环变动和不规则变动四种成分的总和，在加法模型中，四种成分之间是相互独立的，某种成分的变动并不影响其他成分的变动。各个成分都用绝对量表示，并且具有相同的量纲。如要测定某种成分的变动，只须从原时间序列中减去其他影响成分的变动即可。

2. 乘法模型

$$Y_t = T_t \times S_t \times C_t \times I_t \tag{8-2}$$

乘法模型是假设时间序列中每一个指标数值都是长期趋势、季节变动、循环变动和不规则变动四种成分的乘积。在乘法模型中，四种成分之间保持着相互依存的关系。一般而言，长期趋势成分用绝对量表示，具有和时间序列本身相同的量纲，其它成分则用相对量表示。如要测定某种成分的变动，须从原时间序列中除去其他影响成分的变动。

3. 加乘混合模型，比如

$$Y_t = S_t + T_t \times C_t \times I_t \tag{8-3}$$

$$Y_t = T_t \times C_t \times S_t + I_t \tag{8-4}$$

时间序列分解模型的选取需要考虑到现象变化的规律和数据本身的特征，如果季节变动（循环变动、不规则变动）依赖于长期趋势的变化，则宜选用乘法模型或加乘混合模型，否则可以考虑加法模型。

三、时间序列长期趋势分析

长期趋势反映了客观事物的主要变化趋势，是时间序列分析首先要研究的变动趋势。通过测定和分析长期趋势可以认识和把握现象发展变化的规律性，并为季节变动和循环变动的测定以及时间序列预测打下基础。测定长期趋势的方法主要有移动平均法和时间回归法。

1. 移动平均法

移动平均法是对原时间序列求连续若干期的平均数作为某一期的趋势值，逐期移动可求得一系列的移动平均数，形成一个新的、派生的序时平均数时间序列。设观测序列为$Y_t, 1 \le t \le T$，平均期数为N。一次N期移动平均的计算公式为：

$$M_t^{(1)} = \frac{1}{N}(Y_t + Y_{t-1} + \Lambda + Y_{t-N+1}) \tag{8-5}$$

在这个新的时间序列$M_t^{(1)}$中，短期偶然因素引起的变动被削弱，从而呈现出客观现象在较长时间的基本发展趋势。移动平均法可以作为测定长期趋势的一种较为简单的方法，在股市技术分析中有广泛的应用。比如对某只股票的日收盘价格序列分别求一次 5 日、10 日、一个月的移动平均就可以得到其 5 日、10 日、一个月的移动平均股价序列，进而得到 5 日线、10 日线、月线，用以反映股价变动的长期趋势。移动平均的期数一般取序列的某种周期长度，比如 4（季节）、12（月份）等。式（8-5）中把时间序列连续N期的平均数作为最近一期（第t期）的趋势值，也可以作为N期的中间一期的趋势值（即中心化移动平均）。当N为偶数时，中心化移动平均需要进行$N+1$项移动平均，首末两个数据的权重为 0.5，中间数据权重为 1，即

$$M_{t-N/2}^{(1)}=\frac{1}{N}(0.5Y_t+Y_{t-1}+\Lambda+Y_{t-N+1}+0.5Y_{t-N})$$ （N 为偶数） （8-6）

考虑销售额时间序列，其12期中心化移动平均序列如图8-2所示。

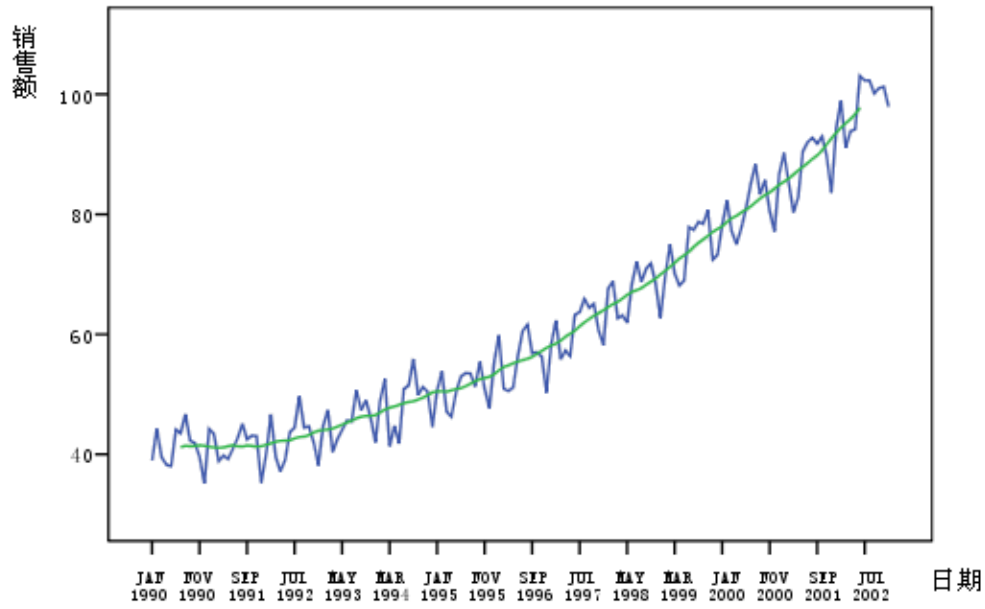


图 8-2　销售额数据的移动平均序列图

2、时间回归法

当时间序列明显包含某种确定性趋势时，通常用时间回归法描述其变化趋势。时间回归法是以所研究的时间序列为因变量，以时间 t 为自变量，利用回归分析原理建立时间序列随时间变化的趋势模型并对未来取值进行预测。这里的时间变量并不具有普通回归中自变量的确切含义，仅仅是时间序列随时间而变化。实际建模中时间变量取值一般用代号（时期数）表示，即$t = 1, 2, 3 \ldots$。

常用来描述时间序列趋势变动的模型有以下几种：

（1）线性方程：$\hat{Y}_t=a+bt$

（2）二次曲线：$\hat{Y}_t=a+bt+ct^2$

（3）指数曲线：$\hat{Y}_t=ab^t$

（4）修正指数曲线：$\hat{Y}_t=K+ab^t,(K>0,a\neq 0,0<b\neq 1)$

（5）Gompertz　(龚帕兹)曲线：$\hat{Y}_t=Ka^{b^t},(K>0,0<a\neq 1,0<b\neq 1)$

（6）Logistic　(逻辑斯谛)曲线：$\hat{Y}_t=\frac{1}{K+ab^t},(K>0,a>0,0<b\neq 1)$

（7）振动曲线：$\hat{Y}_t=f(t)+A\cos(\omega t+\varphi),(f(t)\text{为直线或某种曲线})$

实际中通常先采用数据图法直接观察数据变化的趋势特征，然后选择一类合适的模型进行拟合。选择哪一类模型为宜，往往凭借经验确定，当直观难以确定时用多项式较为方便，也可以拟合几种模型进行比较，利用判定系数 R^2 选择合适的模型。上述模型中的（5）和（6）

常被用来描述事物的成长过程。

再次考虑销售额时间序列，根据图形特征，分别拟合二次曲线和指数曲线，R^2 表明二次曲线效果较好，如图8-3所示。拟合的方程为

$$\hat{Y}_t = 40.851 - 0.009t + 0.003t^2$$

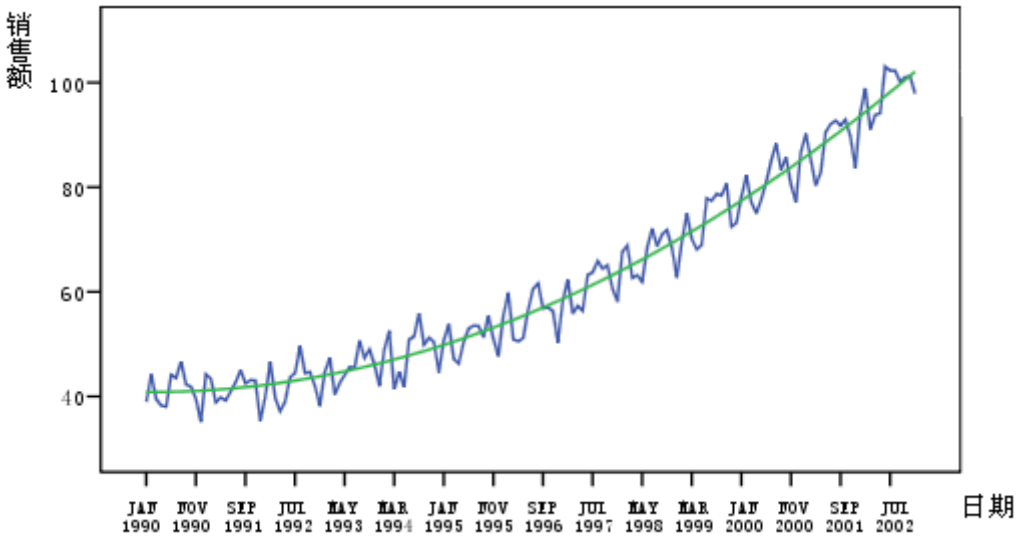


图 8-3　对销售额时间序列数据拟合二次曲线

四　时间序列季节变动分析

时间序列季节变动分析主要是通过计算季节指数（或季节因素）测定季节变动，反映时间序列的季节变动规律，同时为消除时间序列中的季节性因素提供依据。测定季节变动，一般需要先从原时间序列中剔除可能存在的长期趋势，因此需要在一定的模型假定下进行，也有不同的计算方法。实际中乘法模型较为常用，下面以乘法模型为例，介绍移动平均剔除法。

首先，对原时间序列进行 N （周期长度）期中心化移动平均，求得反映长期趋势的移动平均序列；其次，将原时间序列各观测值除以相应的中心化移动平均值，得到剔除长期趋势后的时间序列；再计算剔除长期趋势后的时间序列的同期（同月或季度等）平均值，即为未调整的季节指数；最后，用未调整的季节指数除以剔除长期趋势后的时间序列的总平均值，得到调整后的季节指数，调整后的季节指数之和为 400％（对季度数据）或 1200%（对月度数据）。如果没有季节变动，则各期的季节指数应等于 100%。

【例 8.1】对表 8-1 中销售额时间序列进行季节变动分析。

对该销售额时间序列，分别利用乘法模型和加法模型由 SPSS 软件计算出的季节指数和季节因素如表 8-2 和图 8-4、图 8-5 所示，可以看出，销售旺季为 8 月份，淡季为 12 月份。另外，从图 8-1 可以看出，销售额时间序列的季节变化并未表现出与长期趋势明显的依赖性，因此，使用加法模型分析该销售额时间序列的季节变动较为合适。

表8-2　销售额时间序列的季节变动分析

月份	乘法模型：季节指数（％）	加法模型：季节因素（百万元）

1	101.640	0.972
2	106.828	4.074
3	94.727	-3.028
4	94.091	-3.565
5	94.781	-3.244
6	102.898	1.909
7	104.569	2.711
8	108.162	4.443
9	102.209	1.258
10	103.599	2.131
11	97.455	-1.474
12	89.041	-6.187

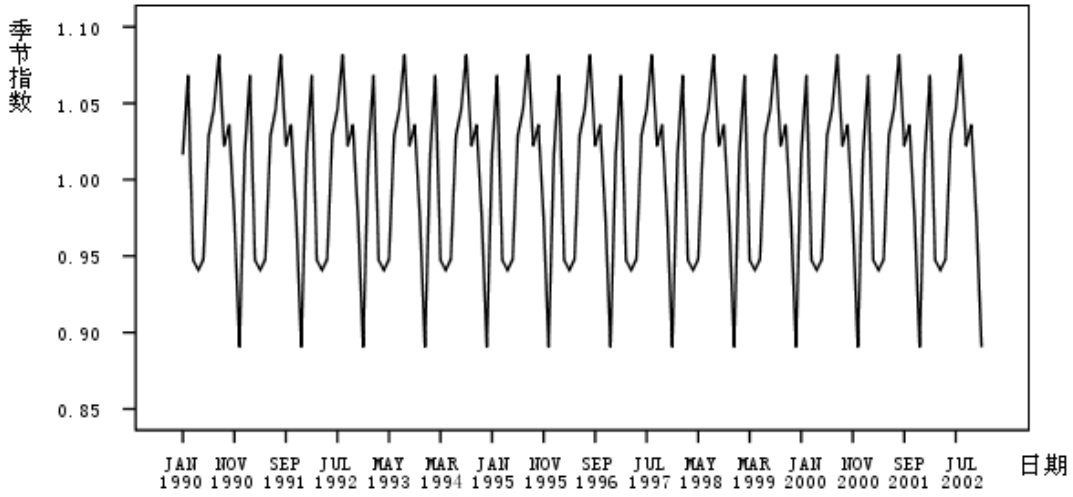


图 8-4　销售额时间序列的季节变动（乘法模型）

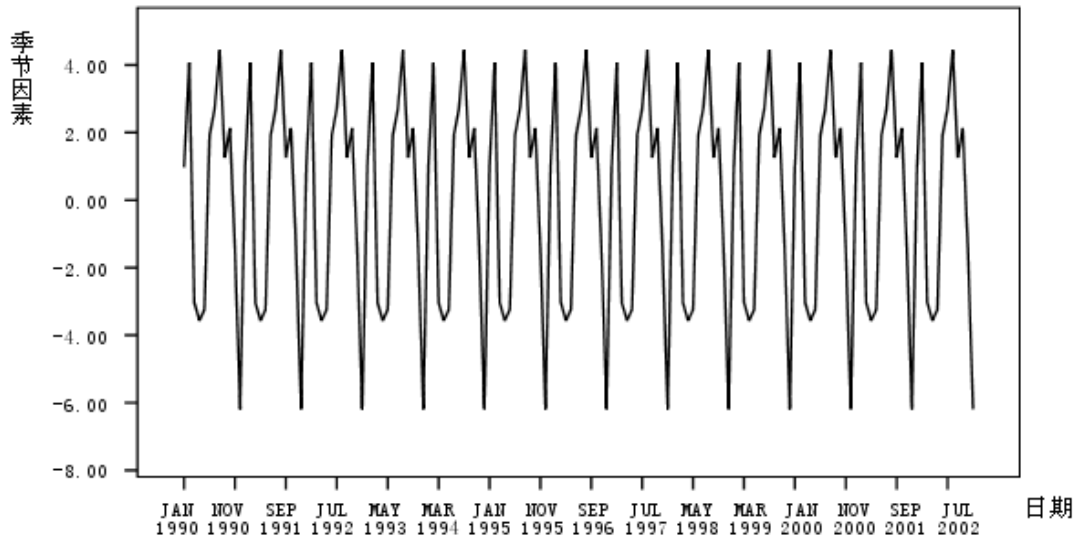


图 8-5 销售额时间序列的季节变动（加法模型）

五 时间序列循环变动分析

时间序列循环变动分析主要是通过测定时间序列循环变动把握时间序列的周期变动规律，通常较多地应用于经济周期波动和商业景气循环研究。实际中常采用剩余法测定循环变动。剩余法的思路是：先求出原时间序列Y_t的长期趋势（T_t）和季节指数（S_t），再从时间序列中除去长期趋势和季节变动，以乘法模型为例，有

$$C_t \times I_t = \frac{Y_t}{T_t \times S_t} \qquad (8\text{-}6)$$

最后用移动平均法除去$C_t \times I_t$中的不规则变动I_t，就得到循环变动C_t。

六、时间序列分解预测法

利用结构模型不仅可以把握时间序列的变化特征，而且可以对时间序列的未来取值进行预测。分解预测法就是依据时间序列的结构模型将序列中的各种非随机成分分离出来，分别进行预测，最后将各部分预测值合成总的预测值。这种方法直观易懂并可以提供较多有用的信息，从不同的方面把握数据的变化特征。

假定时间数列Y_t包含长期趋势T_t、季节变动S_t、循环变动C_t和不规则I_t变动四种成分，以乘法模型为例，分解预测法基本步骤是：（1）对原时间序列进行N（周期长度）期中心化移动平均，求得反映长期趋势和循环变动的移动平均序列T_tC_t；（2）将原时间序列各观测值除以相应的中心化移动平均值，得到剔除长期趋势和循环变动后的时间序列S_tI_t；（3）计算剔除长期趋势和循环变动后的时间序列S_tI_t的同期（同月或季度等）平均值，消除不规则影响，得到未调整的季节指数和调整后的季节指数S_t，反映季节变动；（4）对反映长期趋势和循环变动的移动平均序列T_tC_t用时间回归法建立合适的趋势模型，得到长期趋势值T_t；（5）从反映长期趋势和循环变动的移动平均序列T_tC_t中分离出长期趋势值T_t，得到循环变动指数C_t；（6）合成预测值

$$\hat{Y}_{t+l} = \hat{T}_{t+l} S_{t+l} C_{t+l}$$

其中$\hat{T}_{t+l}$由第（4）步建立的趋势模型得到，S_{t+l}可用同期季节指数代替，C_{t+l}可用半定量化方法预测，即根据分离出的循环变动指数C_t的变化趋势，主观判断C_{t+l}取值的大小。

考虑例 8.1 的销售额数据。为了考察预测效果，利用 1990.1－2001.12 的数据对 2002 年各月的销售额进行预测，这样可以计算预测误差。该销售额序列的循环变动趋势不明显，循环变动指数C_t可以不用考虑。首先对原始序列进行 12 期中心化移动平均，得到T_t；再对T_t序列建立合适的趋势模型，通过比较，二次模型比较合适，估计出的模型为：

$$\hat{Y}_t = 40.782 - 0.007t + 0.003t^2$$

利用估计出的二次模型预测出 2002 年各月份的销售额的趋势值，再乘以季节指数（若利用加法模型，则需加上季节因素），就可以得到 2002 年各月份的销售额的预测值，如表 8-3 所示。加法模型与乘法模型预测的平均绝对误差分别为 1.365 百万元和 2.879 百万元。显然，对该销售额序列，加法模型的预测效果较好。加法模型的预测结果如图 8-6 所示。

表 8-3　销售额时间序列分解法预测

月份	实际值	预测值		绝对误差	
		乘法模型	加法模型	乘法模型	加法模型

JAN 2002	94.28	95.43	94.79	1.15	0.51
FEB 2002	98.89	101.21	98.58	2.32	0.31
MAR 2002	91.09	90.15	92.35	0.94	1.26
APR 2002	93.84	90.08	92.33	3.76	1.51
MAY 2002	94.17	91.45	93.48	2.72	0.69
JUN 2002	103.06	100.10	99.15	2.96	3.91
JUL 2002	102.29	102.87	100.95	0.58	1.34
AUG 2002	102.31	107.64	103.62	5.33	1.31
SEP 2002	100.15	102.08	101.06	1.93	0.91
OCT 2002	101.03	104.38	102.76	3.35	1.73
NOV 2002	101.27	98.81	100.01	2.46	1.26
DEC 2002	97.94	90.89	96.30	7.05	1.64

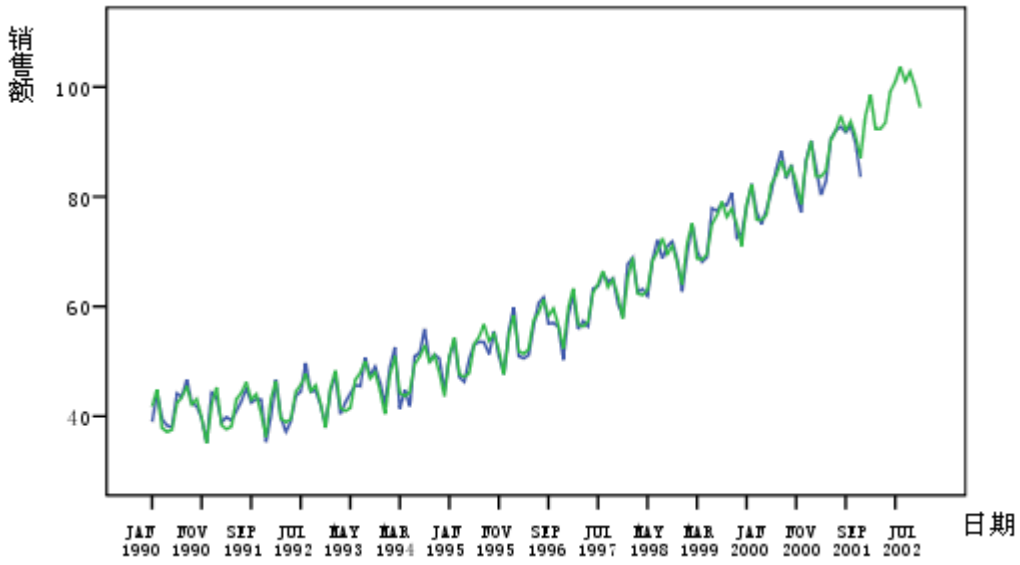


图 8-6　销售额时间序列与分解法预测值

第二节 指数平滑

指数平滑是一种加权移动平均，既可以用来描述时间序列的变化趋势，也可以实现时间序列的预测。指数平滑预测的基本原理是：用时间序列过去取值的加权平均作为未来的预测值，离当前时刻越近的取值，其权重越大。指数平滑方法主要有单参数（一次）指数平滑、双参数指数平滑和三参数指数平滑，分别适用于不同的场合。单参数（一次）指数平滑适用于不包含长期趋势和季节成分的平稳时间序列预测，双参数指数平滑适用于只包含长期趋势的非平稳时间序列预测，三参数指数平滑适用于包含长期趋势和季节成分的非平稳时间序列预测。

一、单参数（一次）指数平滑

单参数指数平滑模型为：

$$F_t = \alpha Y_t + (1-\alpha)F_{t-1} \qquad (8\text{-}7)$$

$$\hat{Y}_{t+1} = F_t$$

式中：F_t 表示时间序列第 t 期的平滑值，Y_t 表示时间序列第 t 期的实际观测值，F_{t-1} 表示时间序列第 $t-1$ 期的平滑值，α 表示平滑系数，并且0< α <1。也就是说，时间序列第 $t+1$ 期的预测值等于第 t 期的实际观测值与预测值的加权平均数。在开始计算时，因为没有第1期的预测值，通常设第1期的预测值等于第1期的观测值，即设 $F_1 = Y_1$。则第2期、第3期、第4期的预测值分别为：

$$\hat{Y}_2 = \alpha Y_1 + (1-\alpha)F_1 = \alpha Y_1 + (1-\alpha)Y_1 = Y_1$$

$$\hat{Y}_3 = \alpha Y_2 + (1-\alpha)F_2 = \alpha Y_2 + (1-\alpha)Y_1$$

$$\hat{Y}_4 = \alpha Y_3 + (1-\alpha)F_3 = \alpha Y_3 + \alpha(1-\alpha)Y_2 + (1-\alpha)^2 Y_1$$

同理，其它时期的预测值依次类推，一般的

$$\hat{Y}_{t+1} = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k Y_{t-k}$$

可见，任何预测值 $\hat{Y}_{t+1}$ 是以前各期实际观测值加权平均的结果。并且

$$\sum_{k=0}^{\infty} \alpha(1-\alpha)^k = 1$$

选择合适的平滑系数 α 是提高预测精度的关键。根据实践经验，α 的取值范围一般以0.1～0.3为宜。如何进一步确定 α 的最佳取值，通常要结合理论分析和模型对比的方法来进行。一般遵循如下的基本准则：如果序列波动较小，则 α 值应取小一些，不同时期数据的权数差别小一些，使预测模型能包含更多历史数据的信息；如果序列趋势波动较大，则 α 值应取得大一些。这样，可以给近期数据较大的权数，以使预测模型更好地适应序列趋势的变化。上述原则可结合模型对比方法来进行。通常将历史数据分成两段，前一段 Y_1, Λ, Y_k，用于建立预测模型，第二段 Y_{k+1}, Λ, Y_T 用于事后预测，以事后预测误差平方和为评价标准，确定最佳 α 值。

二、双参数指数平滑

双参数指数平滑包含两个平滑参数，适用于只包含长期趋势的非平稳时间序列预测。其基本思想是：首先对 Y_t 选定其随时间变化的线性模型，再通过对序列水平和增长量分别进行平滑来估计模型中的参数，是一种双参数平滑法。其预测公式为

$$\hat{Y}_{T+l} = F_t + b_t l \qquad l = 1,2,\Lambda$$

其中

$$F_t = \alpha Y_t + (1-\alpha)(F_{t-1} + b_{t-1})$$

$$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$$

α 和 β 是平滑系数，第一个平滑方程得到原序列经趋势调整的平滑值，第二个平滑方程是对增量 $F_t - F_{t-1}$ 进行指数平滑。初始值取为 $F_1 = Y_1$，$b_1 = \dfrac{Y_m - Y_1}{m-1}$，即前 m 期的平均增长量。

【例 8.2】利用指数平滑法对我国人均原油产量（记为 X_t，见表 8-3，单位：公斤/人）进行预测。

表 8-3　我国历年人均原油产量

年份	X_t	年份	X_t	年份	X_t
1985	118.83	1993	123.25	2001	128.91
1986	122.51	1994	122.57	2002	130.43
1987	123.74	1995	124.54	2003	131.64
1988	124.41	1996	129.22	2004	135.70
1989	123.04	1997	130.68	2005	139.10
1990	121.84	1998	129.61	2006	140.93
1991	122.52	1999	127.63		
1992	121.97	2000	129.09		

从表8-3可以看出，1985年至2006年我国人均原油产量的变化具有向上的长期趋势，可以用双参数指数平滑法进行预测。由SPSS软件搜索出的最终平滑系数α、β分别为1.00和0.001，2007-2010年我国人均原油产量的预测值分别为141.74、142.56、143.37和144.18公斤/人，如图8-7所示。

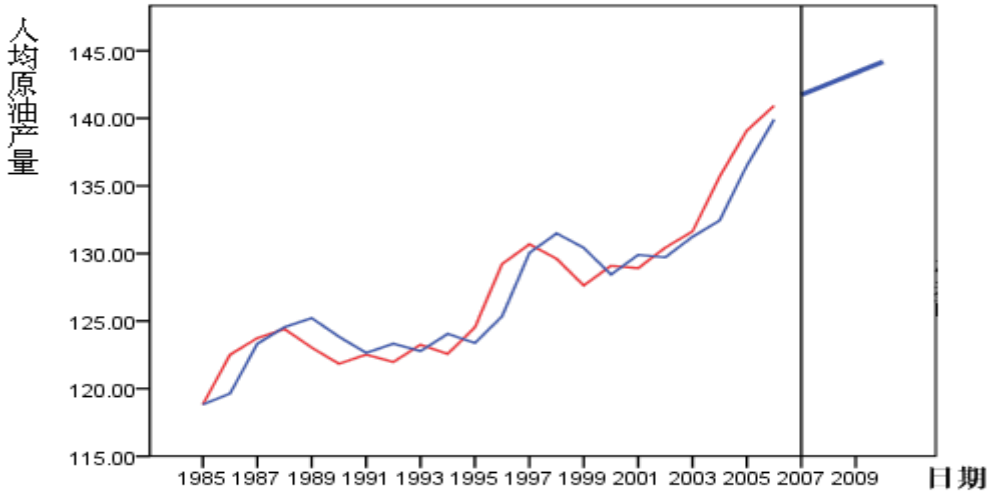


图 8-7　我国人均原油产量预测

三、三参数指数平滑

对既包含长期趋势又包含季节变动的非平稳时间序列进行预测时，常用温特（Winter）指数平滑法，该法包含三个平滑系数。温特指数平滑法是依据时间序列的乘法（或加法）结构模型，在每一步平滑中将原始时间序列分解成趋势成分和季节成分并对它们分别进行平滑。其预测公式为

$$\hat{Y}_{T+l}=(F_t+b_t l)S_{t+l-L} \qquad l=1,2,\Lambda$$

其中

$$F_t=\alpha\frac{Y_t}{S_{t-L}}+(1-\alpha)(F_{t-1}+b_{t-1})$$

$$b_t=\beta(F_t-F_{t-1})+(1-\beta)b_{t-1}$$

$$S_t=\gamma\frac{Y_t}{F_t}+(1-\gamma)S_{t-L}$$

α、β和γ是平滑系数，L是季节变动的周期长度。第一个平滑方程得到不包含季节变动的序列Y_t/S_{t-L}经趋势调整的平滑值，第二个平滑方程是对增量F_t-F_{t-1}进行指数平滑，第三个平滑方程对不包含趋势变动的序列Y_t/F_t（含有不规则影响的季节指数）进行平滑，得到消除不规则影响的季节指数S_t。季节指数共有L个初始值，可由序列的第一个周期数据算得，即

$$\bar{Y}=\frac{1}{L}\sum_{i=1}^{L}Y_i \qquad S_i=Y_i\Big/\bar{Y} \qquad (i=1,2,\Lambda\ ,L)$$

F_t的初始值取为$F_L=\bar{Y}$，b_t的初始值可取为第二个周期头三期数据与第一个周期头三期数据之差的平均值，即

$$b_L=[(Y_{L+1}-Y_1)+(Y_{L+2}-Y_2)+(Y_{L+3}-Y_3)]/3L$$

再次考虑销售额时间序列，该序列包含长期趋势、季节变动和不规则变动，用温特指数平滑法对2003年各月份的销售额进行预测，SPSS软件搜索出的最终平滑系数α、β和γ分别是0.10、0.19和0.34。2003年各月份预测结果依次是：105.15、110.80、102.27、101.48、102.83、112.29、113.97、115.54、113.00、115.11、111.04和105.67（百万元），如图8-8所示。

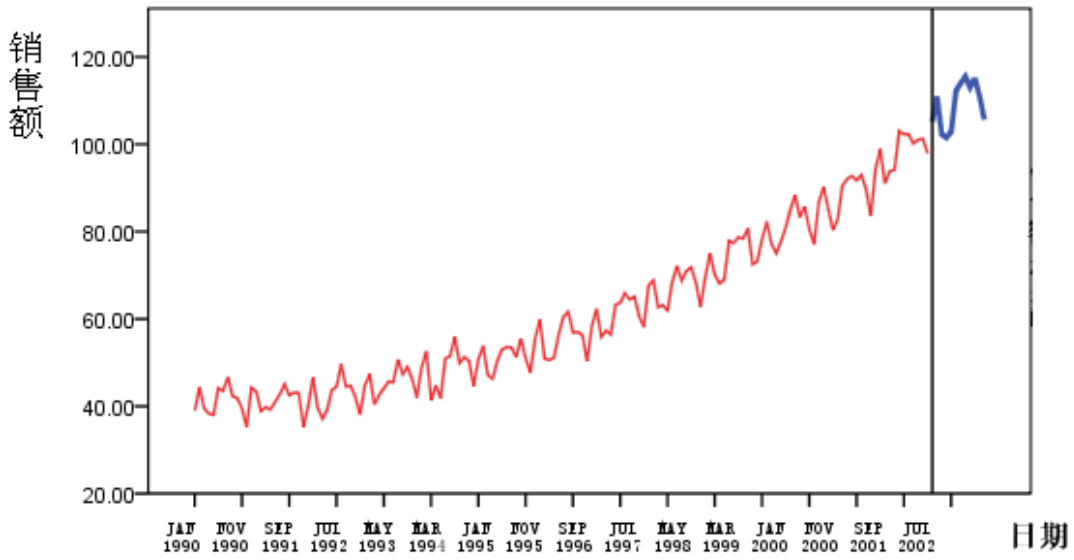


图 8-8　销售额时间序列与温特指数平滑预测值

如果要对时间序列进行更为精确的预测与分析，就需要用到下面介绍的随机时间序列ARIMA模型。可以证明，指数平滑是ARIMA模型的特例。

第三节　ARIMA 模型

用随机过程理论来考察，时间序列可以看作是一个离散随机过程的一次样本实现，时间序列在每一个时间点上的取值具有一定的随机性，根据随机理论研究刻画时间序列的动态规律性（前后期的相关性），就形成随机时间序列分析方法。随机时间序列分析方法能更为精确地刻画时间序列变化的规律。

随机时间序列分析的一个重要概念是平稳性。时间序列平稳性的直观含义是指时间序列没有

明显的长期趋势、循环变动和季节变动。从统计意义上讲，如果序列 X_t 的一、二阶矩存在，而且对任意时刻 t 满足：（1）均值为常数；（2）协方差仅与时间间隔有关，则称该序列为宽平稳时间序列，也叫广义平稳时间序列。实际中研究的时间序列主要是宽平稳时间序列。如果不明确指出，平稳即指宽平稳。反之，不具有平稳性（即序列均值或协方差与时间有关）的序列称之为非平稳序列。下面分别介绍平稳随机时间序列ARMA模型和非平稳随机时间序列ARIMA模型。

一、平稳时间序列模型

1. ARMA模型的基本形式

下面假定 X_t 为零均值平稳时间序列。如果时间序列 X_t 可以表示为其过去值和一个随机冲击 a_t 的线性函数，即

$$X_t = \varphi_1 X_{t-1} + \mathrm{L} + \varphi_p X_{t-p} + a_t \tag{8-9}$$

则式（8-9）称为 p 阶自回归(Autoregressive)模型，简记为AR(p)模型，相应地，时间序列 X_t 称为 p 阶自回归序列。其中 a_t 是互不相关的序列，且均值为零，方差为 σ_a^2（即 a_t 为白噪声序列），一般假定 a_t 服从正态分布。

如果时间序列 X_t 可以表示为当期和过去的随机冲击 a_t 的线性函数，即

$$X_t = a_t - \theta_1 a_{t-1} - \mathrm{L} - \theta_q a_{t-q} \tag{8-10}$$

则式（8-10）称为 q 阶滑动平均(Moving Average)模型，简记为MA(q)模型，相应地，时间序列 X_t 称为 q 阶滑动平均序列。

如果时间序列 X_t 可以表示为当期和过去的随机冲击 a_t 以及其过去值的线性函数，即

$$X_t = \varphi_1 X_{t-1} + \mathrm{L} + \varphi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \mathrm{L} - \theta_q a_{t-q} \tag{8-11}$$

则式(8-11)称为自回归滑动平均(Autoregressive and Moving Average)模型，简记为ARMA(p , q)模型。式（8-11）也常用以下形式表示

$$\Phi(B) X_t = \Theta(B) a_t$$

其中

$$\Phi(B) = 1 - \varphi_1 B - \mathrm{L} - \varphi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \mathrm{L} - \theta_q B^q$$

B 为滞后算子， $B^n X_t = X_{t-n}$ $(n = 0,1,\mathrm{L})$ 。

以上模型中的 $\varphi_1, \varphi_2, \Lambda, \varphi_p$ 与 $\theta_1, \theta_2, \Lambda, \theta_q$ 分别称为自回归参数和滑动平均参数，与 σ_a^2 一起构成模型的待估计参数。

2. ARMA模型的识别与估计

对一个平稳时间序列 $\{ X_t \}$，建立其适应模型的第一步就是模型识别，即判断该序列所适合的模型类型。这里需要用到时间序列的自相关函数（ACF）和偏自相关函数（PACF），自相关函数描述时间序列观测值与其过去的观测值之间的线性相关性，偏自相关函数描述在给定中间观测值的条件下时间序列观测值与其过去的观测值之间的线性相关性。关于这两个概念的细节，有兴趣的读者可以参阅相关的时间序列分析方面的著作。可以证明，三类平稳时间序列的自相关函数和偏自相关函数具有以下不同的统计特性：

模型（序列）	AR(p)	MA(q)	ARMA(p,q)
自相关函数	拖尾	第 q 个后截尾	拖尾
偏自相关函数	第 p 个后截尾	拖尾	拖尾

拖尾是指以指数率单调或振荡衰减，截尾是指从某个开始非常小（不显著非零）。如果一个时间序列是由某一类模型所生成，理论上它就应该具有相应的自相关特征，因而我们可以计算出平稳时间序列的样本自相关函数和样本偏自相关函数，将其特性与以上特性进行比较，进而初步判断序列 X_t 所适合的模型类型。若时间序列的自相关函数和偏自相关函数序列无以上特征，而是出现了缓慢衰减或周期性衰减等情况，则说明序列不是平稳的。需要提及一点，根据ACF和PACF特征判定模型类型有时候并不容易，此时可以借助于下面介绍的信息准则。

利用ACF和PACF判定模型类型的同时也就初步断定了模型的阶数。对于混合的ARMA模型来说，用ACF和PACF判定其阶次有一定的困难，需要用其它的方法来判定。另外，即使对AR或MA模型，用该法判定的阶数也只能是作为初步的参考值，还需结合其它方法确定出精确的阶数。常用的是准则函数定阶法，即确定一个准则函数，建模时按照该准则函数的取值大小确定模型的优劣，使准则函数达到极小值的是最佳模型。该准则函数既考虑到用某一模型拟合时对原始数据的接近程度，同时又考虑到模型中所含待定参数的个数。实际中常用的准则函数是AIC信息准则和BIC信息准则（也称为Schwarz信息准则，记为SIC）：

$$AIC(p,q)=\ln[\hat{\sigma}_a^2(p,q)]+2(p+q)/N$$

$$BIC(p,q)=\ln[\hat{\sigma}_a^2(p,q)]+(p+q)\ln(N)/N$$

其中，$\hat{\sigma}_a^2(p,q)$ 是对序列拟合ARMA（p, q）模型的残差方差，N为观测值的个数。相对于AIC信息准则，BIC信息准则更多的考虑了模型的参数个数。

对时间序列 X_t 所适合的ARMA模型进行初步识别后，接下来就需要估计出其中的参数，以便进一步识别和应用模型。主要的参数估计方法有矩估计法、最小二乘估计法和极大似然估计法等，一般都由计算机软件实现，这里不作介绍。

3. ARMA模型的适应性检验

完成模型的识别、定阶和参数估计后，剩下的问题就是判断这个模型用于描述时间序列是否恰当，也就是模型的适应性检验。模型的适应性检验主要是残差序列 $\{a_t\}$ 的独立性检验。残差序列 $\{a_t\}$ 可由估计出来的模型计算得到。如果残差序列 $\{a_t\}$ 的自相关函数不显著非零，可以认为 $\{a_t\}$ 是独立的。

在前面的讨论中，我们假定所讨论的平稳时间序列是零均值的。实际中我们所遇到的是一段样本观测数据序列，即过程的一个实现，而过程的均值是未知的，我们并不知道过程的均值是否为零。因此，具体建模时，只需给ARMA模型加一个截距项 θ_0 即可。

【例8.3】 一个零均值时间序列 X_t 如图8-9所示，试对其建立ARMA模型。

该序列的ACF和PACF如图8-10所示（图中横线为0±两倍标准差，可以判断ACF和PACF是否显著非零）。可以看出ACF呈拖尾状，PACF第2个后截尾，可初步断定序列适合AR(2)模型。BIC信息准则也表明p=2。由SPSS可以得到参数估计，模型表达式为：

$$X_t=0.42X_{t-1}+0.39X_{t-2}+a_t$$
$$(4.54)\qquad(4.17)$$

括号中为参数的 t 检验值，各参数都是显著的。残差方差的估计为 1.39。由图 8-11 可以看

出残差不存在显著的自相关性，可以认为a_t是独立的，因而模型是适应的。

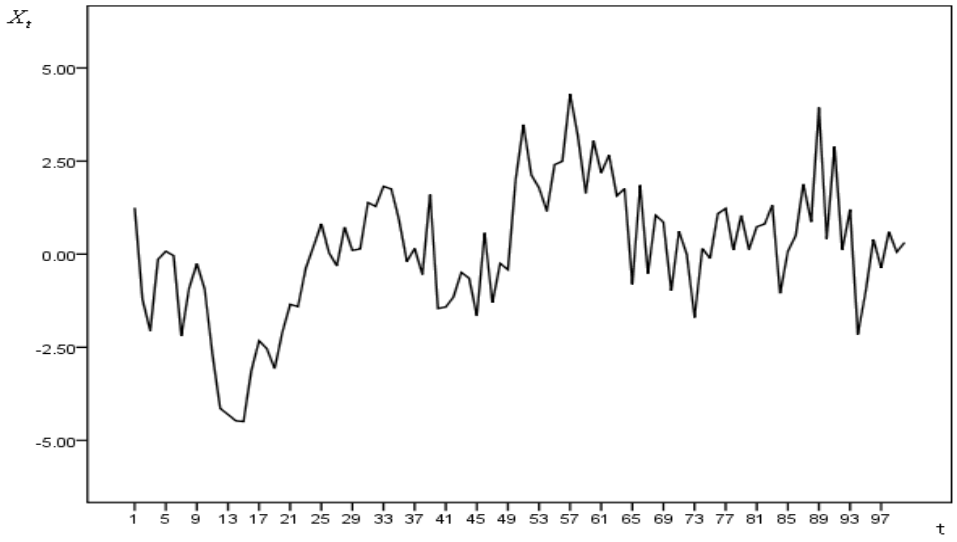


图 8-9　例 8.3 时间序列图

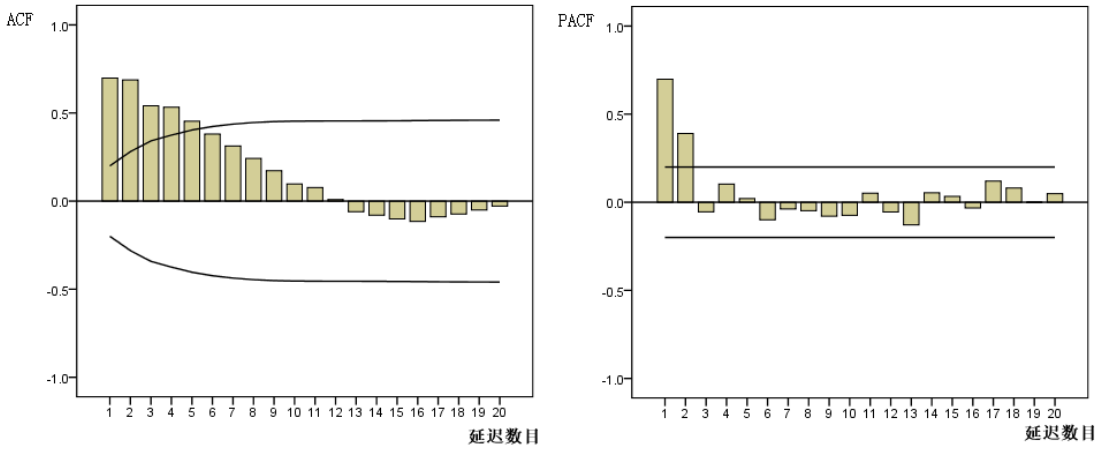


图 8-10　例 8.3 自相关和偏自相关函数图

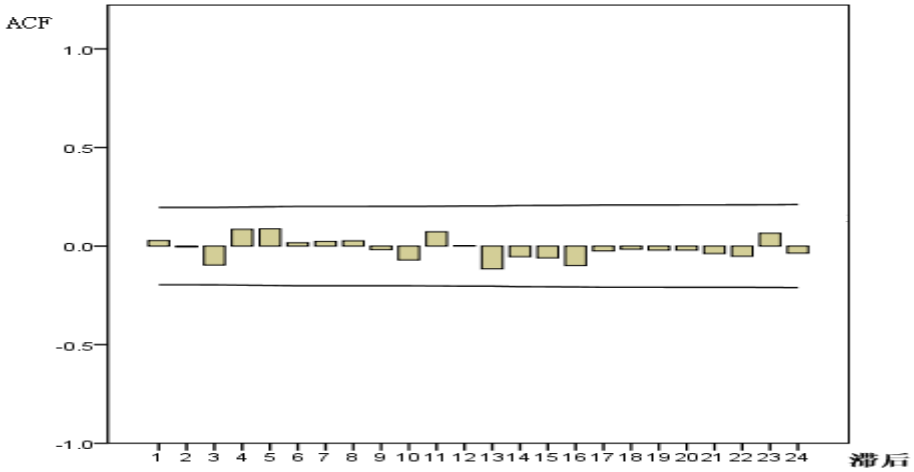


图 8-11　例 8.3 残差的自相关和偏自相关函数图

二、ARIMA 模型

在实际问题中我们常遇到的序列，特别是反映社会、经济现象的序列，大多数并不平稳，而是呈现出明显的趋势性。对于此类时间序列通常采用ARIMA模型进行分析。

ARIMA模型需要用到差分工具。用原序列的每一个观测值减去其前面的一个观测值，就形成原序列的差分序列，记为

$$\nabla^1 X_t = X_t - X_{t-1} \qquad (8\text{-}12)$$

称之为一阶差分。一阶差分可以消除原序列存在的线性趋势。有时候需要进行多次差分（高阶差分），才能够使得变换后的时间序列平稳。一般地，对于 d 阶差分序列 $\nabla^d X_t$（平稳），设其适合ARMA（p, q）模型，即

$$\Phi(B)(\nabla^d X_t) = \Theta(B)a_t \qquad (8\text{-}13)$$

该模型称为求和（或单整integrated）自回归滑动平均模型，记为ARIMA（p,d,q）。之所以称为求和自回归滑动平均模型，是因为求和是差分运算的反运算。

【例8.4】 对某中部省会城市房地产价格数据（2001.1-2005.5）建立适应的ARIMA模型，并进行预测。

该序列有增长的趋势，首先对其进行一阶差分，原序列 X_t 及一阶差分后序列 Y_t 如图8-12所示。差分后序列的自相关和偏自相关函数如图8-13所示。可以看出ACF第一个后截尾，PACF呈拖尾状，初步判定差分后序列适合MA(1)模型，即原序列适合ARIMA(0,1,1)模型。由SPSS可以得到参数估计，模型表达式为：

$$X_t - X_{t-1} = \underset{(2.90)}{25.38} + a_t - \underset{(6.48)}{0.70}a_{t-1}$$

图8-14为残差序列的自相关函数，可以认为 a_t 是独立的，对房地产价格数据建立的ARIMA(0,1,1)模型是适应的。利用该模型可以对房地产价格进行预测，图8-15是房地产价格数据实际值、拟合值以及预测值图示。

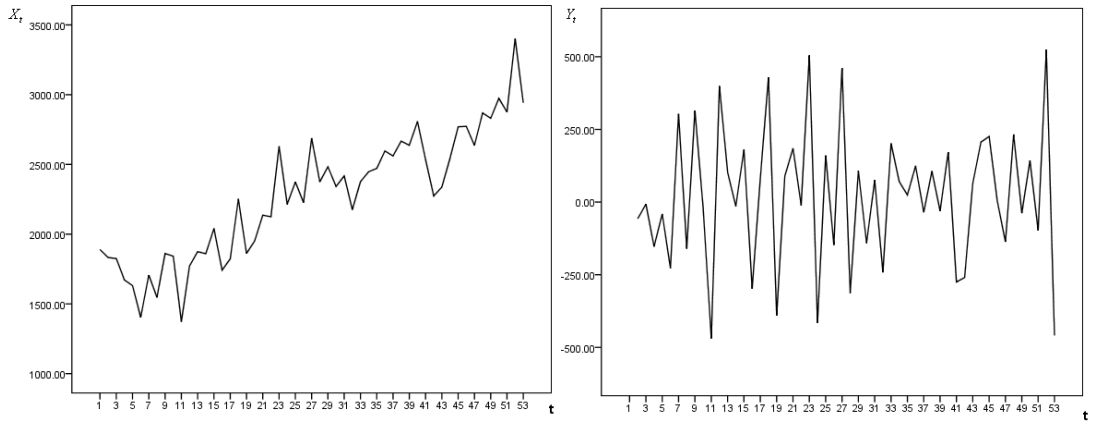


图 8-12　例 8.4 时间序列图

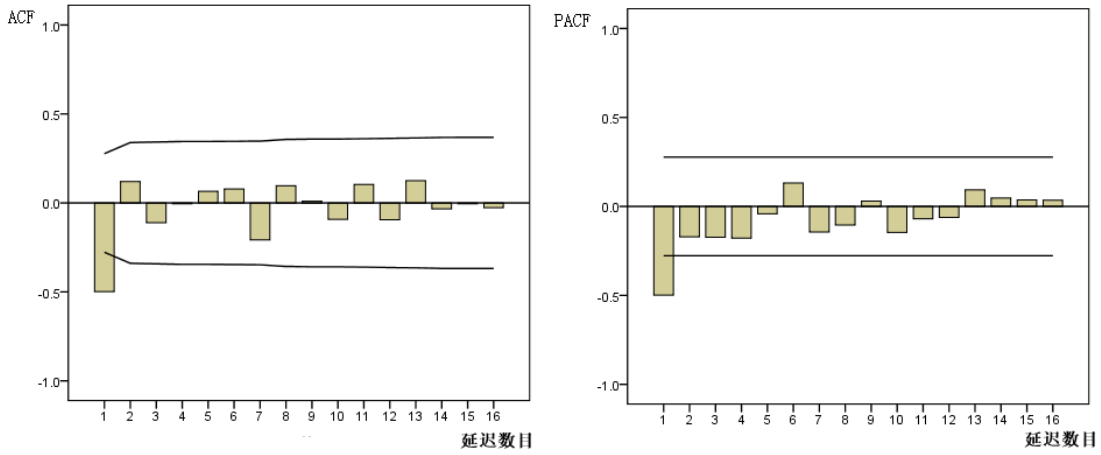


图 8-13　例 8.4 自相关和偏自相关函数图

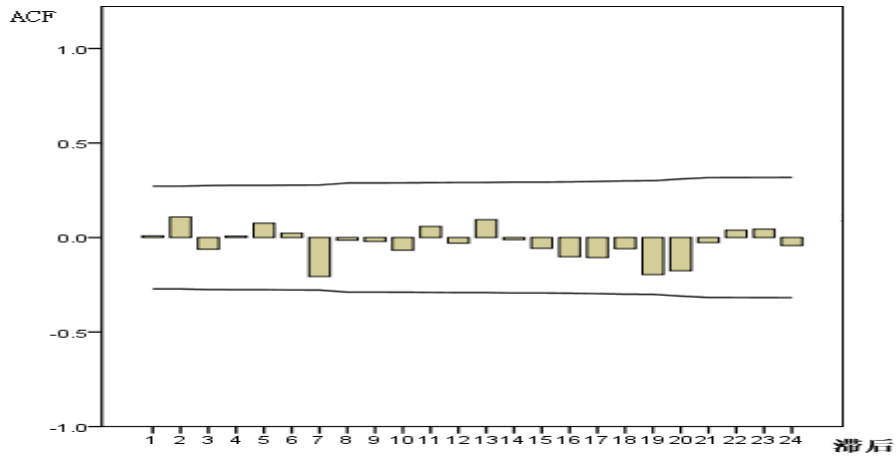


图 8-14　例 8.4 残差序列的自相关函数

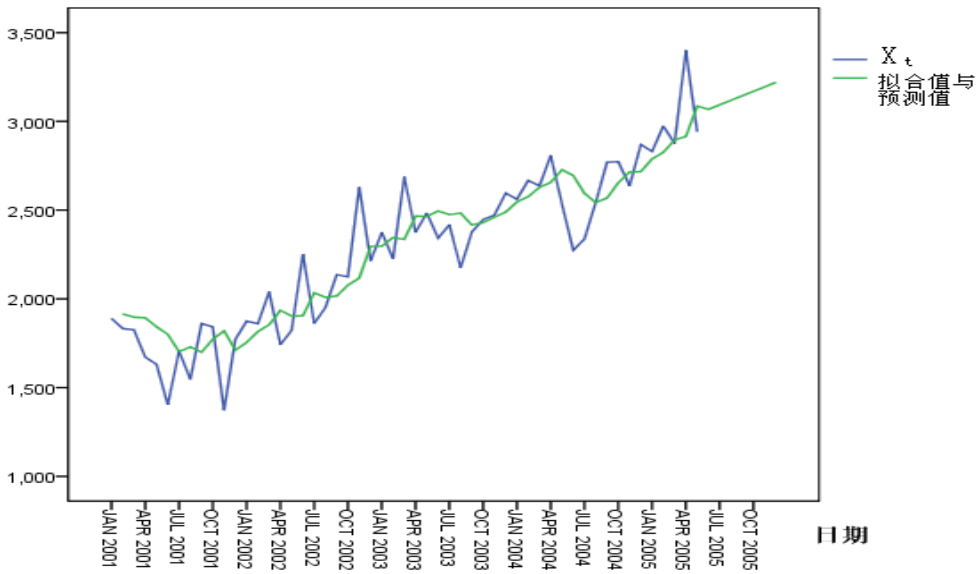


图 8-15　例 8.4 时间序列预测图示

第四节　时间序列分析的软件实现

一、建立时间序列趋势模型

在SPSS数据窗口中依次点击“转换-创建时间序列”，即出现图8-16所示的窗口，在下拉菜单“函数”中选择“中心化移动平均”，在“跨度”框中填写移动平均期数，在左边框中将要分析的时间序列变量通过箭头点进右上方框中，在“名称”框中可以定义新形成的趋势变量名称，最后点击“确定”按钮，就可以在数据窗口形成新的趋势变量。再依次点击“分析-预测-序列图”并选择变量，就可以绘出时间序列图。

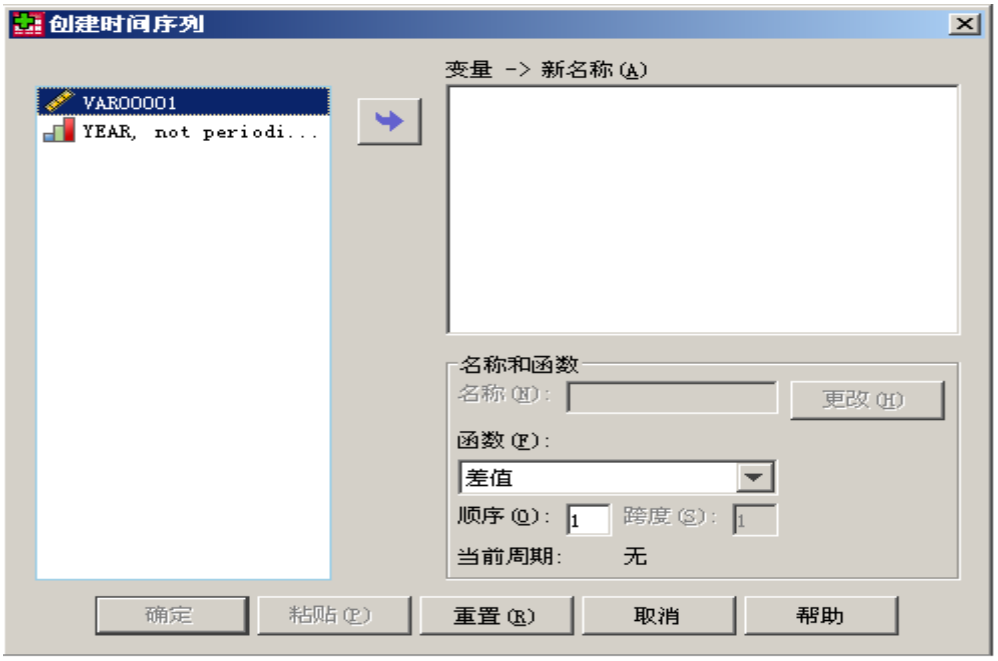


图 8-16　生成移动平均趋势序列

在SPSS数据窗口中依次点击“分析-回归-曲线估计”，即出现图8-17所示的窗口，在左边框中将要分析的时间序列变量通过箭头点进右上方框中，在“自变量”框中选择“时间”，在“模型”框中选择要建立的模型，最后点击“确定”选项，就可以得到时间回归模型的估计结果和图示。

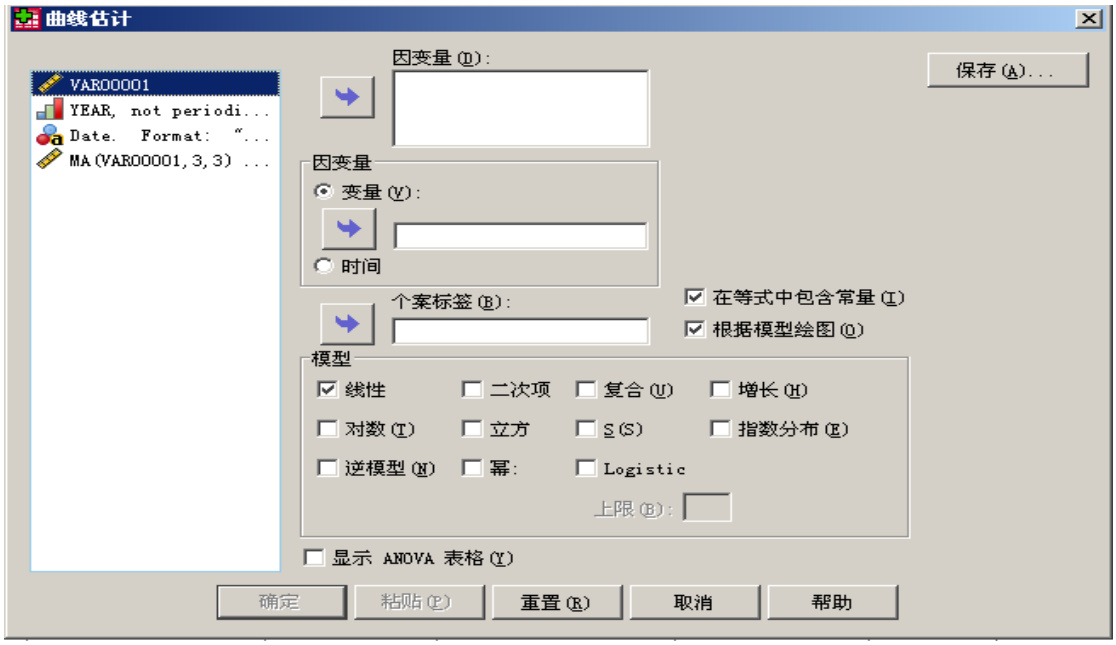


图 8-17　时间序列回归

二、时间序列季节变动分析

在SPSS数据窗口中依次点击“分析-预测-季节性分解”，即出现图8-18所示的窗口，在左边框中将要分析的时间序列变量通过箭头点进右上方框中，在“模型类型”框中选择乘法或加法，最后点击“确定”选项，就可以在数据窗口形成四个新变量：误差、季节性调整序列、季节因子和趋势循环。

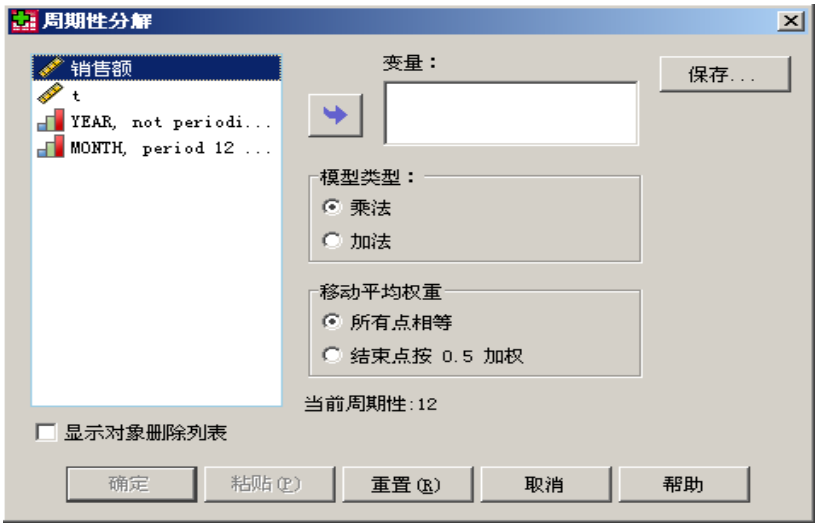


图 8-18　时间序列分解

三、指数平滑预测

在SPSS数据窗口中依次点击“分析-预测-创建模型”，即出现图8-19所示的窗口，在左边框中将要分析的时间序列变量通过箭头点进右上方框中，在“方法”下拉框中选择指数平滑法，点击“条件”按钮，在弹出窗口中选择所用的模型类型，并点击“继续”按钮，再点击右上方“选项”按钮，在弹出窗口中填入预测期，点击右上方“统计量”按钮，在弹出窗口中选择“显示预测值”、“参数估计”等，最后点击“确定”选项，就可以在数据窗口看到预测结果和图示。

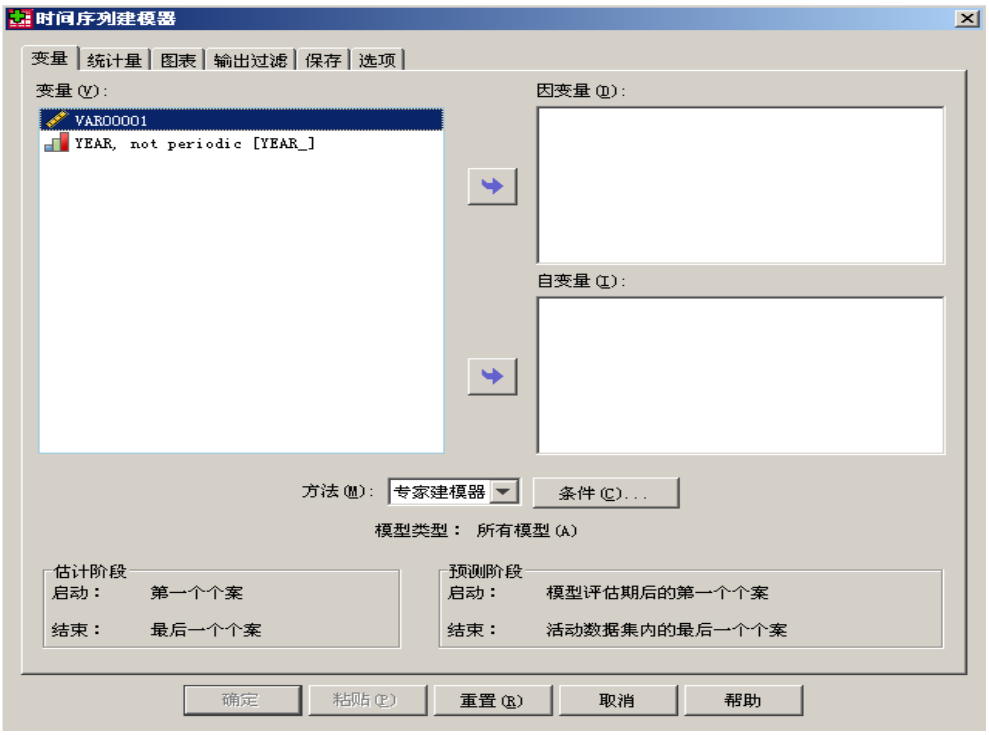


图 8-19　指数平滑预测

四、建立 ARIMA 模型

在SPSS数据窗口中依次点击“分析-预测-自相关”，即出现图8-20所示的窗口，在左边框中将要分析的时间序列变量通过箭头点进右上方框中，点击“选项”按钮，在弹出窗口中填写最大延迟期数，选择Bartlett近似值，并点击“继续”按钮，最后点击“确定”选项，就可以在数据窗口看到自相关和偏自相关函数及其图示。

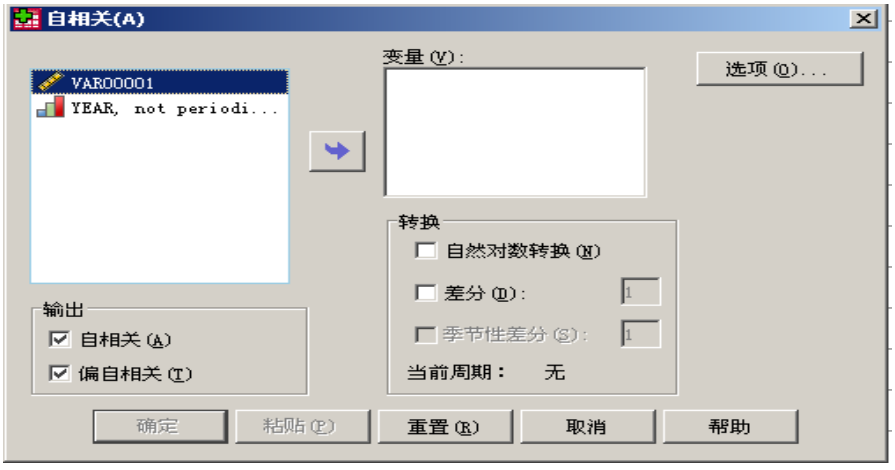


图 8-20　计算自相关和偏自相关函数

在SPSS数据窗口中依次点击“分析-预测-创建模型”，即出现图8-19所示的窗口（？），在左边框中将要分析的时间序列变量通过箭头点进右上方框中，在“方法”下拉框中选择ARIMA模型，点击“条件”按钮，在弹出窗口中填写ARIMA模型的阶数（*p*,*d*,*q*），并点击“继续”按钮，再点击右上方“选项”按钮，在弹出窗口中填入预测期，点击右上方“统计量”按钮，在弹出窗口中选择“显示预测值”、“参数估计”等，最后点击“确定”选项，就可以在数据窗口看到ARIMA模型的建模和预测结果以及图示。

小 结

时间序列分析的主要目的是揭示时间序列中蕴含的统计规律，并对未来取值进行预测。本章主要介绍了时间序列分析的几种方法及其SPSS软件的实现。时间序列变化可以分解为长期趋势、季节变动、循环变动和不规则变动的叠加或耦合，通过对时间序列进行分解既可以把握其各种因素的变动，又可以实现预测。指数平滑方法主要用于时间序列预测，包括单参数（一次）指数平滑、双参数指数平滑和三参数指数平滑，分别适用于不同的场合。单参数（一次）指数平滑适用于不包含长期趋势和季节成分的平稳时间序列预测，双参数指数平滑适用于只包含长期趋势的非平稳时间序列预测，三参数指数平滑适用于包含长期趋势和季节成分的非平稳时间序列预测。ARIMA模型是一种随机时间序列分析方法，它能更为精确地刻画时间序列变化的规律。建立ARIMA模型主要包括差分、ARMA模型识别、模型定阶、参数估计、适应性检验等步骤，适应的模型就可以用来对时间序列进行预测。这些时间序列分析方法都可以通过SPSS软件来实现。

思考和练习

1. 举例说明时间序列的含义。
2. 时间序列分析的目的是什么？
3. 时间序列分析方法有哪几类？
4. 根据影响因素，通常将时间序列分解为哪几种成分？
5. 如何测定时间序列的长期趋势？
6. 如何对时间序列进行季节变动分析？
7. 指数平滑预测方法包括哪几种？分别适合于什么场合？
8. 何为宽平稳时间序列？
9. 建立ARMA模型包括哪几个步骤？
10. 我国第一产业季度增加值数据如表8-5所示，试分析其长期趋势和季节变动。

表 8-5　我国第一产业季度增加值（亿元）

季度\年份	2001	2002	2003	2004	2005	2006
1	1556.00	1592.00	1630.70	2029.00	2287.00	3242.00
2	2960.00	3041.00	3122.90	4148.00	4420.00	5046.00

3	4183.00	4326.00	4733.00	6384.00	6803.00	7282.00
4	5910.90	5924.30	7760.50	8183.00	9208.00	9130.00

11. 运用指数平滑法借助于SPSS软件对我国2007年第一产业季度增加值进行预测。

12. 搜集北京市最近两个月每天的空气污染指数数据，对其建立适应的ARIMA模型，进行超前一期的预测，并借助实际数据计算预测误差，检验预测结果。

第九章　统计指数

为防止金融危机升级为全球性经济衰退，包括中国在内的全球主要央行于2008年10月8日联手降息，作为晴雨表的股市出现剧烈震荡的情景，走势出现分化。美国股市中道琼斯工业指数与标准普尔500指数分别下跌2.00%和1.13%，纳斯达克综合指数收低0.83%；CAC40指数尾盘大跌6.31%，伦敦FTSE 100指数与德国DAX30指数的跌幅分别为5.18%和5.88%；香港与新加坡成为当日赢家，恒生指数与海峡时报指数分别大涨3.31%和3.40%；俄罗斯RTS指数暴跌11.25%，印尼雅加达综合指数暴跌10.38%后，两国政府双双紧急终止股市交易；日经225指数走出过山车行情，小幅低开后，盘中一度大涨近200点，但尾盘依旧下挫45.83点，跌幅达0.50%，收于9157.49点，继续创出2003年6月以来的新低。我国上证指数跳空近40点高开后快速跳水，盘中几番争夺后收于2074.58点，下跌17.64点，跌幅0.84%；深证成指收于6758.22点，下跌166.27点，跌幅2.40%。上述股指走势反映出投资者仍然缺乏信心，各国政府应当继续出台干预政策，快速重塑市场信心，否则可能导致金融危机及信贷紧缩继续恶化。

股票价格指数是一种重要的统计指数，用以衡量股市的波动。此外，地区生产总值指数、居民消费价格指数、商品零售价格指数、工业品出厂价格指数、固定资产投资价格指数、房地产价格指数、农产品生产价格指数等均是常用的统计指数。这些指数在社会经济生活中起着重要的作用，是投资、管理和决策的重要参考依据。本章将阐述统计指数的编制原理及其应用。

第一节　指数编制的基本原理

一、统计指数的概念

统计指数的概念有广义和狭义之分。从广义上说，凡用来反映现象在时间上数量变动程度和方向的相对数，都称为统计指数；从狭义上讲，统计指数是一种特殊的相对数，是用来表明复杂总体数量特征综合变动的一种特殊相对数。

简单总体：构成总体的各事物在数量上能够直接加总，如钢产量。

复杂总体：构成总体的各种事物具有不同的使用价值和计量单位，各事物在数量上不能直接加总，如家电、衣服、食品的数量直接相加就没有经济意义。

二、统计指数的种类

指数从不同的角度，可以作不同的分类：

1. 按指数所反映现象的范围不同，分为个体指数和总指数

反映个别事物变动情况的相对数称为个体指数，反映多种事物综合变动情况的相对数称为总指数。例如苹果的零售价格指数是个体指数，水果的零售价格指数是总指数。广义上的统计指数既包括总指数，也包括个体指数，而狭义上的统计指数仅指总指数。

2. 按指数所反映现象的性质不同，分为数量指数和质量指数

表明现象总体规模变动情况的指数称为数量指数，例如商品销售量指数、职工人数指数等；表明现象总体内涵数量变动情况的指数称为质量指数，例如价格指数、劳动生产率指数等。

3. 按指数是否加权，分为简单指数和加权指数

简单指数是指不加权的指数，加权指数是指以显示各种事物重要性的数值为权数而计算出来的指数。例如，直接用报告期与基期的股票平均价格相比所得到的股票价格指数是简单指数；以发行量或成交量为权数计算出来的股票价格指数是加权指数。

4. 按指数的计算过程不同，分为综合指数和平均指数

综合指数是指先综合，后求比率的指数；平均指数是指先求比率，后求平均的指数。

三、简单指数的编制方法

简单指数是不经过加权计算的指数，对数量指标和质量指标都可以计算。现以价格指数为例说明简单指数法的基本原理。

1. 简单综合指数

设I_p表示简单价格指数，P_0^i和P_1^i（i=1，2，3，……n）分别表示第i种商品在基期和报告期的单位价格，则由简单综合式所编制的价格指数为：

$$I_p=\frac{\frac{1}{n}(P_1^1+P_1^2+P_1^3+\Lambda\ \Lambda+P_1^n)}{\frac{1}{n}(P_0^1+P_0^2+P_0^3+\Lambda\ \Lambda+P_0^n)}=\frac{\sum P_1}{\sum P_0} \qquad (9\text{-}1)$$

【例9.1】三种商品的价格资料如表9-1所示，利用式（9-1）式求价格指数。

表 9-1　三种商品的价格资料

商品名称	2007 年价格（　　）	2008 年价格（　　）
A	100	110
B	80	75
C	5	8

解：

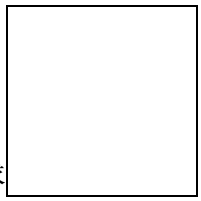
计算结果表明三种商品的价格2008年比2007年上涨了4.32%。

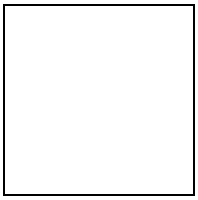
这种方法存在两个缺点：（1）易受价格较高的商品价格变动的影响；（2）没有考虑各种商品的经济重要性。如上例中虽然C商品的价格上涨了60％，但因其单价低，故对价格指数104.32％的影响不大，而且在计算价格指数时是将三种商品同等看待，不考虑三种商品的轻与重如何。因此，简单综合式价格指数在实际中较少使用。

2. 简单平均指数

简单平均价格指数是各种商品价比的简单算术平均数。其计算公式为：

（9-2）

【例 9.2】根据表 9-1 的资料，求  。

解： 

简单平均式价格指数先求比率再平均，克服了简单综合式价格指数的第一个缺点。但它仍存在着与简单综合式价格指数相同的第二个缺点，难以准确地反映物价变动的一般水平。因而，实际中常采用加权指数法编制价格指数来反映物价的变化情况。

四、加权综合指数的编制方法

加权指数用权数来区别指数中所包含商品的经济重要性大小，它比简单指数更为科学。根据计算方法的不同，加权指数区分为加权综合指数和加权平均指数两种。

加权综合指数是总指数的基本形式，它由两个总量对比而来。在计算加权综合指数时，数量指数和质量指数的权数选择上是有区别的：数量指数要用质量指标作为权数；质量指数要用数量指标作为权数。

1. 数量指数

现以产品产量指数为例，说明数量指标综合指数的编制方法和原则。产品产量指数能够综合地反映多种产品产量的变动情况，然而各种产品的使用价值和计量单位均不相同，其产品产量无法直接相加和对比。我们必须借助于另一个因素，使不能直接加总的产量指标过渡到能够相加的价值指标，这个因素就是产品出厂价格。在这里，产品产量称为指数化因素，产品出厂价格称为同度量因素。指数化因素是我们要研究其变化程度的因素，同度量因素使不能相加的产品产量过渡到可以相加的产品产值，起着同度量和权数的双重作用，那些出厂价格越高的产品，其产量与出厂价格之积越大，它对产量综合指数的影响也越大。

找到同度量因素——产品出厂价格后，还要将它固定下来，以单纯反映产品产量的变动。同度量因素既可以固定在报告期，也可以固定在基期。固定在基期时称为拉氏指数，固定在报告期时称为帕氏指数。由此得到下面两个公式：

$$\bar{I}_q = \frac{\sum q_1 p_1}{\sum q_0 p_1} \qquad (9\text{-}3)$$

$$\bar{I}_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} \qquad (9\text{-}4)$$

式中，$\bar{I}_q$ 代表产品产量综合指数，q_0、q_1 分别代表基期和报告期的产品产量，p_0、p_1 分别代表基期和报告期的产品出厂价格。

【例9.3】某企业生产三种产品，其产量及出厂价格资料见表9-2，试编制产品产量综合指数。

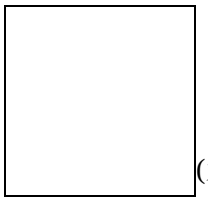
表 9-2　某企业三种产品的产量及出厂价格资料

产品名称	计量单位	产量		出厂价格（元）	
		基期	报告期	基期	报告期

A	件	100	120	440	460
B	吨	500	800	850	950
C	台	600	500	250	220

解：采用式（9-3）

$$\bar{I}_q=\frac{\sum q_1p_1}{\sum q_0p_1}=\frac{120\times460+800\times950+500\times220}{100\times460+500\times950+600\times220}=\frac{925200}{653000}=141.68\%$$

 (元)

计算结果表明，报告期三种产品的产量比基期增长了41.68%，由此增加产值272200元。

采用式（9-4）也可以得到类似结果，计算表明，报告期三种产品的产量比基期增长了38.58%，由此增加产值238800元。

从以上计算中可以看出，对于同一资料采用两个不同的产品产量综合指数公式进行计算，其产量的增长幅度和由此带来的产值增加额都是不相同的。一般认为，编制产品产量综合指数，应将产品出厂价格固定在基期。

上述编制产品产量综合指数的方法，具有普遍意义，也适用于其他数量指数。在一般情况下，编制数量指数以相关的质量指标作为同度量因素，以基期权数来加权。

2. 质量指数

现以产品出厂价格指数为例，说明质量指数的编制方法和原则。

编制产品出厂价格综合指数时，选择产品产量作为同度量因素。在这里，产品出厂价格是指数化因素，产量是同度量因素。产量起着同度量和权数的双重作用，那些产量越大的产品，其出厂价格与产量的乘积越大，它对出厂价格综合指数的影响也越大。

采用不同时期的产量加权，得到两个不同的产品出厂价格综合指数公式：

$$\bar{I}_p=\frac{\sum p_1q_1}{\sum p_0q_1} \qquad (9\text{-}5)$$

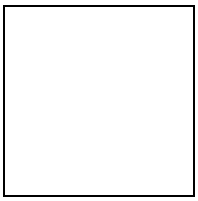
$$\bar{I}_p=\frac{\sum p_1q_0}{\sum p_0q_0} \qquad (9\text{-}6)$$

式中：$\bar{I}_p$代表产品出厂价格综合指数，其余符合同前。

【例9.4】采用表9-2的资料，编制产品出厂价格综合指数。

解：（1）采用报告期产量加权

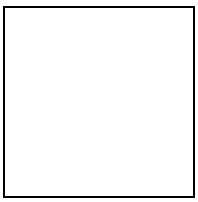
$$\bar{I}_p=\frac{\sum p_1q_1}{\sum p_0q_1}=\frac{925200}{857800}=107.86\%$$



计算结果表明，三种产品的出厂价格报告期比基期增长了7.86%，由此增加67400元产值。

（2）采用基期产量加权

$$\bar{I}_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{653000}{619000} = 105.49\%$$



计算结果表明，三种产品的出厂价格报告期比基期增长了5.49%，由此增加34000元产值。

上述编制产品出厂价格综合指数的方法具有普遍意义，适用于其它质量指数。

需要指出，式（9-3）和（9-5）称为帕氏指数公式，式（9-4）和（9-6）称为拉氏指数公式。拉氏和帕氏公式计算总指数的结果略有不同，在一般意义上无优劣之分。实际中多采用拉氏物量指数和帕氏价格指数公式来编制指数。

除了上面提到的公式之外，综合式加权指数还存在着其他加权方式，由于实际中使用不多，这里不做介绍。

五、加权平均指数

加权综合指数是根据现象内在的关系通过两个同度量后的总量指标对比编制的总指数。它的含义十分明显，既可以说明现象变动的方向和程度，又可以说明现象变动所产生的实际效果。但是编制加权综合指数需要有全面的统计资料，如编制价格指数就要有基期或报告期的数量资料，而编制数量指数则要有基期或报告期的物价资料。而实际统计工作中搜集和取得这些原始资料是十分困难的。因此，实际中除了在范围较小、种类较少、原始资料易于取得的情况下直接采用加权综合指数编制总指数之外，多数情况下需要采用加权平均指数来计算总指数。

加权平均指数是总指数的另一种形式，它是个体指数的加权平均数，其权数是值权。它与加权综合指数既有联系又有区别。其联系在于：在一定的权数下，两种指数之间有变形关系，此时两者在经济意义和数学性质上完全一致。如果计算两种指数时所依据的资料也完全相同，则它们的计算结果也相等。其区别在于：编制加权综合指数的基本问题是同度量因素问题，而加权平均指数不存在这个问题；编制加权综合指数需要全面资料和计算假定值，而加权平均指数不要求使用全面资料，也不计算假定值，具有独立的使用价值。

加权平均指数有算术平均指数和调和平均指数两种基本形式，固定权数的平均指数是其派生形式。

1. 算术平均指数

算术平均指数是以个体指数为变量，以总值（一般为基期总值）资料为权数，对个体指数加权平均计算的总指数。现以产品产量指数为例说明算术平均指数的编制方法。

当已知产品的个体产量指数（K_q）和基期各种产品的实际产值（$q_0 p_0$）资料时，则有：

$$K_q = \frac{q_1}{q_0} \tag{9-7}$$

由式（9-7）得：$q_1 = K_q q_0$

将其代入产量综合指数公式中：

$$\bar{I}_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum K_q q_0 p_0}{\sum q_0 p_0} = \sum K_{q\cdot} \cdot \frac{q_0 p_0}{\sum q_0 p_0} \tag{9-8}$$

式（9-8）是最为常用的算术平均指数公式，基期总值为权数，它可与拉氏物量指数公式相互变形。

【例9.5】根据下表资料编制算术平均物量指数。

表9-3

产品名称	计量单位	产量		基期产值(元)
		基期	报告期	
A	件	100	120	44000
B	吨	500	800	425000
C	台	600	500	150000

解：采用式（9-8）有：

$$\bar{I}_q = \frac{\sum K_q q_0 p_0}{\sum q_0 p_0} = \frac{\frac{120}{100} \times 44000 + \frac{800}{500} \times 425000 + \frac{500}{600} \times 150000}{44000 + 425000 + 150000} = \frac{857800}{619000} = 138.58\%$$

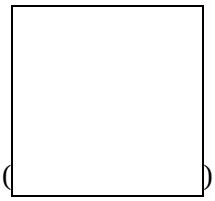
(元)

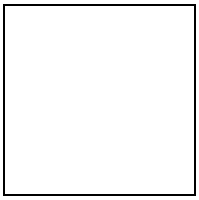
计算结果与例 9.3 相同。

2. 调和平均指数

调和平均指数是以个体指数为变量，以总值（一般为报告期总值）资料为权数，对个体指数加权平均计算的总指数。现以产品出厂价格指数为例，说明调和平均指数的编制方法。

当已知产品的个体出厂价格指数( )和报告期各种产品的实际产值


资料时，则有：



（9-9）

由式（9-9）得：$p_0=\frac{p_1}{K_p}$

将其代入价格综合指数公式中：

$$\bar{I}_p=\frac{\sum p_1q_1}{\sum p_0q_1}=\frac{\sum p_1q_1}{\sum\frac{1}{K_p}p_1q_1} \tag{9-10}$$

式（9-10）是最为常用的调和平均式价格指数公式，其权数为报告期总值，它可与帕氏价格指数公式相互变形。

【例 9.6】根据下表资料编制调和平均式价格指数。

表 9-4

产品名称	计量单位	价格（元）		报告期产值（元）
		基期	报告期	
A	件	440	460	55200
B	吨	850	950	760000
C	台	250	220	110000

解：调和平均式价格指数为：

$$\bar{I}_p=\frac{\sum p_1q_1}{\sum\frac{1}{K_p}p_1q_1}=\frac{55200+760000+110000}{\frac{440}{460}\times55200+\frac{850}{950}\times760000+\frac{250}{220}\times110000}=\frac{925200}{857800}=107.86\%$$

$$\sum p_1q_1-\sum\frac{1}{K_p}p_1q_1=925200-857800=67400(\text{元})$$

计算结果与例 9.4 相同。

将式（9-10）中的个体价格指数换为个体物量指数，得到式（9-11），它是调和平均物量指数公式，可与帕氏物量指数公式相变形。

$$\bar{I}_q=\frac{\sum p_1q_1}{\sum\frac{q_0}{q_1}p_1q_1}=\frac{\sum p_1q_1}{\sum\frac{1}{K_q}p_1q_1} \tag{9-11}$$

同理，将式（9-8）中的个体物量指数换为个体价格指数，得到式（9-12），它是算术平均式

价格指数公式，可与拉氏价格指数公式相变形。

$$\bar{I}_p = \frac{\sum \frac{p_1}{p_0} q_0 p_0}{\sum q_0 p_0} = \frac{\sum K_p q_0 p_0}{\sum q_0 p_0} \tag{9-12}$$

3. 固定权数的平均指数

计算加权算术平均指数时，在实践中常常将权数（w）相对固定，对物价或数量因素的个体指数进行加权平均而求得总指数，公式形式为：

$$\bar{I}_p = \frac{\sum \frac{p_1}{p_0} w}{\sum w} \tag{9-13}$$

$$\bar{I}_q = \frac{\sum \frac{q_1}{q_0} w}{\sum w} \tag{9-14}$$

固定权数的平均指数形式，在国内外的指数实践中得到了广泛的应用，如西方国家的工业生产指数、消费品价格指数，我国的商品零售价格总指数、职工生活费指数等都是用固定权数的平均指数形式编制的。这里的固定权数 w，是指经过调整计算后在一定时期（1 年、5 年甚至 10 年）内保持不变的权数。固定权数的资料可以根据有关的普查、抽样调查或全面统计报表资料调整计算来确定，权数的表现形式为相对数（比重），在加权平均的方法上以算术平均形式居多。

【例 9.7】有甲、乙、丙三类商品，其代表规格品的个体价格指数分别为 95%、110%和 125%。根据过去的资料经分析调整后，三类商品的分类销售额在全部销售中所占比重分别固定为 30%、25%和 45%。以此比重为权数，求三类商品的价格总指数。

解：三类商品的价格总指数为：

$$\bar{I}_p = \frac{\sum \frac{p_1}{p_0} w}{\sum w} = \frac{95\% \times 30\% + 110\% \times 25\% + 125\% \times 45\%}{30\% + 25\% + 45\%} = 112.25\%$$

第二节　常用的统计指数及其应用

一、居民消费价格指数

1. 编制居民消费价格指数的意义

居民消费价格指数（Consumer Price Index，CPI）也称消费者价格指数或生活费用指数，是综合反映各种消费品和生活服务价格的变动趋势和程度的重要经济指数。全国、省、自治区、直辖市和 500 多个市县按月度和年度编制，国家统计局定期发布居民消费价格指数。

编制居民消费价格指数可以观察和分析消费品的零售价格和服务价格变动对城乡居民实际生活的影响，为各级政府和部门掌握消费品价格状况，研究和制定价格、工资、货币和消费等政策，进行宏观经济调控提供依据。同时，它也是反映通货膨胀、进行国民经济核算的重要指标和根据。

2. 编制居民消费价格指数的方法

目前，我国的居民消费价格指数是将居民消费划分为居民日常吃、穿、用、住、行、娱等八大类，共 263 个基本分类（93SNA，国际分类标准），从中选定约 700 种商品和服务项目，采用固定加权算术平均指数法来编制的。编制居民消费价格指数一般需要经过下列步骤：

（1）对生活消费品和服务项目进行分类。现行的居民消费价格指数编制方案将生活消费品

和服务项目分为食品、烟酒及用品、衣着、家庭设备用品及服务、医疗保健及个人用品、交通和通信、娱乐教育文化用品及服务、居住等八大类；在八大类中划分中类，如交通和通信分为交通、通信两类；再在中类中进一步划分小类，如通信又分为通信工具、通信服务两类等等。于是，价格指数有了总指数、类指数和个体指数三个层次。

（2）选取代表商品和代表规格品及服务项目。商品种类繁多，要定期调查每种商品的价格水平难以办到。因此国家统计局在各类商品中选定 325 种必报商品和服务项目编制居民消费价格指数。有些商品规格繁多，其价差也较大，因此需要从中选择若干规格品作为该商品的代表，通过调查其价格，计算个体价格指数，用以反映该商品价格水平的变化情况。在选择代表规格品时要遵循下列原则：①与社会生产和人民生活关系密切；②销售量大；③生产前景较好，市场供应比较稳定；④价格变动趋势和程度对其它规格品具有较强的影响与牵引力；⑤所选的代表规格品之间差异大。

（3）选择指数公式。居民消费价格指数的汇总计算采用加权算术平均指数公式，即：

$$\text{单项商品或服务项目的价格指数}=\frac{\overline{p}_1}{\overline{p}_0} \tag{9-15}$$

式中 $\overline{p}_1$ 代表报告期平均价格，$\overline{p}_0$ 代表基期平均价格。

$$\text{类指数}=\sum k\cdot\frac{w}{\sum w} \tag{9-16}$$

在计算小类指数时，k 为个体指数，w 为各项商品或服务项目的权数；在计算中类指数时，k 为小类指数，w 为各小类商品或服务项目的权数；在计算大类指数时，k 为中类指数，w 为各中类商品或服务项目的权数；在计算居民消费价格总指数时，k 为大类指数，w 为大类商品或服务项目的权数。

（4）确定权数。权数一般用千分数表示，各类商品和服务项目的权数之和应等于 1000。权数是根据城乡居民消费模式、消费习惯，参照抽样调查原理选中的近 12 万户城乡居民家庭（城市近 5 万户，农村近 7 万户）的消费支出数据，结合其他相关资料确定的。随着人民生活水平的提高，消费结构在不断变化。为此，我国的 CPI 权数每年都做一些小调整，每五年做一次大调整。

为了取得商品和服务项目的价格资料，我国采用分层抽样的方式，在全国 31 个省（区、市）抽选出 500 多个市、县作为调查地区，从中选择了 50000 多个调查网点进行经常性的调查。国家统计局直属的全国调查系统采取定人、定时、定点的直接调查方式，由近 3000 名专职物价调查员到不同类型、不同规模的农贸市场和商店现场采集价格资料。在选择调查地区时，主要考虑分布是否合理并兼顾不同经济区域。对于与居民生活密切相关、价格变动比较频繁的商品，至少每五天调查一次价格，从而保证了 CPI 能够及时、准确地反映市场价格的变动情况。

【例 9.8】　根据表 9-5 的资料计算居民消费价格指数。

表 9-5　居民消费价格指数计算资料，2006 年=100

序号	项　　目	基期平均价格（元）$\overline{p}_0$	报告期平均价格（元）$\overline{p}_1$	个体指数（%）	权数（‰）

一	食品			?	420
	（一）粮食			?	350
	1. 细粮			?	650
	（1）大米， 梗米标一，Kg	3.5	3.71	?	600
	（2）面粉,标准,Kg	2.4	2.52	?	400
	2. 粗粮			104.80	350
	（二）副食品			125.40	450
	（三）茶及饮料			126.00	110
	（四）其他食品			114.80	90
二	烟酒及用品			100.60	60
三	衣着			95.46	150
四	家庭设备用品及服务			102.70	110
五	医疗保健和个人用品			110.43	30
六	交通和通信			98.53	40
七	娱乐教育文化用品及服务			101.26	50
八	居住			110.60	140

解：

大米价格指数=3.71/3.5=106%，面粉价格指数=2.52/2.4=105%

细粮类价格指数=106%*0.6+105%*0.4=105.6%

粮食类价格指数=105.6%*0.65+104.8%*0.35=105.32%

食品类价格指数=105.32%*0.35+125.4%*0.45+126%*0.11+114.8%*0.09=117.484%

$$\text{居民消费价格指数}=\frac{\sum kw}{\sum w}=\frac{108796}{1000}=108.796\%$$

或：

消费价格指数= 117.484%*0.42+100.60%*0.06+……+110.60%*0.14=108.796%

需要指出，商品零售价格指数与居民消费价格指数的编制方法和步骤基本相同，只是两者包括的内容有所不同。商品零售价格指数是反映一定时期内城乡商品零售价格变动趋势和程度的相对数，反映的是工业、商业、餐饮业和其它零售企业向城乡居民、机关团体出售生活消费品和办公用品价格的综合变动，其变动直接影响到城乡居民的生活支出和国家的财政收入，影响居民购买力和市场供需的平衡，影响到消费与积累的比例关系。因此，该指数可以从一个侧面对上述经济活动进行观察和分析；而居民消费价格指数是反映城乡居民支付生活消费品和服务项目消费价格的综合变动。

3. 居民消费价格指数的应用

在很多情况下我们从统计年鉴中直接得到的宏观经济总量数据，如 GDP、总消费、总投资等等都是以当年价格计算的，而我们在经济分析中需要首先剔除价格因素的影响，这时就需要用相应的价格指数来“缩减”现价指标。在对不同的序列进行缩减时，需要选用与相应序

列一致的价格指数序列。例如对总消费进行缩减时，应选用居民消费价格指数；对固定资产投资进行缩减时应当选用“固定资产投资价格指数”，等等。如果没有完全一致的价格指数，可以选择一个比较接近的价格指数。

经济学中一般把按当年价格计算的指标称为名义（nominal）指标，将以可比价格计算的指标称为实际（real）指标。

$$\text{实际指标} = \frac{\text{名义指标}}{\text{相应的定基价格指数}}$$

居民消费价格指数主要用以观察和分析消费品的零售价格和服务项目价格变动对城乡居民实际生活费支出的影响程度。居民消费价格指数可以用来对现价消费数据进行缩减。

例【9.9】 已知我国 2000 年—2007 年当年价格的最终消费支出和环比居民消费价格指数（表 9-6），计算以 2000 年价格表示的各年的最终消费。

表 9-6 我国 2000 年—2007 年最终消费支出和环比居民消费价格指数

年份	最终消费支出（亿元）	居民消费价格指数（%）
2000	61516.0	100.4
2001	66878.3	100.7
2002	71691.2	99.2
2003	77449.5	101.2
2004	87032.9	103.9
2005	97822.7	101.8
2006	110595.3	101.5
2007	128444.6	104.8

资料来源：《中国统计年鉴》（2008）

解：以 2000 年为基期的定基指数等于相应环比指数的连乘积（见表 9-7 第 3 栏），计算过程与结果列在表 9-7 中。

表 9-7 以 2000 年价格表示的各年的最终消费：计算步骤和结果

年份	最终消费支出	居民消费价格指数	定基价格指数（%）	最终消费支出（亿元）
	（亿元）	（%）	2000 年=100	2000 年价格
	（1）	（2）	（3）	（4）=（1）/（3）
2000	61516	100.4	100.00	61516.00
2001	66878.3	100.7	100.70	66413.41
2002	71691.2	99.2	99.89	71770.15
2003	77449.5	101.2	101.09	76614.40
2004	87032.9	103.9	105.03	82864.80
2005	97822.7	101.8	106.92	91491.49
2006	110595.3	101.5	108.52	101912.37
2007	128444.6	104.8	113.73	112938.19

二、生产指数

为了反映国民经济的发展变化以及经济周期的变动，世界各国都在编制各种生产指数（Production Index）。生产指数是指反映生产的发展趋势及变动程度的相对数，它是一种物

量指数，一般采用代表产品产量个体指数的加权算术平均指数公式来编制。即：

$$\text{生产物量指数} = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} \qquad (9\text{-}17)$$

由于研究目的和范围不同，生产指数可分为工业生产指数、农业生产指数、建筑业生产指数等。

1. 工业生产指数

工业生产指数（Industrial Production Index）反映工业生产的发展变化情况，是经济动态最敏感的指数，可以作为经济分析和商情预测的尺度。世界各国都十分重视并定期编制工业生产指数，它是西方国家普遍用来计算和反映工业发展速度的指标，也是景气分析的首选指标。但是，各国的工业生产指数包括的生产部门范围不尽相同，选择的代表产品也有差异。此外，在计算产量时，有以生产量衡量，也有以销售量衡量的。

工业生产指数原则上采用修改后的拉氏物量指数公式（9-18）。不过，在实际编制工业生产指数时，一般采用固定权数的加权算术平均指数公式（9-19），即：

（9-18）

（9-19）

式中 p_n 为可比价格，q_0、q_1 分别表示基期和报告期工业产品的数量，$p_n q_n$ 为权数，是可比价格增加值（基期增加值）。

可见，计算工业生产指数有两个关键要素，一个是产品产量，另一个是权数。在确定产品产量时不可能涵盖全部产品，只能适度选取代表产品，选取的产品数量太少不足以反映产品结构，反之又会造成工作量过大；权数是对产品的个体指数在生产指数形成过程中的重要性进行界定的指标，权数正确与否直接影响总指数的准确程度。在工业总体中比重大的产品或行业其权数大一些，对生产指数的影响也大，即权数可用各种产品增加值在该类产品增加值中的比重或各部门增加值在全部工业增加值中的比重表示。编制工业生产指数的基期一般五年变更一次，每变更一次基期，有关年份的生产指数都要按新的基期年份调整，以保证资料的可比性。

【例 9.10】某地区工业增加值资料见下表。试编制工业生产指数，反映 1949－1995 年工业发展情况。

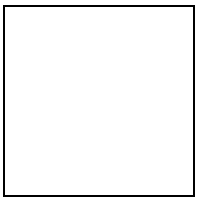
表 9-8　某地区工业增加值资料　单位：亿元

	1949 年	1957 年	1971 年	1981 年	1990 年	1995 年
1952 年不变价格	350	680				
1957 年不变价格		650	2000			
1970 年不变价格			2050	4200		
1980 年不变价格				4400	23500	
1990 年不变价格					24800	81600

解：我们不能直接用 1995 年的总产值对比 1949 年的总产值，以反映 1949－1995 年该地区工业生产的发展情况，因为两者的价格是不可比的。我们首先应当根据各交替年（即采用新旧不变价格的年份）的资料计算出换算系数，将两年的工业增加值价格调整一致时方可直接对比。换算系数的计算公式为：

$$\text{换算系数}=\frac{\text{交替年按新不变价格计算的增加值}}{\text{交替年按旧不变价格计算的增加值}}$$

所以，1995 年对 1949 年工业生产指数



上式中 379.12 即为按 1990 年不变价格计算的 1949 年工业增加值。

世界上大多数国家十分重视工业生产指数的编制，除了编制表明全部工业生产发展速度的总指数之外，还编制各种分类指数以表明各类工业生产的发展变化情况，并分析其对工业生产总变动的影响。分类标准可以是国民经济部门、工业产品耐用性、生产资料与消费资料等。我国在 1997 年之前一直沿用工业不变价总产值的计算方法来计算工业生产指数。随着我国社会主义市场经济体制的建立和统计管理体制的不断健全，原有的工业总产值计算方法渐渐偏离了工业经济发展的实际。国家统计局决定从 1997 年开始以生产指数法代替现行的不变价格法试算工业生产指数，但到目前为止尚未正式推出该指数。

2. 农业生产指数

农业生产指数（Agricultural Production Index）是反映农业生产的发展趋势和程度的相对数，表明一定时期全部农畜产品实物产量的增减变化，一般包括了所有重要的农产品。农业生产指数采用固定权数的加权算术平均指数公式，即按式（9-19）计算，以各代表产品的可比价格增加值作为权数。

由于农业生产周期较长，一般都是一年编制一次农业生产指数。世界各国除编制农业生产总指数之外，还编制各种分类指数，如农作物生产指数、牲畜家禽生产指数、农产品生产指数等。农业受自然条件的影响比较大，因而有些国家除了以某一年度作为基期以外，还以某几个年度的平均数作为基期。联合国还在各国农业生产指数的基础上，编制和发表全世界和分地区的农业生产指数。由于农产品中的食用部分直接关系到人民的生活消费，所以各国十分重视农产品生产指数，并对其中的食用部分编制食物生产指数，以反映农业生产的各类食用产品产量的增减变化情况。

三、股票价格指数

1. 股票价格指数的概念和作用

股票价格指数（Stock Price Index）简称股价指数，它是反映股票市场（简称股市）价格变动趋势和程度的综合性指标。由于政治、经济、股票供求关系以及投资者预期等诸多因素的影响，股票价格经常处于变动之中。为了能够综合地反映这种变化，世界各地的股票市场都编制股票价格指数。

通过观察个股股票价格指数的升降，投资者可以了解各个股票价格的变动情况，在此基础上结合股票价格总指数的跌涨，即可迅速把握股市总体价格水平的变化趋势。此外，股市是经济发展的“晴雨表”，股票价格指数的变动也是反映一个国家或地区政治、经济发展状况的

灵敏信号。正如本章开篇中所述，从 2006 年春季开始逐步显现，2007 年 8 月开始席卷美国、欧盟和日本等世界主要金融市场的美国“次贷危机”，严重影响了这些国家的经济发展，这种现象在股市上反映得十分突出。从这里我们可以充分认识到股价指数的变化能够十分灵敏地映射出某一国家或地区的经济运行状况。

2. 股票价格指数的编制要求及方法

股票价格指数的编制应符合客观、准确、灵敏的要求。在编制过程中，要处理好以下几个问题：

（1）正确选择具有代表性的若干种股票（又称样本）作为计算对象。在选择样本股票时，必须综合考虑行业分布、市场影响力、股票等级、数量适当等因素。

（2）采用恰当的计算方法进行编制计算。计算方法应具有高度适应性和较好的敏感性，能对不断变化的股市行情作出相应的调整或修正。此外，还需要有科学的计算依据和手段，计算依据的时间间隔必须一致，计算手段也须不断完善，使股价指数能客观、准确地反映股市行情。

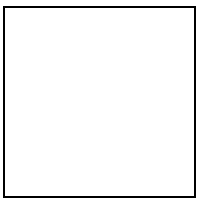
（3）选好计算股价指数的基期。基期应具有较好的代表性和均衡水平，使各个不同时期的股价指数具有可比性，如实反映股市活动的变化情况。

一般说来，全世界各个股市的股票价格指数的具体编制过程都不完全相同，采用的编制方法也多种多样，但基本的编制方法是简单指数法和加权指数法，主要用式（9-1）、（9-2）、（9-5）和（9-6）计算。公式中的 q（即权数）既可以是发行量，也可以是成交量。

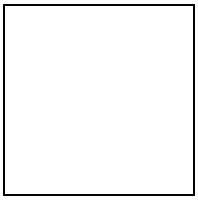
从国际上来说，常见的几种代表性的股票价格指数包括：美国的道琼斯股票价格平均指数、标准普尔股票价格指数、纽约证券交易所股票价格指数、纳斯达克综合指数，欧洲的伦敦 FTSE 100 指数、法国 CAC40 指数以及德国 DAX30 指数，亚洲的香港恒生指数、日经指数、上证指数、深证指数、澳大利亚 S&P/ASX 200 指数、韩国首尔综合指数、印尼雅加达指数、马来西亚 KLSE 综合指数等。这里我们主要介绍一下我国几种股票指数的计算方法。

（1） 上证综合指数（简称上证指数，Shanghai Composite Index）

上证综合指数（简称上证指数，Shanghai Composite Index）由上海证券交易所编制，反映上海证券交易所上市股票价格变动的动态指标。上证指数以开业时上市的全部 8 种股票为样本，以正式开业日 1990 年 12 月 19 日为基期，采用市价总额加权计算得到。其计算公式为：



具体的计算方法是，基期和报告期 8 种股票的市价分别乘以发行量，相加后求出基期和报告期市价总值，再相除乘以 100 得出报告期股价指数。如遇报告期前一日出现增资扩股或新增股票（删除旧股），则须相应修正，公式中的分母成为新基准市价总值。



上证指数采用股票发行量作为权数，基本符合目前上海证券市场的股价状况，从而成为分析上海股市变化的重要依据。

随着上市品种的逐步丰富，上海证券交易所在这一综合指数的基础上，从 1992 年 2 月起分别公布 A 股指数和 B 股指数，1993 年 5 月 3 日起正式公布工业、商业、地产业、公用事业

和综合五大类分类股价指数。

（2）深证成分股指数与深证综合指数

深证成分股指数（简称深证成指，Shenzhen Component Index）与深证综合指数（简称深证综指，Shenzhen Composite Index）均是深圳证券交易所编制和发布的股价指数，只是前者是一种成份股指数，是从上市的所有股票中抽取具有市场代表性的 40 家上市公司的股票作为计算对象，并以流通股为权数计算得出的加权股价指数，综合反映深交所上市 A、B 股的股价走势；后者以深圳证券交易所挂牌上市的全部股票为计算范围，以计算日总股本数（即发行量）为权数加权计算股价指数，综合反映深交所全部 A 股和 B 股上市股票的股价走势。此外，深证综指以 1991 年 4 月 3 日为基日，基日指数定为 100 点，1991 年 4 月 4 日开始发布；深证成指的基日为 1994 年 7 月 20 日，基日指数定为 1000 点，于 1995 年 1 月 23 日开始试发布，1995 年 5 月 5 日正式启用。

深证成指的基本公式为：

$$\text{股价指数}=\frac{\text{报告期成份股总市价}}{\text{基期成份股总市价}}\times 1000$$

当有新股上市时，在其上市后的第二天即纳入成份股计算。当某一成份股暂停买卖时，便将其剔除于指数计算之外，若有成份股在交易时间突然停盘，将取其最后成交价格计算即时指数，直到收市后再进行必要的调整。

上证指数和深证指数是反映我国沪深股票市场股票价格走势的基本指标。除此之外，目前，反映上海证券交易所股票价格变动的指数还有上证 A 股指数、上证 B 股指数、上证 180、上证 50 以及上证工业指数、商业指数、地产指数、制造指数、纺织指数等行业指数；反映深圳证券交易所股票价格变动的指数有深证 A 股指数、深证 B 股指数等综合指数，以及成份 A 股指数、成份 B 股指数、工业类指数、商业类指数、金融类指数、地产类指数、公用事业类指数、综合企业类指数等分类成份股指数。这些指数的编制原理和方法基本相同，只是编制和发布单位、入选股票样本、股指基日和基日指数等有所区别，这里不再赘述。

（3） 沪深 300 指数

沪深 300 指数（Hushen 300 Index）是由上海证券交易所和深圳证券交易所共同开发的中国 A 股市场股价指数，用以综合反映沪深市场 A 股整体走势的跨市场股票价格指数。它的样本选自沪深两个证券市场, 覆盖了沪深市场六成左右的市值，其成份股票为我国 A 股市场中代表性强、流动性高、交易活跃的主流投资股票，涉及银行、钢铁、石油、电力、煤炭、水泥、家电、机械、纺织、食品、酿酒、化纤、有色金属、交通运输、电子器件、商业百货、生物制药、酒店旅游、房地产等数十个主要行业的龙头企业，具有良好的市场代表性，能够反映 A 股市场总体发展趋势。

指数成分股的选择空间是：上市交易时间超过一个季度；非 ST，*ST 股票，非暂停上市的股票；公司经营状况良好，最近 1 年无重大违法违规事件、财务报告无重大问题；股票价格无明显的异常波动或市场操纵；专家委员会没有异议。入样股票的选取标准是：规模大、流动性好。为此，首先要计算样本空间股票最近 1 年（新股为上市以来）的日平均总市值、日均流通市值、日均流通股份数、日均成交金额和日均成交股份数 5 个指标，再将上述指标的比重按 2:2:2:1:1 进行加权平均，然后将计算结果从高到低排序，选取排名在前 300 的股票。此外，依据样本稳定性和动态跟踪相结合的原则，沪深 300 指数每半年调整一次成份股，每次调整比例一般不超过 10%。样本调整设置缓冲区，排名在 240 名内的新样本优先进入，排名在 360 名之前的老样本优先保留。当样本股公司退市时，自退市日起，从指数样本中剔除，由过去最近一次指数定期调整时的候选样本中排名最高的尚未调入指数的股票替代。

沪深 300 指数以 2004 年 12 月 31 日为基日，以该日 300 只成份股的调整市值为基期，基期

指数定为 1000 点，自 2005 年 4 月 8 日起正式发布。计算公式为：

$$\text{股价指数} = \frac{\text{报告期成份股的调整市值}}{\text{基期成分股的调整市值}} \times 1000$$

其中，调整市值 = ∑（市价×调整股本数），调整股本数采用分级靠档的方法对成份股股本进行调整。例如，某股票流通股比例（流通股本 / 总股本）为 7%，低于 20%，则采用流通股本为权数；某股票流通比例为 35%，落在区间（30，40）内，对应的加权比例为 40%，则将总股本的 40%作为权数。

第三节　指数体系与因素分析

一、指数体系

不论是自然界，还是社会生活中，许多现象之间都存在着相互联系和相互影响的客观关系，事物之间的这种必然联系可以从数量上加以测定。在统计中，我们把经济上有联系、数量上保持一定关系的若干指数所形成的整体称为指数体系。例如，商品销售额指数＝商品销售量指数×商品销售价格指数，总产值指数=产量指数×产品价格指数，等等。

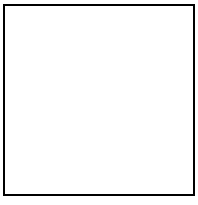
指数体系有两个主要作用：（1）根据指数体系中各指数之间的联系进行相互推算。例如，我国商品销售量指数往往就是用商品销售额指数除以商品销售价格指数推算出来的；（2）分析研究现象总变动中各个因素变动对其的影响方向和程度。

二、总值变动的因素分析

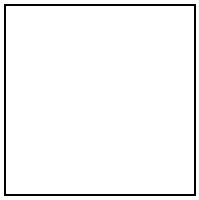
利用指数体系可以对总值变动的影响因素进行分析。总值变动与各因素的关系表现在两个方面：（1）总指指数等于各因素指数的乘积；（2）总值变动的绝对额等于各因素导致的变动额之和。

在总值变动因素分析中，指数体系中可以包含两个或者更多个因素指数。这里我们用两个因素的情况加以说明。

在两个因素的指数体系中，两个因素指数通常一个是数量指数、另一个是质量指数，而且必须一个是拉氏指数，另一个是帕氏指数。最常用的分解公式为：价值指数=拉氏数量指数×帕氏价格指数。用公式表示：



（9-20）



（9-21）

在两因素的分解方案中，也可以将价值指数分解为帕氏数量指数与拉氏价格指数的乘积。但为了统一，通常采用前一种分解方案。

总值变动的因素分析包括以下内容：计算被分析指标的总变动程度和绝对额；计算各因素的

变动程度和对绝对额的影响；对影响因素进行综合分析，即总变动程度等于各因素变动程度之连乘积，总变动绝对额等于各因素变动影响绝对额之和。

【例 9.11】　根据表 9-2 的资料，进行两因素分析。

解：（1）总产值的变动情况分析

总产值指数＝

（2）分析产量(　　)变动对总产值的影响

产品产量指数＝

（3）分析产品出厂价格(　　)变动对总产值的影响

产品出厂价格指数=

计算结果表明，总产值报告期比基期增长 49.47%，增加绝对额为 306200 元。其中，产品产量报告期比基期增长 38.58%，从而使产值增加 238800 元；产品出厂价格报告期比基期增长 7.86%，从而增加产值 67400 元。即：

149.47%=138.58%×107.86%

306200 元=238800 元+67400 元

小 结

本章主要介绍统计指数的编制原理及其应用。内容包括：

1. 统计指数的概念、特点和作用。统计指数的概念有广义和狭义之分。从广义上说，凡用来反映现象在时间上数量变动程度和方向的相对数，都称为统计指数；从狭义上讲，统计指数是一种特殊的相对数，用来反映不能直接加总和对比的多因素所构成的现象的综合变动。统计指数具有相对性、综合性和平均性的特点，能够综合反映现象的变动方向和变动程度，能够分析受多种因素影响的现象的总变动中各个因素的影响方向和影响程度。

2. 统计指数的种类。统计指数按所反映现象的范围不同，分为个体指数和总指数；按所反映现象的性质不同，分为数量指数和质量指数；按是否加权，分为简单指数和加权指数；按计算过程不同，分为综合指数和平均指数；按采用的基期不同，分为定基指数和环比指数。

3. 统计指数的编制方法。(1) 简单指数。简单指数是不经过加权计算的指数，分为简单综合指数与简单平均指数两种。(2) 加权综合指数。分为数量指数和质量指数，其编制有两种典型的方法：一种是将同度量因素固定在基期，称为拉氏指数公式；另一种是将同度量因素固定在报告期，称为帕氏指数公式。在一般情况下，编制数量指数以相关的基期质量指标作为同度量因素，而编制质量指数以相关的报告期数量指标作为同度量因素。(3) 加权平均指数。它是总指数的另一种形式，是为了克服编制加权综合指数容易受到客观资料的限制而采用的指数形式。加权平均指数是个体指数的加权平均数，其权数是值权，分为加权算术平均指数和加权调和平均指数。在计算加权平均指数时，在实践中常常将权数相对固定，形成固

定权数的平均指数，其用途更加广泛。

4. 常用的统计指数。主要介绍了居民消费价格指数、生产指数以及股票价格指数。居民消费价格指数是综合反映各种消费品和生活服务价格的变动趋势和程度的重要经济指数；生产指数是指反映生产的发展趋势及变动程度的相对数，它是一种物量指数，一般采用代表产品产量个体指数的加权算术平均指数公式来编制；股票价格指数是反映股市价格变动趋势和程度的综合性指标，我国有代表性的股票价格指数包括上证综合指数、深证成分股指数、深证综合指数、沪深300指数等。

5. 指数体系及其应用。指数体系是指经济上有联系、数量上保持一定关系的若干指数所形成的整体。利用指数体系，可以对总值的变动进行两因素或多因素分析。

思考与练习

1. 何谓统计指数？它有哪些特性和作用？
2. 简单指数分哪两类？它们是如何编制的？各有什么优缺点？
3. 加权指数分哪两类？它们之间有何区别和联系？
4. 编制加权综合指数的一般原则是什么？
5. 何谓指数体系？它有哪些作用？
6. 何谓居民消费价格指数？它有什么作用？
7. 何谓生产指数？
8. 何谓股票价格指数？它有什么作用？我国和其他国家与地区有代表性的股票价格指数有哪几种？
9. 某股票市场的四种股票价格资料如表9-9。分别利用公式（9.1）至（9.4）编制该股市股票价格指数，并比较计算结果。

表9-9 某股票市场的四种股票价格

股票	价格（元）	
	基期	报告期
A	80	114
B	135	160
C	100	120
D	110	132

10. 某企业产品产量和单位成本资料如表9-10。

表9-10 某企业产品产量和单位成本资料

产品名称	计量单位	产量	单位成本（元）

		基期	报告期	基期	报告期
甲	套	100	138	1100	1050
乙	件	90	90	1000	1000
丙	吨	70	60	3000	3100

试求：

（1）各种产品的产量个体指数；

（2）各种产品的单位成本个体指数；

（3）总成本指数；

（4）分析由于产品产量、单位成本变化对总成本的影响。

11. 某农副产品收购站三种鲜果产品收购资料如表9-11。

表9-11　某农副产品收购站三种鲜果产品收购资料

果品	计量单位	价格（元）		10 月份收购额（万元）
		9 月份	10 月份	
A	公斤	15	20	1000
B	公斤	20	25	900
C	公斤	10	8	440

要求：

（1）计算三种鲜果产品的收购价格指数；

（2）分析由于收购价格变动对农民收入的影响。

12. 某企业三种商品的销售额和销售量动态资料如表9-12。

表9-12　某企业三种商品的销售额和销售量动态资料

商品名称	计量单位	2007 年销售额（万元）	2008 年比 2007 年销售量增长%
甲	公斤	240	10
乙	件	460	20
丙	台	300	15

要求：

（1）计算三种商品的销售量总指数；

（2）分析由于销售量变化对企业销售额的影响。

13. 简单推算：

（1）已知生产费用总额增长20%，产品产量增长25%，试计算单位产品成本的变动程度。

（2）在价格上涨10%的情况下，商品销售量又增加了5%，那么，商品销售额增长了多少?

第十章　主成分分析与因子分析

一个有经验的裁缝加工一件上衣，需要测量上体长、手臂长、胸围、颈围、肩宽、腰围等14个指标，但在批量生产中，测量每个人的14个指标是不可能的，怎么办呢？人们发现，这14个指标之间具有相关性，如果从这些指标中构造出少数几个指标，只要根据这少数的几个主要指标加工出的上衣就能适合大多数人的体型，即这少数几个指标充分把握了上衣的主要特征。事实上，采用主成分分析和因子分析便能找到两个不相关的指标“型和号”，根据这两个指标加工出的上衣，特体除外，95%以上的人都能穿。从14个指标中构造出两个不相关的指标的过程就称为降维。在现实中类似的降维事例是很多的，在统计学中主要利用因子分子和主成分分析实现对数据的降维处理。这一章我们将介绍因子分析和主成分分析如何实现降维，以及在SPSS中如何实现这两种方法。

第一节　主成分分析

一、主成分分析的基本思想

1. 基本思想和数学模型

在对某一事件进行研究时，常常会涉及到与此相关的多个变量，而这些变量之间往往存在着相关性，很多的变量以及变量间的相关性大大增加了研究的复杂程度。主成分分析就是在解决上述问题过程中产生的，目的在于用少数几个不相关的主成分来代表原来的多个变量，以方便我们对问题的分析。

所谓的主成分就是指多个变量的线性组合，不同的主成分之间相互无关。假设有n个样品，每个样品有p个变量分别为 $X_1, X_2, ..., X_p$ ，则主成分的个数最多可以有p个，用公式表示为：

$F_i = a_{1i}X_1 + a_{2i}X_2 + ... + a_{pi}X_p \quad i = 1,2,..., p$ 。

方程应满足下列条件：

（1） $a_{1i}^2 + a_{2i}^2 + ... + a_{pi}^2 = 1$。

（2） F_i 与 F_j （ $i \neq j; i, j = 1,2,..., p$ ）不相关。

（3） F_1 到 F_p 方差依次递减。

第一个条件对系数加以限制使得方差不会任意增大。如果不对系数加以限制，方差可以趋于无穷大就变得没有意义了，同时第一个条件也使得每个主成分都是原始变量的凸函数。
第二个条件也是主成分分析的灵魂所在，进行主成分分析的依据就是原始变量间的相关性，用不相关的主成分表达相关的原始变量的信息来实现降维，也是在提取主成分时不会提取重复的信息。
第三个条件中每个主成分的方差衡量了每个主成分所能表示的原始变量信息的多少， F_1 到 F_p 方差依次递减为提取主成分提供了方便。在提取主成分时，可以根据方差的大小确定主成分的个数。这个条件可以保证被提取的每个主成分都比不被提取的主成分包含更多的原始变量信息，以保证在降维的同时最大限度的提取原始信息。

2. 主成分的几何意义

通过前面的介绍，了解到主成分在代数观点上就是原始变量的线性组合，而在几何上可表示为是对原始变量进行线性变换，从而实现以较少的维度表达大部分原始变量的信息。为了方便在坐标中表现降维的几何过程，下面以二元为例来说明主成分的几何意义。
设有*n*个样品，每个样品有两个变量 X_1 和 X_2，这样画出这*n*个样品的散点图如图10-1。

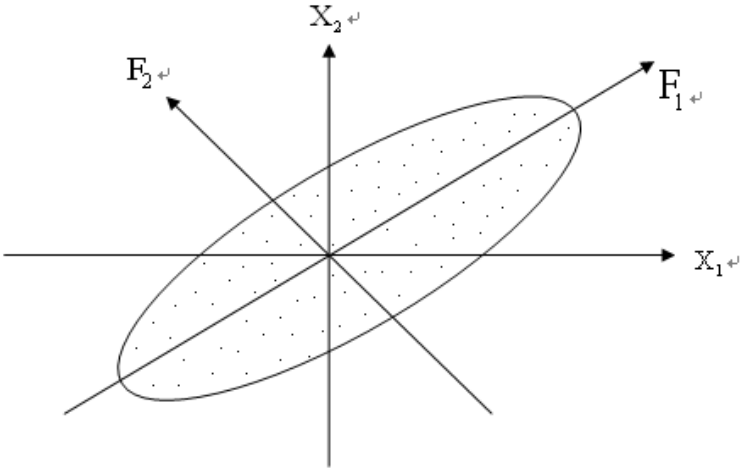


图10-1　二维空间主成分示意图

从图10-1可以看到数据都集中在椭圆的区域内。在水平轴 X_1 和 X_2 的两个方向上，我们看到数据点是很分散的，把原始变量作线性变换就相当于把原坐标轴进行旋转，把坐标轴旋转到与椭圆的长短轴平行的 F_1 和 F_2 方向上。相对于长轴如果短轴上的波动可以忽略时，就可以只用长轴的变量来表示原始两个变量的信息，即：把原来的 X_1 和 X_2 两个变量信息只用 F_1 来表示，也就完成了降维的过程。一种极端的情况是：如果短轴趋近于0时，只用一个长轴变量就可以提取几乎所有的原始信息。
在多元的情况下是类似于二元的多元空间中椭球体的主轴问题，计算要比二元的情况复杂的多，但思想是相同的，在计算机的辅助下可以很简单的实现对多元的降维，具体的实施在下面的软件实现中会有详细的介绍。

二、主成分分析的步骤和结果分析

主成分可以按以下步骤计算得出：计算原始变量的相关系数矩阵R；计算相关系数矩阵R的特征值，并按从大到小的顺序排列，记为$\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_p$；计算特征值对应的特征向量，即为主成分$F_1$到$F_p$相应的系数。

把原始变量的值代入主成分表达式中，可以计算出主成分得分。注意在计算主成分得分时需要先对原始变量进行标准化。得到各主成分得分后，可以把各个主成分看作新的变量代替原始变量，从而达到降维的目的。

对于第k个主成分，其对方差的贡献率为$\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$。前$k$个主成分贡献率的累计值称为累计贡献率。

主成分个数的确定通常有两种方式：（1）根据大于1的特征值的个数确定主成分的个数；（2）根据主成分的累计贡献率确定主成分的个数，使累计贡献率>85%或者其他值。最常见的情况是主成分的个数为2-3个。

下面通过例子来说明主成分分析如何帮助我们实现数据的降维。

【例10.1】某公司欲从48名应聘者中选出6人，公司拟定15条评价标准，根据应聘者回答问题的情况分别进行评分，评分取值范围为0～10分，0分为最低评价，10分为最高评价。部分应聘者的得分情况参见表10-1（数据文件：应聘.sav）。

表10-1　应聘者测验得分表

求职者编号	简历格式	外观	学术能力	兴趣爱好	自信心	洞察力	诚信度	销售能力	工作经验	工作魄力	志向抱负	理解能力	潜能	求职渴望度	适应力
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5
5	6	8	8	8	4	4	9	5	8	5	5	8	8	7	7
6	7	7	7	6	8	7	10	5	9	6	5	8	6	6	6
7	9	9	8	8	8	8	8	8	10	8	10	8	9	8	10
8	9	9	9	8	9	9	8	8	10	9	10	9	9	9	10
9	9	9	7	8	8	8	8	5	9	8	9	8	8	8	10
10	4	7	10	2	10	10	7	10	3	10	10	10	9	3	10
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

表10-1的数据包含简历格式、外貌、学习能力、兴趣爱好、自信心、洞察力、诚信度、推销能力、工作经验、工作魄力、志向抱负、理解力、潜能、求职渴望度、适应力等15个指标，在SPSS中应用主成分分析得到分析结果如下。

表10-2的相关矩阵给出了各个原始变量间的相关系数，从表中可以看出部分原始变量间存在很大的相关性，最大相关系数为0.883，进行主成分分析是合理的。

表10-2　相关矩阵

	简历格式	外貌	学习能力	兴趣爱好	自信心	洞察力	诚信度	销售能力	工作经验	工作魄力	志向抱负	理解能力	潜能	求职渴望度	适应力
简历格式	1.00	0.24	0.04	0.31	0.09	0.23	-0.11	0.27	0.55	0.35	0.29	0.34	0.37	0.47	0.59
外貌	0.24	1.00	0.12	0.38	0.43	0.37	0.35	0.49	0.14	0.34	0.55	0.51	0.51	0.28	0.38
学习能力	0.04	0.12	1.00	0.00	0.00	0.00	-0.03	0.06	0.27	0.09	0.04	0.20	0.29	-0.32	0.14
兴趣爱好	0.31	0.38	0.00	1.00	0.30	0.48	0.65	0.36	0.14	0.39	0.35	0.50	0.61	0.69	0.33
自信心	0.09	0.43	0.00	0.30	1.00	0.81	0.41	0.80	0.02	0.70	0.84	0.72	0.67	0.48	0.25
洞察力	0.23	0.37	0.00	0.48	0.81	1.00	0.36	0.82	0.15	0.70	0.76	0.88	0.78	0.53	0.42
诚信度	-0.11	0.35	-0.03	0.65	0.41	0.36	1.00	0.24	-0.16	0.28	0.22	0.39	0.42	0.45	0.00
销售能力	0.27	0.49	0.06	0.36	0.80	0.82	0.24	1.00	0.26	0.82	0.86	0.78	0.75	0.56	0.56
工作经验	0.55	0.14	0.27	0.14	0.02	0.15	-0.16	0.26	1.00	0.34	0.20	0.30	0.35	0.25	0.69
工作魄力	0.35	0.34	0.09	0.39	0.70	0.70	0.28	0.82	0.34	1.00	0.78	0.81	0.79	0.61	0.62
志向抱负	0.29	0.55	0.04	0.35	0.84	0.76	0.22	0.86	0.20	0.78	1.00	0.78	0.77	0.54	0.44
理解能力	0.34	0.51	0.20	0.50	0.72	0.88	0.39	0.78	0.30	0.81	0.78	1.00	0.85	0.53	0.53
潜能	0.37	0.51	0.29	0.61	0.67	0.78	0.42	0.75	0.35	0.79	0.77	0.85	1.00	0.54	0.57
求职渴望度	0.47	0.28	-0.32	0.69	0.48	0.53	0.45	0.56	0.25	0.61	0.54	0.53	0.54	1.00	0.40
适应力	0.59	0.38	0.14	0.33	0.25	0.42	0.00	0.56	0.69	0.62	0.44	0.53	0.57	0.40	1.00

表10-3是SPSS输出的表格之一，称为“解释的总方差”。表中的第一列就是各个主成分对应的特征值，并给出了各主成分对解释原始变量总方差的贡献率和累积贡献率。从表中可以看出，前四个成分共解释原始变量总方差的81.5%。

表10-3　解释的总方差

成分	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.514	50.092	50.092	7.514	50.092	50.092
2	2.056	13.709	63.801	2.056	13.709	63.801

3	1.456	9.705	73.506	1.456	9.705	73.506
4	1.198	7.986	81.492	1.198	7.986	81.492
5	.739	4.928	86.420			
6	.495	3.297	89.717			
7	.351	2.342	92.059			
8	.310	2.066	94.125			
9	.257	1.713	95.838			
10	.185	1.233	97.071			
11	.153	1.018	98.088			
12	.098	.650	98.739			
13	.089	.592	99.331			
14	.065	.431	99.762			
15	.036	.238	100.000			

提取方法：主成分分析。

图10-2的碎石图显示的是特征值的大小。按照特征值大于1的原则，本例中看到提取了前四个主成分。前四个主成分已经足够描述应聘者的素质水平，共解释了总方差的81.5%。

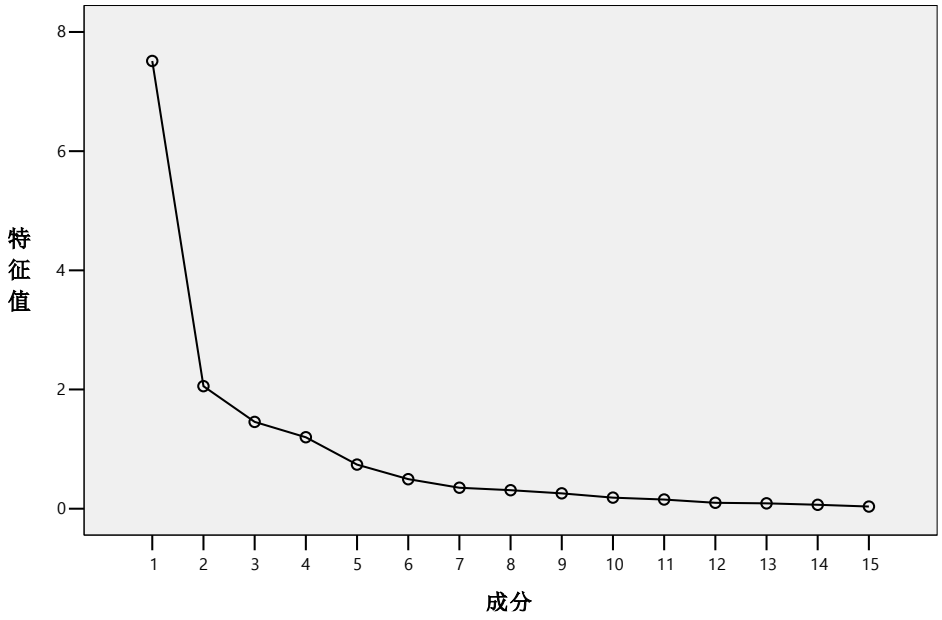


图10-2　四个成分特征值的碎石图

表10-4是SPSS输出的成分矩阵，成分矩阵并不是我们需要的主成分系数。把表10-4中的各列除以表10-3中对应特征值的平方根，就可以得到主成分的系数（表10-5）。

根据表10-5，可以写出各个主成分的表达式。例如第一主成分为：

第一主成分=0.162× 简历格式分+0.213× 外貌分+0.040× 学习能力分+0.225× 兴趣爱好分+0.290× 自信心分+0.315× 洞察力分+0.158× 诚信度分+0.324× 销售能力分+0.134× 工作经验分+0.315× 工作魄力分+0.318× 志向抱负分+0.331× 理解能力分+0.333× 潜能分+0.259× 求职渴望度分+0.236× 适应力分。

根据主成分的表达式，可以对每个应聘者根据各指标得分算出其主成分得分，按照主成分得分对应聘者进行分析并进行解释。在计算主成分得分时，首先应该分别对十五个指标的数据列加以标准化，将标准化的指标值代入主成分表达式，得到各应聘者的主成分得分。

表10-4　成分矩阵

	成分			
	1	2	3	4
简历格式	.445	.615	.381	-.103
外貌	.584	-.051	-.028	.287
学习能力	.110	.340	-.519	.696
兴趣爱好	.617	-.186	.562	.378
自信心	.796	-.357	-.291	-.189
洞察力	.863	-.188	-.181	-.078
诚信度	.433	-.581	.343	.456
销售能力	.889	-.042	-.224	-.217
工作经验	.367	.793	.100	.074
工作魄力	.864	.066	-.096	-.171
志向抱负	.872	-.098	-.252	-.218
理解能力	.909	-.033	-.141	.082
潜能	.914	.032	-.088	.206
求职渴望度	.711	-.118	.564	-.220
适应力	.647	.603	.108	-.022

提取方法 :主成分分析法。

表10-5　主成分的系数表

	成分			
	1	2	3	4
简历格式	0.162	0.429	0.315	-0.094
外貌	0.213	-0.035	-0.023	0.262
研究能力	0.040	0.237	-0.430	0.636
兴趣爱好	0.225	-0.130	0.466	0.345
自信心	0.290	-0.249	-0.241	-0.173
洞察力	0.315	-0.131	-0.150	-0.071
诚信度	0.158	-0.405	0.284	0.416
推销能力	0.324	-0.029	-0.186	-0.198
工作经验	0.134	0.553	0.083	0.068
工作魄力	0.315	0.046	-0.080	-0.156
志向抱负	0.318	-0.068	-0.209	-0.199
理解能力	0.331	-0.023	-0.117	0.075
潜能	0.333	0.022	-0.073	0.188
求职渴望度	0.259	-0.082	0.467	-0.201
适应力	0.236	0.421	0.089	-0.020

在第一主成分中销售能力、工作魄力、志向抱负、理解能力、潜力的系数比较大，可以把第一主成分看作是直观影响工作业绩的指标。在第二主成分中工作经验和适应力的系数比较大，可以把第二主成分看成是影响工作业绩的经验指标。类似地，第三主成分中兴趣爱好和求职渴望度的系数比较大，可以把第三主成分看作是影响工作业绩的主观指标。第四主成分主要反映了求职者的学习能力。这样就把原始变量中的十五个指标降为四个主成分。

第二节　因子分析

一、因子分析的基本思想和数学模型

因子分析（factor analysis）是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。

例如，在企业形象或品牌形象的研究中，消费者可以通过一个有24个指标构成的评价体系，评价百货商场的24个方面的优劣。但消费者主要关心的是三个方面，即商店的环境、商店的服务和商品的价格。因子分析方法可以通过24个变量，找出反映商店环境、商店服务水平和商品价格的三个潜在的因子，对商店进行综合评价。

这三个潜在因子和24个变量的关系可以表示为：$x_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i$，$i = 1,\Lambda\ ,24$。我们称$F_1$、$F_2$、$F_3$为不可观测的潜在因子或者公因子。24个变量共享这三个因子，但是每个变量又有自己的个性，不被包含的部分ε_i，称为特殊因子。

因子分析是主成分分析的推广和发展，但是因子分析和主成分分析之间也是有区别的。主成分分析是将许多的变量通过线性组合综合为较少的几个主成分，通过原始数据可以直接计算出各个主成分的数值；而因子分析中的因子是从原始变量中提取出来的，能够表现原始变量信息的不可直接观测出来的综合变量。从数学上来说，主成分是原始变量的线性组合，而原始变量可以表示成公因子的线性组合。

因子分析不仅可用来研究变量之间的关系，还可以用来研究样品之间的关系，对变量进行研究称为*R*型因子分析，对样品进行研究称为*Q*型因子分析。从数学模型来看*R*型和*Q*型因子分析都是一样的，不同在于研究变量时是从相关系数矩阵出发的，而研究样品时是从相似系数矩阵出发的。以下主要介绍*R*型因子分析的数学模型。

$X_1, X_2, ..., X_p$为经过标准化后的原始数据的*p*个变量，现在假设原始变量都受到*q*个公因子和一个特殊因子的影响。模型的一般形式如下：

$$X_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \Lambda\ \alpha_{iq}F_q + \varepsilon_i, \quad i = 1,2,..., p$$

$F_1, F_2, ..., F_q$是彼此独立的公因子，ε_i是各个变量的特殊因子，a_{ij}是各个变量的因子载荷。因子载荷a_{ij}是第*i*个变量与第*j*个公因子的相关系数，反映了第*j*个公因子对第*i*个变量的影响程度。

模型需要满足以下的条件：

（1）$q \le p$。

（2）$\mathrm{cov}(F, \varepsilon) = 0$，即*F*和$\varepsilon_i$是不相关的。

（3）$\mathrm{var}(X_i) = a_{i1}^2 + ... + a_{iq}^2 + \mathrm{var}(\varepsilon_i) = 1$，$i = 1,2,..., p$。

（4）ε_i之间不相关，且方差不同。

因子分析中公因子的个数是少于原始变量的个数，这样才能达到简化模型的目的，特殊因子ε_i可以理解为原始变量提取公因子后所剩的信息，所以特殊因子和公因子间是不相关的，所

有公因子的方差加上特殊因子的方差就解释了原始变量的所有方差，所以有 $\text{var}(X_i)=a_{i1}^2+...+a_{iq}^2+\text{var}(\varepsilon_i)=1$。

定义 $h_i^2=\sum_{j=1}^{m}a_{ij}^2$ 为变量X_i的共同度。一个变量的共同度越接近1，说明该变量被公因子解释的程度越高，因子分析的效果越好。

因子载荷矩阵中各列元素的平方和 $S_j=\sum_{i=1}^{p}a_{ij}^2$ 称为公因子F_j对原始变量X的方差贡献，可以衡量因子F_j的相对重要性。

二、因子分析的步骤和结果分析

因子载荷矩阵的估计方法有多种，估计结果并不唯一。最常用的方法之一是主成分法：求解变量X的前m个主成分，进行简单的数学变换就可以得到因子载荷矩阵。与主成分分析类似，可以根据因子的累计贡献率确定因子的个数。

因子分析中得出的各个因子如果有明确的含义，则因子分析的模型会更加易于解释和有实际意义。在因子分析中可以对因子载荷矩阵进行旋转，使每个变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小。通过旋转，因子可以有更加明确的含义。常用的一种方法是方差最大旋转。

以上我们主要解决了用公因子的线性组合来表示一组观测变量的有关问题。如果我们要使用这些因子做其他的研究，比如把得到的因子作为自变量来做回归分析，对样本进行分类或评价，就需要计算每个个体在每个因子上的得分。

要计算因子得分，需要估计以下表达式：$F_j=b_{j0}+b_{j1}X_1+\Lambda+b_{jp}X_p$。如果对变量都进行了标准化，则模型中没有常数项。因子得分有多种计算方法，常用的一种是回归法。

根据以上分析，完整的因子分析过程可以包含以下主要步骤：

（1）根据问题选取原始变量；

（2）求其相关阵R，探讨其相关性；

（3）从R求解初始公因子F及因子载荷矩阵A(主成分法)；

（4）因子旋转，分析因子的含义；

（5）计算因子得分函数和因子得分；

（6）根据因子得分值进行进一步分析（例如综合评价）。

在SPSS中因子分析和主成分分析是同一个模块，分析的结果也很相似，不过在因子分析中需要进行因子旋转。为了便于将因子分析与主成分分析相比较，本节仍采用上一节中的数据，对48名应聘者十五个指标的测试得分进行因子分析，分析中采用主成分法提取因子，采用方差最大旋转对因子载荷矩阵进行旋转，结果分析如下。

表10-6输出的“公因子方差”表显示了各个变量的共同度。从表中可以看出，除了“外貌”变量外其他变量的共同度都比较高。

表10-6　公因子方差

	初始	提取
简历格式	1.000	.732
外貌	1.000	.427
学习能力	1.000	.882
兴趣爱好	1.000	.874
自信心	1.000	.882

洞察力	1.000	.819
诚信度	1.000	.851
销售能力	1.000	.889
工作经验	1.000	.780
工作魄力	1.000	.789
志向抱负	1.000	.880
理解能力	1.000	.853
潜能	1.000	.886
求职渴望度	1.000	.885
适应力	1.000	.795

提取方法：主成分分析。

表10-7是特征值和各因子方差贡献率的输出结果。和主成分分析不同的是，这个表中多了一栏，反映的是旋转后的因子的贡献率和累积贡献率。从表中可以看出，提取4个因子时可以解释总方差的81.5%；旋转后因子的贡献率会发生一定变化。

表10-7　解释的总方差

成分	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.514	50.092	50.092	7.514	50.092	50.092	5.766	38.443	38.443
2	2.056	13.709	63.801	2.056	13.709	63.801	2.726	18.175	56.618
3	1.456	9.705	73.506	1.456	9.705	73.506	2.386	15.904	72.523
4	1.198	7.986	81.492	1.198	7.986	81.492	1.345	8.969	81.492
5	.739	4.928	86.420						
6	.495	3.297	89.717						
7	.351	2.342	92.059						
8	.310	2.066	94.125						
9	.257	1.713	95.838						
10	.185	1.233	97.071						
11	.153	1.018	98.088						
12	.098	.650	98.739						
13	.089	.592	99.331						
14	.065	.431	99.762						
15	.036	.238	100.000						

提取方法：主成分分析。

为了更好地对因子进行解释，往往需要对因子进行旋转，与主成分分析的结果相比，旋转后公因子解释原始数据的能力并没有提高，但因子载荷矩阵及因子得分系数矩阵都发生了变化，因子载荷矩阵中的元素更趋向于0或者±1了。旋转之后的因子往往能够有更加明确的含义。旋转成分矩阵给出了旋转后标准化的原始变量表示成各因子的线性组合的系数矩阵（表10-8），利用表中数据有：

标准化的简历格式分≈0.116×第一个因子+0.830×第二个因子+0.109×第三个因子-0.136×第四个因子

标准化的外貌分≈0.440×第一个因子+0.151×第二个因子+0.399×第三个因子+0.227×第四个因子

……

从表10-8可以看出，第一个因子在变量自信心、洞察力、销售能力、工作魄力、志向抱负、理解能力和潜力上有较大的系数，可以抽象为应聘者主客观工作能力因子；第二个因子在变量简历格式、工作经验、适应力上有较大的系数，可以抽象为应聘者对客观环境的适应力因子；第三个因子在兴趣爱好、诚信度和求职渴望度上有较大的系数，可以抽象为应聘者的兴趣和诚信因子；第四个因子在学习能力上有较大的系数，可以抽象为应聘者的学习能力因子。

表10-8　旋转成分矩阵

	成分			
	1	2	3	4
简历格式	.116	.830	.109	-.136
外貌	.440	.151	.399	.227
学习能力	.064	.128	.007	.928
兴趣爱好	.220	.245	.871	-.081
自信心	.916	-.107	.163	-.065
洞察力	.863	.097	.255	.002
诚信度	.219	-.242	.863	.001
销售能力	.910	.223	.103	-.041
工作经验	.087	.851	-.055	.211
工作魄力	.800	.349	.156	-.052
志向抱负	.918	.159	.100	-.041
理解能力	.811	.255	.331	.143
潜能	.747	.326	.413	.224
求职渴望度	.440	.363	.534	-.524
适应力	.383	.797	.076	.084

提取方法 :主成分分析法；旋转方式：方差最大化旋转。

成分得分系数矩阵给出了根据原始变量计算因子得分的系数矩阵（表10-9），可以用来计算因子得分，注意计算中需要先对原始变量进行标准化。利用表10-9中的数据可以得到计算因子得分的公式：

因子1≈-0.099×简历格式分+0.016×外貌分-0.020×学习能力分-0.159×兴趣爱好分+0.251×自信心分+0.185×洞察力分-0.093×诚信度分+0.217×销售能力分-0.082×工作经验分+0.155×工作魄力分+0.228×志向抱负分+0.129×理解能力分+0.080×潜能分-0.026×求职渴望度分-0.014×适应力分。

因子2≈0.373×简历格式分-0.008×外貌分+0.003×学习能力分+0.072×兴趣爱好分-0.173×自信心分-0.077×洞察力分-0.157×诚信度分-0.020×销售能力分+0.372×工作经验分+0.053×工作魄力分-0.050×志向抱负分-0.005×理解能力分+0.031×潜能分+0.126×求职渴望度分+0.311×适应力分。

因子3和因子4也可以用同样的方法得到。

表10-9　成分得分系数矩阵

	成分			
	1	2	3	4
简历格式	-.099	.373	.016	-.139
外貌	.016	-.008	.168	.189
学习能力	-.020	.003	.067	.698

兴趣爱好	-.159	.072	.483	-.005
自信心	.251	-.173	-.105	-.050
洞察力	.185	-.077	-.033	-.001
诚信度	-.093	-.157	.494	.081
销售能力	.217	-.020	-.144	-.054
工作经验	-.082	.372	-.050	.110
工作魄力	.155	.053	-.090	-.062
志向抱负	.228	-.050	-.147	-.051
理解能力	.129	-.005	.035	.106
潜能	.080	.031	.106	.172
求职渴望度	-.026	.126	.185	-.381
适应力	-.014	.311	-.044	.021

提取方法 :主成分分析法。

可以根据每个应聘者在不同因子上的得分来衡量应聘者在各项考核指标方面的综合能力。例如：在进行因子分析时把因子得分保存为变量，然后按第一因子得分排序，结果见表10-10。从表10-10可以看出，48个应聘者中编号为10的应聘者的主客观工作能力是最强的，编号为42的应聘者的主客观工作能力是最差的。也可以类似的对应聘者的其它特征进行分析。

表10-10　根据应聘者第一因子得分的排序结果

应聘者编号	*FAC1_1*	*FAC2_1*	*FAC3_1*	*FAC4_1*
10	2.014	-0.493	-1.628	1.444
11	1.963	-0.542	-2.681	1.519
12	1.577	-0.817	-0.951	1.550
37	1.150	-1.584	-0.757	0.002
…	…	…	…	…
35	-1.656	-0.417	0.191	-1.126
41	-1.744	2.045	-2.068	0.873
42	-2.040	2.227	-2.730	0.382

第三节　主成分与因子分析的软件实现

本节主要介绍SPSS中实现主成分分析与因子分析的具体步骤，以及过程中各个选项的相关统计意义和具体说明。当然对于初学者可以完全使用系统的默认值进行最简单的因子分析，不过了解了相关选项的统计意义，能帮助您得到更满意的结果。

在SPSS中主成分和因子分析是在一个模块中完成的。采用主成分方法提取因子、不进行因子旋转，就可以得到主成分分析需要的结果。

在用SPSS读取数据后，按照“分析”→“降维”→“因子分析”顺序打开对话框，如图10-3。把要进行分析的变量移到右边的“变量”框中。在主对话框下面有“描述”、“抽取”、“旋转”、“得分”和“选项”五个选项钮。下面分别介绍五个选项中常用到的设置。

图 10-3　因子分析过程的主对话框

1. “描述”选项钮

在主对话框中单击“描述”选项钮就得到如图10-4对话框，在“统计量”栏中，选中“单变量描述性”复选框时，分析结果会输出参与分析的每个变量的均值、标准差等单变量的描述统计量。“原始分析结果”复选项给出因子提取时，分析变量的公因子方差。在“相关矩阵”栏中，“系数”给出了原始分析变量间的相关系数矩阵，“显著性水平”给出每个相关系数相对于0的单尾假设检验的显著性水平。

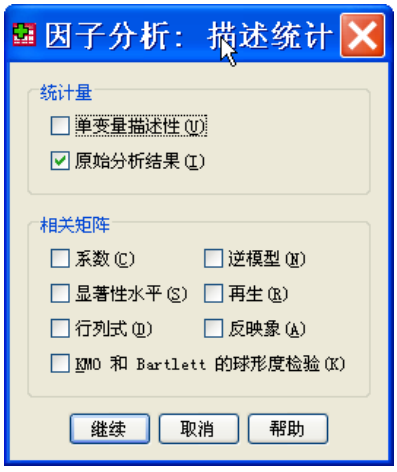


图10-4　“描述”子对话框

2. “抽取”选项钮

在主对话框中单击“抽取”选项钮就得到如图10-5对话框。

“方法”框是一组指定的提取方法的选项，可以从主成分法、未加权的最小平方法、最大似然法等多种方法中选择，默认的是主成分法。

“输出”中指定与因子提取相关的输出，包括未经旋转的因子提取结果和碎石图。

“抽取”栏中可以设定因子或主成分的个数。默认是按照特征值大于1提取；也可以设定固定的因子个数。

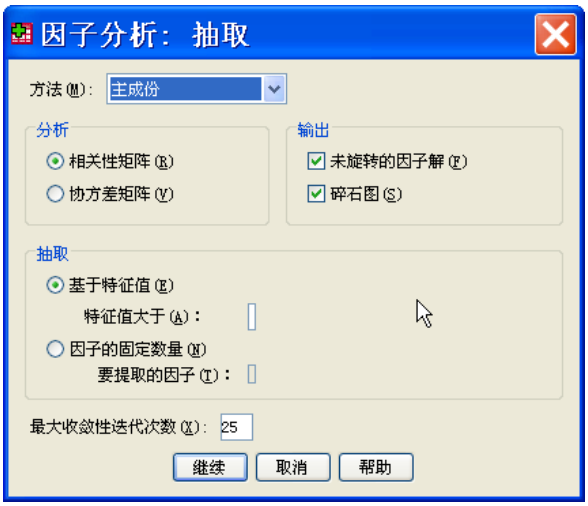


图10-5　“抽取”子对话框

3. “旋转”选项钮

在主对话框中单击“旋转”选项钮就得到如图10-6对话框：在“方法”中选择旋转的方法，其中的最大方差法，是一种正交旋转，是常使用的旋转方法。在“输出”中选择有关的输出结果。

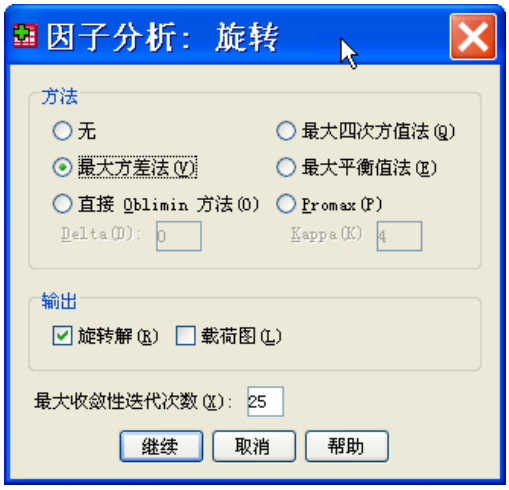


图10-6　“旋转”子对话框

4. “得分”选项钮

在主对话框中单击“得分”选项钮就得到如图10-7的对话框：选中复选框“保存为变量”将因子得分作为新变量保存在数据文件中。“方法”中最常用的是回归法。

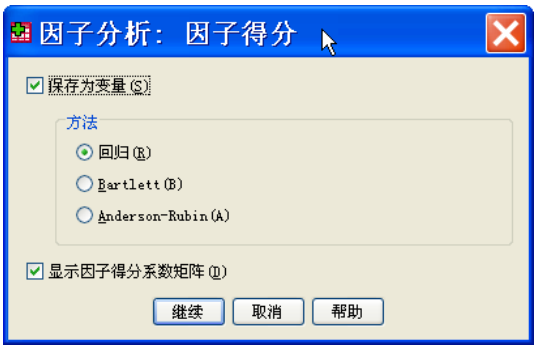


图10-7　“得分”子对话框

5. “选项”选项钮

在主对话框中单击“选项”按钮就得到如图10-8对话框：这里可以选择缺失值的处理方法。

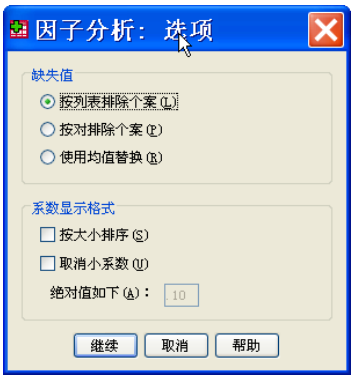


图10-8　“选项”子对话框

本节主要介绍SPSS中如何实现主成分分析与因子分析的具体步骤，以及过程中各个选项的相关统计意义和具体说明。限于篇幅，一些方法和操作没有做进一步的解释，请参考软件的帮助文件或其他相关资料。

小　结

本章主要介绍主成分分析和因子分析的思想，主成分分析和因子分析在SPSS软件中的实现，软件输出结果的解释和说明，从而达到降维或综合评价的目的，同时主成分分析法也是其它诸多多元统计分析方法的应用基础。主成分分析和因子分析的关系主要表现在以下几方面：主成分分析和因子分析都是用于将多个相关变量简化为少数几个综合指标的多元统计分析方法，可以在尽可能保留变量信息的基础上实现变量的降维，因此两种方法的分析结果对原始变量的依赖性高，只有原始变量之间具有较强相关性时，变量的降维才有效果。

对于一个确定的原始数据矩阵，主成分分析可以得到唯一确定的主成分系数矩阵，各主成分可以直接表示为对应的特征向量与相应原始变量的线性组合。因子分析是主成分分析的推广和发展，但是因子分析和主成分分析之间也是有区别的。主成分分析是将许多的变量通过线性组合综合为较少的几个主成分，通过原始数据可以直接计算出各个主成分的数值；而因子分析中的因子是从原始变量中提取出来的，能够表现原始变量信息的不可直接观测出来的综合变量。从数学上来说，主成分是原始变量的线性组合，而原始变量可以表示成其公因子和特殊因子的线性组合。

主成分分析得到的主成分唯一确定但不一定具有实际含义。因子分析中因子载荷不是唯一

的，通过因子旋转，可以对因子载荷进一步简化，得到具有明确实际意义的公因子，从而更容易解释。

思考与练习

1. 收集我国31个省（区、市）2007年反映经济发展情况的八项指标的数据，具体采用的指标包括：地区生产总值、工业总产值、固定资产投资、职工平均工资、居民消费水平、货物周转量、居民消费价格指数、商品零售价格指数。并对这八项指标利用主成分分析法进行降维。
2. 对第1题的数据进行因子分析，解释分析结果的经济意义，并比较主成分分析和因子分析的结果。
3. 对第1题的数据，采用因子分析法对我国各地区的经济发展状况进行综合评价，并进行排序。

第十一章 聚类分析与判别分析

聚类分析与判别分析是两类常用多元分析方法。聚类分析可以将个体或对象分类，使得同一类中的对象之间的相似性比与其他类的对象的相似性更强；而判别分析则可以根据已掌握的样本信息建立判别函数，当遇到新的样本点时根据判别函数可以判断该样本点所属的类别。

第一节 聚类分析

一、聚类分析的基本思想

“物以类聚，人以群分”。分类处理，在现实中极为普遍。

在生物、经济、社会、人口等领域的研究中，存在着大量量化分类研究。例如：在生物学中，为了研究生物的演变，生物学家需要根据各种生物不同的特征对生物进行分类；在经济研究中，为了研究不同地区城镇居民生活中的收入和消费情况，往往需要划分不同的类型去研究；在人口学研究中，需要构造人口生育分类模式、人口死亡分类状况，以此来研究人口的生育和死亡规律。

历史上，这些分类方法多半是人们主要依靠经验作定性分类，致使许多分类带有主观性和任意性，特别是对于多因素、多指标的分类问题，定性分类的准确性不好把握。为了克服定性分类存在的不足，人们把数学方法引入分类中，形成了数值分类学，进而产生了聚类分析这一最常用的技巧。

聚类分析将个体或对象分类，使得同一类中的对象之间的相似性比与其他类的对象的相似性更强。其目的在于：使类内对象的同质性最大化和类间对象的异质性最大化。

聚类分析通常可以分为两种：Q 型聚类和 R 型聚类。Q 型聚类是对观测个体的分类，R 型聚类是对变量的分类。二者在数学上是对称的，没有本质区别。

二、符号说明

多元统计分析中要注意区分样本和变量。
每个样品有 p 个指标（变量）从不同方面描述其性质，形成一个 p 维的向量，可以把 n 个样品看成 p 维空间中的 n 个点。

我们用记号 X_{jk} 表示第 k 个变量第 j 次观测值（或称第 j 个项目的测量值），即：

X_{jk}=第 k 个变量第 j 次观测值

p 个变量的 n 个观测值可表示如下：

	变量1	变量2	L	变量k	L	变量p
观测1	X_{11}	X_{12}	L	X_{1k}	L	X_{1p}
观测2	X_{21}	X_{22}	L	X_{2k}	L	X_{2p}
M	M	M		M		M
观测j	X_{j1}	X_{j2}	L	X_{jk}	L	X_{jp}
M	M	M		M		M
观测n	X_{n1}	X_{n2}	L	X_{nk}	L	X_{np}

记为：

$$\mathbf{X}=\begin{pmatrix} X_{11} & X_{12} & \mathrm{L} & X_{1k} & \mathrm{L} & X_{1p} \\ X_{21} & X_{22} & \mathrm{L} & X_{2k} & \mathrm{L} & X_{2p} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} & & \mathrm{M} \\ X_{j1} & X_{j2} & \mathrm{L} & X_{jk} & \mathrm{L} & X_{jp} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} & & \mathrm{M} \\ X_{n1} & X_{n2} & \mathrm{L} & X_{nk} & \mathrm{L} & X_{np} \end{pmatrix}$$

记 $X^j=(X_{j1},X_{j2},\mathrm{L}\ ,X_{jp})'\in R^p$，表示第 j 个样品，它表示 p 维空间的一个点。则有：

$$\mathbf{X}_{(n\times p)}=\begin{pmatrix} (X^1)' \\ (X^2)' \\ \mathrm{M} \\ (X^n)' \end{pmatrix}$$

记 $X_i=\begin{pmatrix} X_{1i} \\ X_{2i} \\ \mathrm{M} \\ X_{ni} \end{pmatrix}\in R^n$，表示第 i 个变量所有 n 个观测值，则有：

$$\mathbf{X}_{(n\times p)}=\left(X_1,\ \ X_2,\ \ \mathrm{L}\ ,\ \ X_p\right)$$

在不引起混淆的情况下，我们也以 $X_1,\ \ X_2,\ \ \mathrm{L}\ ,\ \ X_p$ 表示变量。

三、相似性度量

在聚类之前，要首先分析样品间的相似性。

一般说，研究的样品或指标（变量）之间是存在着程度不同的相似性（亲疏关系）。于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间的相似程度的统计量，以这些统计量为划分类型的依据，把一些相似程度较大的样品（或指标）聚合为一类，把另外一些彼此之间相似程度较大的样品（或指标）又聚合为另外一类，等等。

因而对相似性的描述成为聚类分析的基础。

相似性度量的工具一般可以采用距离和相似系数。距离常用来度量样品间相似性，相似系数常用来度量变量间相似性。

1．样品间相似性度量

两个样品间相似程度就可用 p 维空间中的两点距离公式来度量。两点距离公式可以从不同角度进行定义，令 d_{ij} 表示样品 X^i 与 X^j 的距离，常用以下距离公式：

（1）绝对距离

$$d_{ij}(1)=\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right| \tag{11-1}$$

（2）平方欧氏距离

$$d_{ij}(2)=(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^2)^{1/2} \tag{11-2}$$

（3）切比雪夫距离

$$d_{ij}(\infty)=\max_{1\le k\le p}\left|X_{ik}-X_{jk}\right| \tag{11-3}$$

（4）明考夫斯基距离（明氏距离）

$$d_{ij}(q)=(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^q)^{1/q} \tag{11-4}$$

绝对距离、平方欧氏距离与切比雪夫距离都是明氏距离的特例（$q=1,2,\infty$）。

明氏基距离主要有以下两个缺点：

①明氏距离的值与各指标的量纲有关，而各指标计量单位的选择有一定的人为性和随意性。

②明氏距离的定义没有考虑各个变量之间的相关性和重要性。实际上，明考夫斯基距离是把各个变量都同等看待，将两个样品在各个变量上的离差简单地进行了综合。

考虑到明氏距离的缺陷，可以采用兰氏距离和马氏距离。

（5）兰氏距离

兰思和威廉姆斯(Lance & Williams)所给定的一种距离，其计算公式为

$$d_{ij}(L)=\sum_{k=1}^{p}\frac{\left|X_{ik}-X_{jk}\right|}{X_{ik}+X_{jk}} \tag{11-5}$$

这是一个自身标准化的量，由于它对大的奇异值不敏感，使其特别适合于高度偏倚的数据，

有助于克服明氏距离的第一个缺点。但它也没有考虑指标之间的相关性。

（6）马氏距离

印度著名统计学家马哈拉诺比斯(P．C．Mahalanobis)所定义了一种距离，其计算公式为：

$$d_{ij}^2(M)=(X^i-X^j)'\boldsymbol{\Sigma}^{-1}(X^i-X^j) \tag{11-6}$$

其中，X^i与X^j为第 i 个和第 j 个样本，列向量，来自均值向量为 μ ，协方差为Σ（＞0）的总体。

马氏距离又称为广义欧氏距离。显然，马氏距离与上述各种距离的主要不同就是它考虑了观测变量之间的相关性。如果各变量之间相互独立，即观测变量的协方差矩阵是对角矩阵，则马氏距离就退化为用各个观测指标的方差的倒数作为权数的加权平方欧氏距离。马氏距离还考虑了观测变量之间的变异性，不再受各指标量纲的影响。将原始数据作线性变换后，马氏距离不变。

马氏距离计算的困难在于协方差矩阵的计算。通常总体的协方差矩阵未知，可以用样本数据估计。

一般说来，同一批数据采用不同的距离公式，会得到不同的分类结果。通常选择距离公式应注意遵循以下的基本原则：

（1）要考虑所选择的距离公式在实际应用中有明确的意义。如欧氏距离就有非常明确的空间距离概念，马氏距离有消除量纲影响的作用。

（2）要综合考虑对样本观测数据的预处理和将要采用的聚类分析方法。如在进行聚类分析之前已经对变量作了标准化处理，则通常就可采用欧氏距离。

（3）实际中，聚类分析前不妨试探性地选择几个距离公式分别进行聚类，然后对聚类分析的结果进行对比分析，以确定最合适的距离测度方法。

2．变量相似性的度量

变量间的相似性有两种度量方法：夹角余弦和相关系数。

（1）夹角余弦

两变量 X_i 与 X_j 看作 p 维空间的两个向量，这两个向量间的夹角余弦可用下式进行计算

$$\cos\theta_{ij}=\frac{\sum_{k=1}^{p}X_{ik}X_{jk}}{\sqrt{(\sum_{k=1}^{p}X_{ik}^2)(\sum_{k=1}^{p}X_{jk}^2)}} \tag{11-7}$$

显然，｜cos θ_{ij}｜ ≤1。

（2）相关系数

相关系数经常用来度量变量间的相似性。变量 Xi 与 Xj 的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)(X_{jk}-\bar{X}_j)}{\sqrt{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)^2\sum_{k=1}^{p}(X_{jk}-\bar{X}_j)^2}} \tag{11-8}$$

显然也有，｜r_{ij}｜ ≤1。

无论是夹角余弦还是相关系数，它们的绝对值都小于 1，作为变量近似的度量工具，我们把它们统记为 c_{ij}。当｜c_{ij}｜=1 时，说明变量 X_i 与 X_j 完全相似；当｜c_{ij}｜近似于 1 时，说明变量 X_i 与 X_j 非常密切；当｜c_{ij}｜= 0 时，说明变量 X_i 与 X_j 完全不一样；当｜c_{ij}｜近似于 0 时，

说明变量 X_i 与 X_j 差别很大。

据此，我们把比较相似的变量聚为一类，把不太相似的变量归到不同的类内。

在实际聚类过程中，为了计算方便，我们把变量间相似性的度量公式作如下变换：

$$d_{ij} = 1 - |c_{ij}| \tag{11-9}$$

或者

$$d_{ij}^2 = 1 - c_{ij}^2 \tag{11-10}$$

四、系统聚类法

1．系统聚类的基本思路

系统聚类思路是：假设总共有 n 个样品（或变量），第一步将每个样品（或变量）独自聚成一类，共有 n 类；第二步根据所确定的样品（或变量）“距离”公式，把距离较近的两个样品（或变量）聚合为一类，其它的样品（或变量）仍各自聚为一类，共聚成 $n-1$ 类；第三步将“距离”最近的两个类进一步聚成一类，共聚成 $n-2$ 类；……，以上步骤一直进行下去，最后将所有的样品（或变量）全聚成一类。

2．类间距离与系统聚类方法

在进行系统聚类之前，我们首先要定义类与类之间的距离，不同的类间距离定义产生了不同的系统聚类法。常用的类间距离定义有 8 种，与之相应的系统聚类法也有 8 种，分别为最短距离法、最长距离法、中间距离法、重心法、类平均法、可变类平均法、可变法和离差平方和法。它们的归类步骤基本上是一致的，主要差异是类间距离的计算方法不同。以下用 d_{ij} 表示样品 X^i 与 X^j 之间距离，用 D_{ij} 表示类 G_i 与 G_j 之间的距离。

最短距离法定义类与类之间的距离为两类最近样品的距离，即为

$$D_{\mathrm{pq}} = \min_{X^i \in G_{\mathrm{p}}, X^j \in G_{\mathrm{q}}} d_{ij} \tag{11-11}$$

设类 G_i 与 G_j 合并成一个新类记为 G_r，则 G_k 与 G_r 的距离为

$$D_{kr} = \min_{X^i \in G_k, X^j \in G_r} d_{ij} = \min\{\min_{X^i \in G_k, X^j \in G_p} d_{ij}, \min_{X^i \in G_k, X^j \in G_q} d_{ij}\} = \min\{D_{kp}, D_{kq}\} \tag{11-12}$$

类似的，最长距离法定义类与类之间的距离为两类最远样品的距离；重心法定义类与类之间的距离为两个类的重心之间的距离，等等，我们不再详述。

离差平方和法也称为 Ward 法。按照这种方法，在进行聚类时先计算某两个类各自的类内离差平方和，然后计算把这两个类合并后的类内离差平方和，计算出两个类合并前后类内离差平方和的增加量。最后，将类内离差平方和增加最小的两个类进行合并，依此类推。

下面我们用最短距离法来说明系统聚类的步骤。

（1）定义样品之间距离，计算样品的两两距离，得一距离阵记为 D（0），开始每个样品自成一类，显然这时 $D_{ij}=d_{ij}$。

（2）找出距离最小元素，设为 D_{pq}，则将 G_p 和 G_q 合并成一个新类，记为 G_r，即 $G_r = \{G_p, G_q\}$。

（3）按（11-12）计算新类与其它类的距离。

（4）重复（2）、（3）两步，直到所有元素。并成一类为止。如果某一步距离最小的元素不止一个，则对应这些最小元素的类可以同时合并。

【例 11.1】设有五个样品，每个只测量一个指标，分别是 1，2，3，7，9。试用最短距离法将它们分类。

（1）样品采用绝对值距离，计算样品间的距离阵 D（0），见表 11-4。

表 11-4 距离阵 D（0）

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	2	1	0		
G_4	6	5	4	0	
G_5	8	7	6	2	0

（2）D（0）中最小的元素是 $D_{12}=D_{23}=1$，于是将 G_1、G_2 与 G_3 合并成 G_6，并利用（11-14）式计算新类与其它类的距离阵 D（1），见表 11-5。

表 11-5 距离阵 D（1）

	G_6	G_4	G_5
G_6	0		
G_4	4	0	
G_5	6	2	0

（3）在 D（1）中最小值是 $D_{45}=2$，由于 G_4 与 G_5 合并成一个新类 G_7，G_7 与其它类的距离阵 D（2），见表 11-6。

表 11-6 距离阵 D（2）

	G_6	G_7
G_6	0	
G_7	4	0

（4）最后将 $G6$ 和 $G7$ 合并成 $G8$，这时所有的五个样品聚为一类，聚类过程终止。

五、K-均值聚类分析

系统聚类法需要计算出不同样品或变量的距离，还要在聚类的每一步都要计算“类间距离”，相应的计算量自然比较大；特别是当样本的容量很大时，需要占据非常大的计算机内存空间，这给应用带来一定的困难。而 K-均值法是一种快速聚类法，采用该方法得到的结果比较简单易懂，对计算机的性能要求不高，因此应用也比较广泛。

K-均值法是麦奎因（MacQueen，1967）提出的，这种算法的基本思想是将每一个样品分配给最近中心（均值）的类中，具体的算法至少包括以下三个步骤：

1．将所有的样品分成 K 个初始类；

2．通过欧氏距离将某个样品划入离中心最近的类中，并对获得样品与失去样品的类，重新计算中心坐标；

3．重复步骤 2，直到所有的样品都不能再分配时为止。

K-均值法和系统聚类法一样，都是以距离的远近亲疏为标准进行聚类的，但是两者的不同之处也是明显的：系统聚类对不同的类数产生一系列的聚类结果，而 K-均值法只能产生指定类数的聚类结果。具体类数的确定，离不开实践经验的积累；有时也可以借助系统聚类法以一部分样品为对象进行聚类，其结果作为 K-均值法确定类数的参考。

下面通过一个具体问题说明 K-均值法的计算过程。

【例 11.2】假定我们对 A、B、C、D 四个样品分别测量两个变量（表 11-7），试将以上的样品聚成两类。

表 11-7　两个变量的观测结果

样品	变量	
	X_1	X_2
A	6	4
B	-2	2
C	3	3
D	-1	-1

第一步：按要求取 K=2，为了实施均值法聚类，我们将这些样品随意分成两类，比如（A、B）和（C、D），然后计算这两个聚类的中心坐标，见表 11-8 所示。

表 11-8　两个聚类的中心坐标

聚类	中心坐标	
	$\bar{X}_1$	$\bar{X}_2$
（A、B）	2	3
（C、D）	1	1

表 11-8 中的中心坐标是通过原始数据计算得来的，比如（A、B）类的，

$$\bar{X}_1 = \frac{5+(-1)}{2} = 2$$

等等。

第二步：计算某个样品到各类中心的欧氏平方距离，然后将该样品分配给最近的一类。对于样品有变动的类，重新计算它们的中心坐标，为下一步聚类做准备。先计算 A 到两个类的平方距离：

$$d^2(A,(AB)) = (6-2)^2 + (4-3)^2 = 17$$

$$d^2(A,(CD)) = (6-1)^2 + (4-1)^2 = 34$$

由于 A 到（A、B）的距离小于到（C、D）的距离，因此 A 不用重新分配。计算 B 到两类的平方距离：

$$d^2(B,(AB)) = (-2-2)^2 + (2-3)^2 = 17$$

$$d^2(B,(CD)) = (-2-1)^2 + (2-1)^2 = 10$$

由于 B 到（A、B）的距离大于到（C、D）的距离，因此 B 要分配给（C、D）类，得到新的聚类是（A）和（B、C、D）。更新中心坐标如表 11-9 所示。

表 11-9　　更新后的中心坐标

聚类	中心坐标	
	$\bar{X}_1$	$\bar{X}_2$
（A）	6	4
（B、C、D）	0	2

第三步：再次检查每个样品，以决定是否需要重新分类。计算各样品到各中心的距离平方，得结果见表 11-10。

表 11-10　样品聚类结果

聚类	样品到中心的距离平方

	A	B	C	D
（A）	0	68	10	74
（B、C、D）	40	4	10	10

到现在为止，每个样品都已经分配给距离中心最近的类，因此聚类过程到此结束。最终得到K=2 的聚类结果是 A 独自成一类，B、C、D 聚成一类。

第二节 判别分析

判别分析（discriminant analysis），是在已知的分类之下，判断新的样本隶属于那一类的多元分析方法。

一、判别分析的基本思想

判别分析的基本思想是：根据已掌握的每个类别的若干样本的数据信息，建立判别公式和判别准则。当遇到新的样本点时，根据总结出来的判别公式和判别准则，即能判别该样本点所属的类别。

已知分类可用总体表示，所以，判别分析是判断样本属于那个总体的方法。

判别分析用途甚多，如医学疾病诊断、动植物分类、商品等级划分和商业银行客户评级等。常用的判别分析方法有：距离判别法、Fisher 判别法、Bayes 判别法和逐步判别法。我们这里只介绍距离判法和 Fisher 判别法。

二、距离判别法

1．两总体的距离判别法（协方差阵相同）

先考虑两个总体的情况，设有两个协差阵 Σ 相同的 p 维正态总体 G_1 和 G_2，对给定的样本 Y，判别一个样品 Y 是来自哪一个总体，一个最直观的想法是计算 Y 到两个总体的距离。一般用马氏距离来指定判别规则，即有

$$\begin{cases} Y \in G_1, & \text{如} d^2(Y, G_1) < d^2(Y, G_2), \\ Y \in G_2, & \text{如} d^2(Y, G_2) < d^2(Y, G_1) \\ \text{待判}, & \text{如} d^2(Y, G_1) = d^2(Y, G_2) \end{cases} \tag{11-13}$$

样本 Y 到两类的距离之差为

$$\begin{aligned} & d^2(Y,G_2) - d^2(Y,G_1) \\ &= (Y-\mu_2)'\Sigma^{-1}(Y-\mu_2) - (Y-\mu_1)'\Sigma^{-1}(Y-\mu_1) \\ &= Y'\Sigma^{-1}\mathbf{y} - 2Y'\Sigma^{-1}\mu_2 + \mu_2'\Sigma^{-1}\mu_2 \\ &-(Y'\Sigma^{-1}\mathbf{y} - 2Y'\Sigma^{-1}\mu_1 + \mu_1'\Sigma^{-1}\mu_1) \\ &= 2Y'\Sigma^{-1}(\mu_1-\mu_2) - (\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \\ &= 2[Y - \frac{(\mu_1+\mu_2)}{2}]'\Sigma^{-1}(\mu_1-\mu_2) \end{aligned} \tag{11-14}$$

令

$$\bar{\mu}=\frac{\mu_1+\mu_2}{2},\quad \alpha=\Sigma^{-1}(\mu_1-\mu_2)=(a_1,a_2,\mathrm{L}\ ,a_p)'$$

$$W(Y)=(Y-\bar{\mu})'\alpha=a_1(Y_1-\bar{\mu}_1)+\mathrm{L}\ +a_p(Y_p-\bar{\mu}_p)=\boldsymbol{\alpha}'Y-\boldsymbol{\alpha}'\bar{\boldsymbol{\mu}} \tag{11-15}$$

则判别函数可以表示为：

$$\begin{cases} Y\in G_1, & \text{如}W(Y)>0 \\ Y\in G_2, & \text{如}W(Y)<0 \\ \text{待判}, & \text{如}W(Y)=0 \end{cases} \tag{11-16}$$

显然，$W(Y)$是Y的线性函数，线性判别函数使用起来最为方便，应用也最为广泛。

在实际应用中，协方差矩阵一般是未知的，需要预先估计。

具体判别步骤如下：

（1）分别计算各组的离差矩阵 S_1 和 S_2

（2）计算 $\hat{\Sigma}=\dfrac{S_1+S_2}{n_1+n_2-2}$

（3）计算类的均值 μ_1,μ_2

（4）计算 $\hat{\Sigma}^{-1},\mu_1-\mu_2,\dfrac{\mu_1+\mu_2}{2}$

（5）计算

判别函数的系数：$\Sigma^{-1}(\mu_1-\mu_2)$，

判别函数的常数：$(\dfrac{\mu_1+\mu_2}{2})'\ \Sigma^{-1}(\mu_1-\mu_2)$

（6）生成判别函数，将检验样本代入，得分，判类。

2．两总体的距离判别法（协方差阵不同）

判别准则是：

$$\begin{cases} Y\in G_1, & \text{如}d^2\left(Y,\ G_1\right)<d^2\left(Y,\ G_2\right), \\ Y\in G_2, & \text{如}d^2\left(Y,\ G_2\right)<d^2\left(Y,\ G_1\right) \\ \text{待判}, & \text{如}d^2(Y,G_1)=d^2(Y,G_2) \end{cases} \tag{11-17}$$

样本 Y到两类的距离之差为

$$\begin{aligned} &d^2(Y,G_2)-d^2(Y,G_1) \\ &=(Y-\mu_2)'\Sigma_2^{-1}(Y-\mu_2)-(Y-\mu_1)'\Sigma_1^{-1}(Y-\mu_1) \end{aligned} \tag{11-18}$$

3．多总体的距离判别法

设有 k 个总体 $G_1,G_2,\Lambda\ ,G_k$，其均值和协方差矩阵分别是 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Lambda\ ,\boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\Lambda\ ,\boldsymbol{\Sigma}_k$，而且 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\Lambda\ =\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$。对于一个新的样品 $\mathbf{y}$，要判断它来自哪个总体。

该问题与两个总体的距离判别问题的解决思想一样。计算新样品 $\mathbf{X}$ 到每一个总体的距

离，即

$$\begin{aligned} d^2(Y,G_\alpha) &= (\mathbf{y}-\boldsymbol{\mu}_\alpha)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}_\alpha) \\ &= Y'\boldsymbol{\Sigma}^{-1}Y - 2\boldsymbol{\mu}_\alpha'\boldsymbol{\Sigma}^{-1}Y + \boldsymbol{\mu}_\alpha'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\alpha \\ &= Y'\boldsymbol{\Sigma}^{-1}Y - 2(\mathbf{I}_\alpha'Y + C_\alpha) \end{aligned} \tag{11-19}$$

这里$\mathbf{I}_\alpha = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\alpha$，$C_\alpha = -\frac{1}{2}\boldsymbol{\mu}_\alpha'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\alpha$，$\alpha = 1,2,\Lambda\ ,k$

取线性判别函数为

$$W_\alpha(\mathbf{X}) = \mathbf{I}_\alpha'Y + C_\alpha,\quad \alpha = 1,2,\Lambda\ ,k \tag{11-20}$$

相应的判别规则为

$$Y \in G_i \quad \text{如果}\ W_i(Y) = \max_{1\le\alpha\le k}(\mathbf{I}_\alpha'Y + C_\alpha) \tag{11-21}$$

针对实际问题，当$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Lambda\ ,\boldsymbol{\mu}_k$和$\boldsymbol{\Sigma}$均未知时，可以通过相应的无偏估计量替代。

二、Fisher 判别法

从距离判别法，我们已经看到判别规则是一个线性函数，由于线性判别函数使用简便，因此我们希望能在更一般的情况下，建立一种线性判别函数。Fisher 判别法是根据方差分析的思想建立起来的一种能较好区分各个总体的线性判别法，由 Fisher 在 1936 年提出。该判别方法对总体的分布不做任何要求。

Fisher 判别法是一种投影方法，把高维空间的点向低维空间投影。在原来的坐标系下，可能很难把样品分开，而投影后可能区别明显。一般说，可以先投影到一维空间（直线）上，如果效果不理想，在投影到另一条直线上（从而构成二维空间），依此类推。每个投影可以建立一个判别函数。

下面给出 Fisher 判别法的详细步骤。

1．两个总体的 Fisher 判别函数

从两个总体中抽取具有 p 个指标的样品观测数据，借助于方差分析的思想构造一个线性判别函数：

$$C(Y) = C_1Y_1 + C_2Y_2 + \mathrm{L}\ + C_pY_p = C'Y \tag{11-22}$$

其中系数$C_1,C_2,\mathrm{L}\ ,C_p$确定的原则是使两组间的组间离差最大，而每个组的组内离差最小。

当建立了判别式以后，对一个新的样品值，我们可以将他的 p 个指标值代入判别式中求出 Y 值，然后与判别临界值比较，就可以将该样品归类。

设有 2 个总体G_1,G_2，其均值和协方差矩阵分别是μ_1,μ_2和$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$。

可以证明，Fisher 判别函数系数

$$C = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) \tag{11-23}$$

若总体均值与方差未知，可通过样本进行估计。

设从第一个总体G_1取得n_1个样本，从第二个总体G_2取得n_2个样本，记两组样本均值分别为$\overline{X}^{(1)}$、$\overline{X}^{(2)}$，样本离差阵为$S^{(1)}$、$S^{(2)}$。

显然，μ_1,μ_2的无偏估计为$\overline{X}^{(1)}$、$\overline{X}^{(2)}$。$(\Sigma_1+\Sigma_2)^{-1}$的估计有两种方式。

第一种估计方式是分别估计

$$\hat{\Sigma}_1=\frac{1}{n_1-1}S^{(1)},\quad \hat{\Sigma}_2=\frac{1}{n_2-1}S^{(2)} \tag{11-24}$$

判别函数为

$$\begin{aligned} C(Y)&=Y^{'}(\hat{\Sigma}_1+\hat{\Sigma}_2)^{-1}(\hat{\mu}_1-\hat{\mu}_2)\\ &=Y^{'}(\frac{1}{n_1}S^{(1)}+\frac{1}{n_2}S^{(2)})^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)}) \end{aligned} \tag{11-25}$$

第二种估计方式是联合估计

$$\Sigma_1\hat{+}\Sigma_2=\frac{1}{n_1+n_2-2}(S^{(1)}+S^{(2)}) \tag{11-26}$$

于是判别函数

$$C(Y)=Y^{'}(n_1+n_2-2)(S^{(1)}+S^{(2)})^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)}) \tag{11-27}$$

当$n_1=n_2$时，两种方法是等价的；当n_1与n_2相差不大时，两种方法近似；当n_1与n_2相差很大时两种方法相差较远。在等协方差阵的情况下，显然第二种方法更合理一些。目前采用较多的是第二种方法。

2．多个总体的 Fisher 判别函数

Fisher 判别法致力于寻找一个最能反映组和组之间差异的投影方向，即寻找线性判别函数。

设有k个总体G_1,G_2,Λ,G_k，其均值和协方差矩阵分别是$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Lambda,\boldsymbol{\mu}_k$和$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\Lambda,\boldsymbol{\Sigma}_k$。

在$X\in G_i$的条件下，有

$$\begin{aligned} &E(C'Y)=E(C'Y\mid G_i)=C'E(Y\mid G_i)=C'\mu_i, \qquad i=1,2,\mathrm{L}\ ,k\\ &D(C'Y)=D(C'Y\mid G_i)=C'D(Y\mid G_i)C=C'\boldsymbol{\Sigma}_iC, \quad i=1,2,\mathrm{L}\ ,k \end{aligned} \tag{11-28}$$

令

$$B=\sum_{i=1}^{k}(C'\mu_i-C'\bar{\mu})^2=C'\sum_{i=1}^{k}(\mu_i-\bar{\mu})^2C=C'B_0C$$

$$E=\sum_{i=1}^{k}C'\boldsymbol{\Sigma}_iC=C'(\sum_{i=1}^{k}\boldsymbol{\Sigma}_i)C=C'E_0C\ 。$$

B 相当于组间差，E 相当于组内差。运用判别分析的思想，构造

$$\Delta(C)=\frac{C'B_0C}{C'E_0C} \tag{11-29}$$

若求得$\Delta(C)$极大值，即可得到判别函数。显然，B_0,E_0均为非负定矩阵。$\Delta(C)$的极大值为方程

$$\left|B_0-\lambda E_0\right|=0 \tag{11-30}$$

的最大特征根，而系数向量 C 为最大特征根对应的特征向量。
若总体均值与方差未知，可通过样本进行估计。具体估计方法较为复杂，有兴趣读者可以参考有关书籍。
3. 判别规则

如果我们得到判别函数 $C(Y)=C'Y$，对于一个新的样本 Y，可以构造一个判别规则：

$$Y\in G_i,\ \text{当}\left|C'Y-C'\mu_i\right|=\min_{1\le j\le k}\left|C'Y-C'\mu_j\right|$$

第三节　聚类分析与判别分析的软件实现

一、聚类分析的 SPSS 实现

1．系统聚类法

【例 11.3】我们用 SPSS 软件自带的数据文件 World95.sav 来做一个实例分析。其目的是根据亚洲国家的经济发展水平和文化教育水平，对亚洲国家进行分类研究，这里我们进行聚类分析（在 World95.sav 数据中筛选出亚洲国家，使用“数据”→“选择个案”→“选择”中选入“地区=3”）。
详细步骤如下：
（1）打开数据。使用菜单中“文件”→“打开”命令，然后选中要分析的数据 World95.sav。在这个数据文件中，我们选择的变量有国家或地区、城市人口比例、平均女性寿命、平均男性寿命、非文盲人口比例、人均国内生产总值，国家（地区）来标识本例中的 17 个亚洲国家或地区，并以其他 5 个变量进行 Q 型聚类分析，即对国家进行聚类。
（2）在菜单中的选项中选择“分析”→“分类”，“分类”命令下有三个聚类分析命令，一是“两步聚类”，二是“K-均值聚类”，三是“系统聚类”。这里我们选择系统聚类法。
（3）在系统聚类法中，我们看到“分群”下有两个选项，“个案”指样品聚类或 Q 型聚类；“变量”指变量聚类或 R 型聚类。这里我们选择对样品进行聚类。

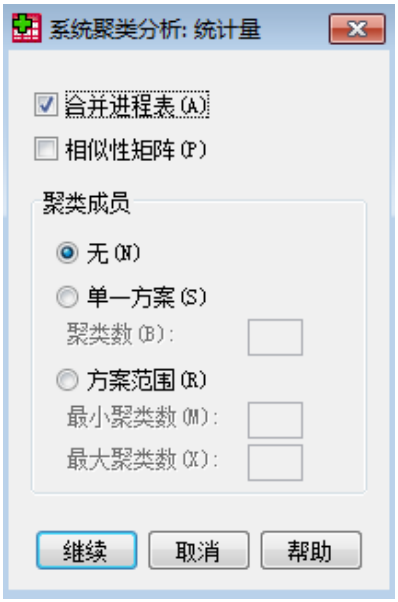
图 11-1　系统聚类主对话框

（4）“输出”下面有两个选项，分别是“统计量”、“输出图形”，我们可以选择所需要输出

的统计量和图形。

（5）在系统聚类法中底下有四个按钮，分别是“统计量”、“绘制”、“方法”和“保存”。

（a）在“统计量”中，有“合并进程表”、“相似性矩阵”。由“聚类成员”可以指定聚类的个数，“无”选项不指定聚类个数，“单一方案”指定一个确定类的个数，“方案范围”指定类的个数的范围（如从分 3 类到分 5 类）。



系统聚类子对话框：统计量

(b)在“绘制”（图）中，有“树状图”（也称谱系聚类图）、“冰柱图”、方向有“水平”和“垂直”两种。

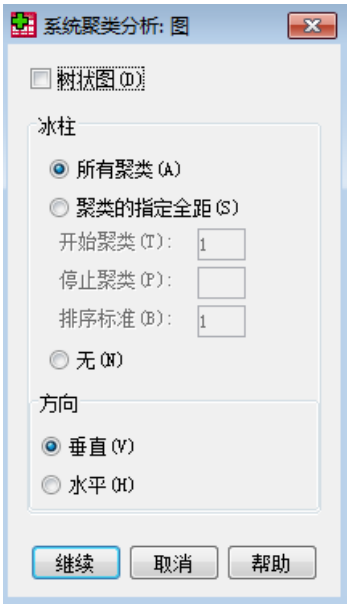


图 11-2 系统聚类子对话框：图

(c)在“方法”子对话框中，“聚类方法”可以选择组间连接、组内连接等，“度量标准”可以依不同变量类型选择。

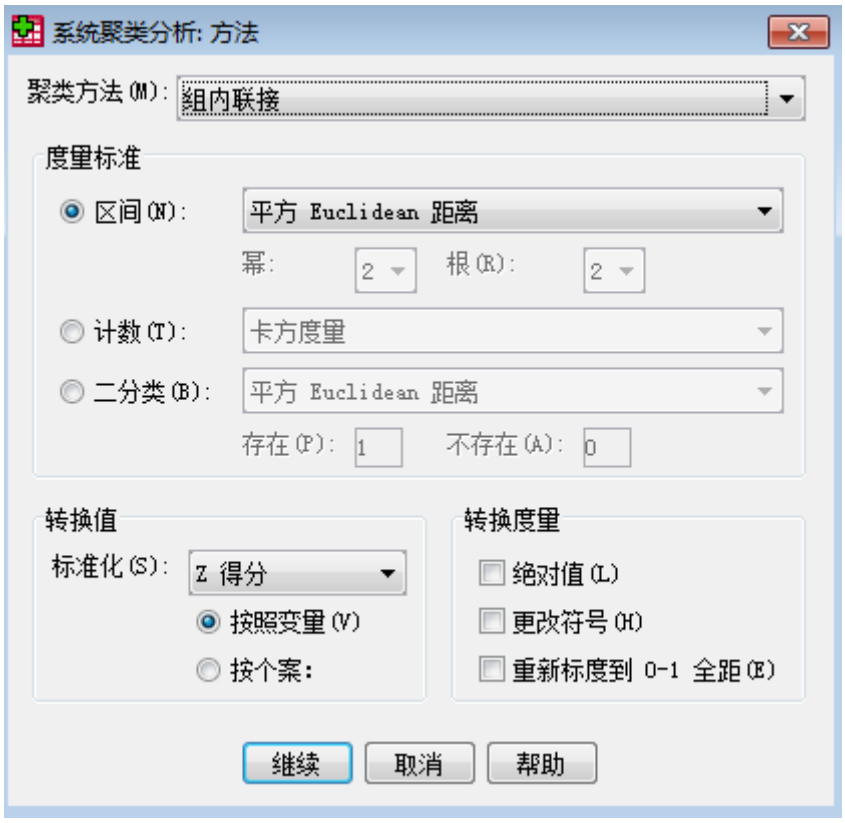


图 11-3 系统聚类子对话框：方法

选好每个选项后，点“确定”就可以执行了。这里我们将原始变量标准化（在“方法”选项下“转换值”的“标准化”空白框内，选择 Z 得分），在“统计量”选项中选择“合并进程表”，聚类方法选择组内联结法，计算距离选择平方欧氏距离，输出冰柱图和树状聚类图。图 11-4 是 SPSS18.0 绘制的树形图。这是分析聚类结果的最直观的图形之一。在图 11-4 中，从距离等于 5 的地方画一条垂直于横轴的直线，这条直线与树形图有 4 个交点，说明在这一距离上可以将样本点分为 4 类。我们可以从图中看出，这时朝鲜、韩国和台湾在一个组，香港、新加坡、日本在一个组，等等。

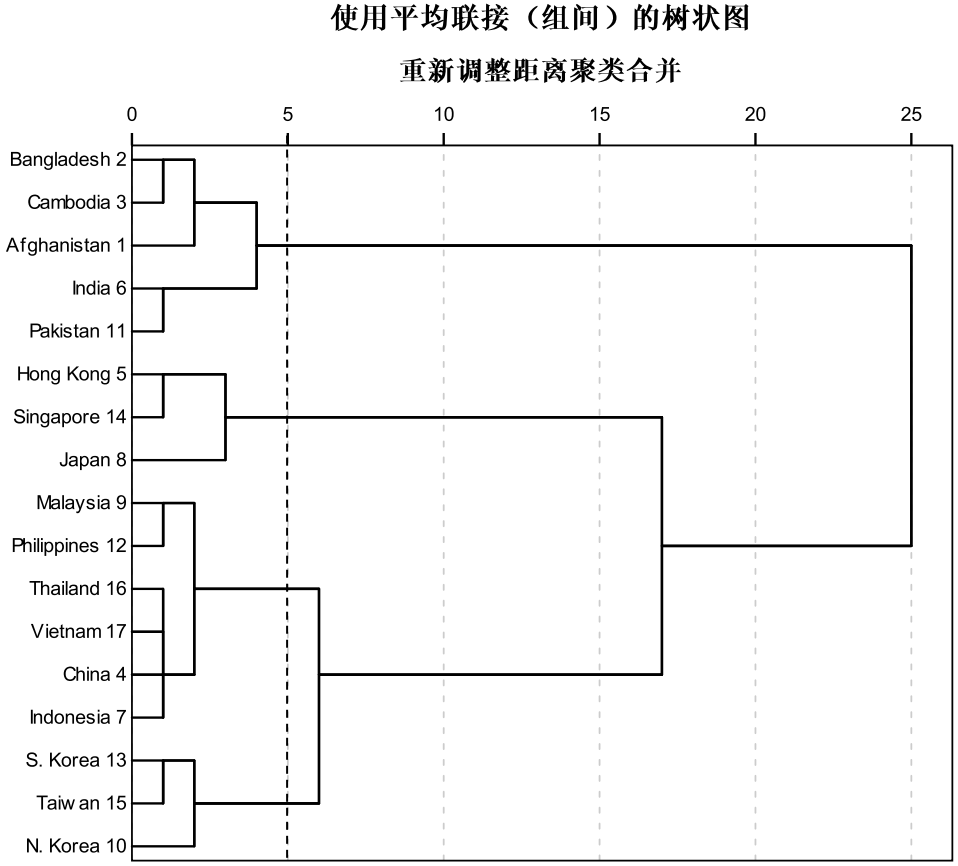


图 11-4 聚类结果的树形图

2.K-均值聚类法

【例 11.4】同样我们使用上面的数据文件 World95.sav，从中筛选出亚洲国家，试图将亚洲国家按经济和文教水平分为 3 类。可以使用快速聚类法对样品进行聚类。

我们使用的变量有国家或地区、城市人口比例、平均女性寿命、平均男性寿命、非文盲人口比例、人均国内生产总值，以国家（地区）来标识本例中的 17 个亚洲国家或地区，并以其他 5 个变量进行 Q 型聚类分析，即对国家（地区）进行聚类。

在 SPSS 软件中选择“分析”→“分类”→“K-均值聚类分析”。进入 K-均值聚类对话框以后，将上面 5 个变量选入“变量”，“个案标记依据”选择国家。将“聚类数”定为 3。我们可以在“选项”中选择“初始聚类中心”、“ANOVA 表”、“每个个案的聚类信息”（相关软件操作参见图 11-5）最后单击“确定”，可以得到相应的聚类结果（表 11-11）。根据表 11-11，香港、日本、新加坡在一个组，韩国和台湾在一个组，其余 12 个国家是一个组。

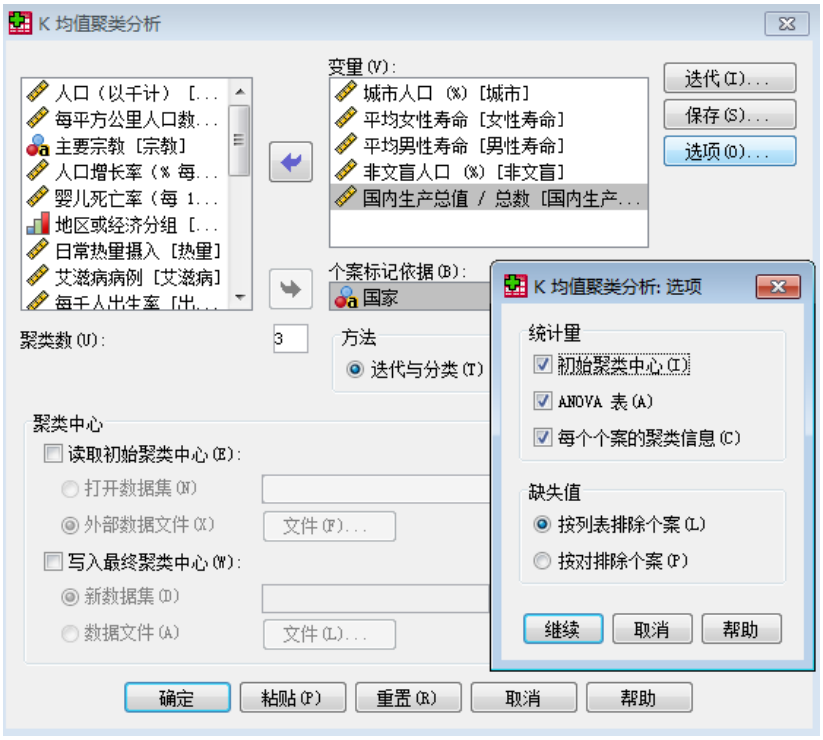


图 11-5 K-均值聚类分析对话框

表 11-11 各个国家的聚类结果

案例号	国家	聚类	距离
1	Afghanistan	1	571.615
8	Bangladesh	1	573.924
19	Cambodia	1	516.229
24	China	1	398.151
47	Hong Kong	2	1856.036
50	India	1	500.047
51	Indonesia	1	94.543
57	Japan	2	3363.045
66	Malaysia	1	2220.274
69	N. Korea	1	230.069
76	Pakistan	1	370.165
80	Philippines	1	96.542
86	S. Korea	3	214.034
89	Singapore	2	1507.033
96	Taiwan	3	214.034
98	Thailand	1	1025.608
108	Vietnam	1	545.396

二、判别分析的 SPSS 实现

我们以 Fisher 判别法为例，简要介绍判别分析的 SPSS 实现。

【例 11.5】一个城市的居民家庭，按其有无割草机可以分为两组，有割草机的一组记为 y=1，没有割草机的一组记为 y=0。家庭有无割草机与两个变量有关：X_1：家庭收入；X_2：房前屋后草地面积。割草机工厂欲判断一些家庭是否购买割草机。试用判别分析进行初步研究。数据如表 11-12。

表 11-12 购买割草机数据

有割草机家庭		*无割草机家庭*		*待判家庭*	
X_1	X_2	X_1	X_2	X_1	X_2
20.00	9.20	27.00	10.00	20.00	6.33
28.50	8.40	25.00	9.80	28.00	8.78
21.60	10.80	17.60	10.40		
20.50	10.40	21.60	8.60		
29.00	11.80	14.40	10.20		
36.70	9.60	28.00	8.80		
36.00	8.80	16.40	8.80		
27.60	11.20	19.80	8.00		
23.00	10.00	22.00	9.20		
31.00	10.40	15.80	8.20		
17.00	11.00	11.00	9.40		
27.00	10.00	17.00	7.00		

操作步骤如下：将表 11-12 整理成 SPSS 需要的格式，增加一个分组变量，待判家庭的组别未知。在 SPSS 中读入数据，然后在 SPSS 窗口中选择“分析”→“分类”→“判别”，调出判别分析主界面，将左边的变量列表中的 y 选入分组变量中，点击“定义范围”按钮，定义分组变量的取值范围。本例中在最小值和最大值中分别输入 0 和 1。

将 X_1、X_2 选入自变量中，并选择“一起输入自变量”单选按钮，即使用所有自变量进行判别分析。单击“统计量”按钮，指定输出的描述统计量和判别函数系数。选中“函数关系”栏中的 Fisher（F）和未标准化。单击“分类”按钮，定义判别分组参数和选择输出结果。选择“输出”栏中的“个案结果”，和“摘要表”。

单击主对话框中的“保存”，选中“预测组成员”，可以将分类结果保存到数据表中。

相关操作参见图 11-6 到图 11-8。部分输出结果见表 11-13 到表 11-16。

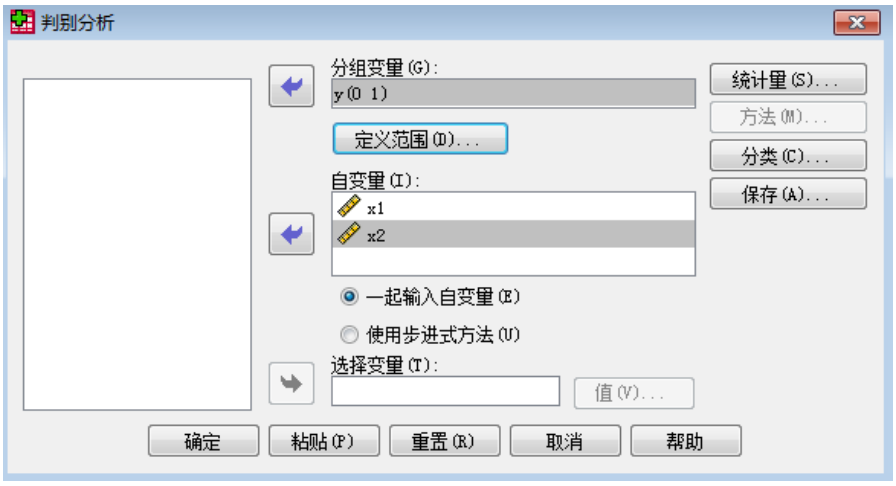


图 11-6 判别分析主对话框

图 11-7 判别分析子对话框：统计量

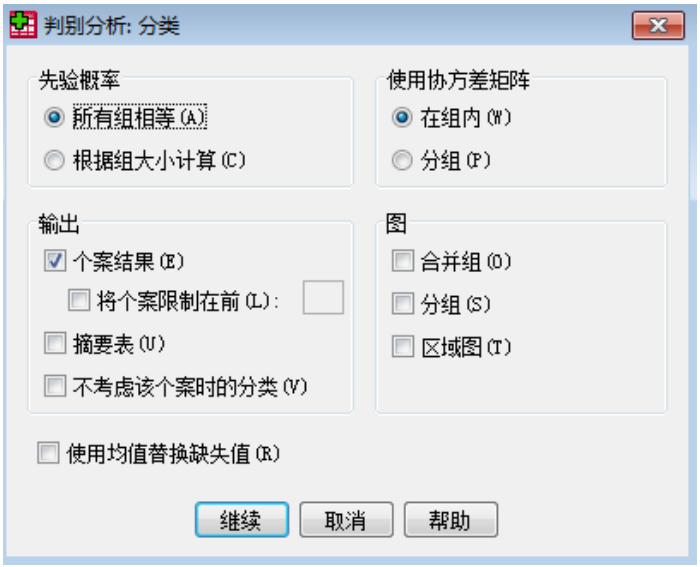


图 11-8 判别分析子对话框：分类

表 11-13 是非标准化的判别式函数，写成公式为：$D=-10.057+0.135x_1+0.724x_2$。表 11-14 是两个类重心的判别函数值。按照判别函数进行分类时，根据判别函数计算样品的判别函数

值，这个值与哪一个类的重心更接近就分入哪个类。

表11-13 典型判别式函数系数

	函数
	1
x1	.135
x2	.724
(常量)	-10.057

非标准化系数

表11-14 组质心处的函数

	函数
y	1
.00	-.862
1.00	.862

在组均值处评估的非标准化典型判别式函数

表 11-15 是 Bayes 判别的两个分类函数。在进行判别时把待判样品的数据代入分类函数，哪个组的值最大就分入哪个组。有几个组就有几个分类函数。根据表 11-15，两个分类函数分别为： $D_0 = -51.210 + 0.794x_1 + 9.459x_2$； $D_1 = -68.546 + 1.027x_1 + 10.707x_2$。

表 11-15 分类函数系数

	y	
	.00	1.00
x1	.794	1.027
x2	9.459	10.707
(常量)	-51.210	-68.546

Fisher 的线性判别式函数

表 11-16 是判别分析的摘要表，这个表可以看出判别分析效果的好坏。从表中我们可以看出，根据判别函数对已知类别的家庭进行回判时，类别为 0 的 12 户家庭中 9 个家庭（75%）分类正确；类别为 1 的 12 户家庭中 10 个家庭（83.3%）分类正确；总体来说 24 个家庭中有 79.2%分类正确。在未知类别两个家庭中分别由 1 个家庭被分入两个类（可以中保存到数据表中的分类结果中知道具体的判别结果）。

表11-16 判别分析的摘要表

		y	预测组成员		合计
			.00	1.00	
初始	计数	.00	9	3	12
		1.00	2	10	12

	未分组的案例	1	1	2
%	.00	75.0	25.0	100.0
	1.00	16.7	83.3	100.0
	未分组的案例	50.0	50.0	100.0

a. 已对初始分组案例中的 79.2% 个进行了正确分类。

小 结

聚类分析与判别分析是两类常用多元分析方法，可将观测对象分成不同的集合或将新的观测值分配到事先分好的各组之中。

聚类分析将个体或对象分类，使得同一类中的对象之间的相似性比与其他类的对象的相似性更强。其目的在于：使类内对象的同质性最大化、类间对象的异质性最大化。

聚类分析通常可以分为两种：Q 型聚类和 R 型聚类。Q 型聚类是对观测值分类，R 型聚类是对变量分类。二者在数学上是对称的，没有本质不同。

判别分析的基本思想是：根据已掌握的每个类别的若干样本的数据信息，建立判别公式和判别准则。当遇到新的样本点时，根据总结出来的判别公式和判别准则，即能判别该样本点所属的类别。

常用的判别分析方法有：距离判别法、Fisher 判别法、Bayes 判别法和逐步判别法。

思考与练习

1. 简述聚类分析与判别分析的区别。

2. SPSS 自带数据集 judges.sav 是中、美、法等 7 个国家的裁判和未经严格培训的体育爱好者在评判体育比赛中对选手的评分情况。请根据评分的差异将他们分为适当的若干类。

3. 根据 16 种饮料的热量、咖啡因、钠及价格四种变量的值，进行聚类分（数据见表 11-17）。要求：给出分成 3 类和 4 类的结果。

表 11-17　16 种饮料的统计数据

热量	*咖啡因*	*钠*	*价格*
207.2	3.3	15.5	2.8
36.8	5.9	12.9	3.3
72.2	7.3	8.2	2.4
36.7	0.4	10.5	4.0
121.7	4.1	9.2	3.5
89.1	4.0	10.2	3.3
146.7	4.3	9.7	1.8
57.6	2.2	13.6	2.1
95.9	0.0	8.5	1.3
199.0	0.0	10.6	3.5
49.8	8.0	6.3	3.7
16.6	4.7	6.3	1.5

38.5	3.7	7.7	2.0
0.0	4.2	13.1	2.2
118.8	4.7	7.2	4.1
107.0	0.0	8.3	4.2

4. 根据死亡率状况和平均预期寿命，15 个地区被分为 3 类。现又得到 4 个地区的相关数据（表 11-18）。表中 X1-X5 分别表示 0 随岁、1 岁、10 岁、55 岁和 80 岁的死亡率，X6 表示平均预期寿命。要求出判别函数表达式，对 4 个待判地区进行归类，并分析判别函数的判别效果。

表 11-18　19 个地区的死亡率和平均预期寿命

X1	*X2*	*X3*	*X4*	*X5*	*X6*	*类别*
34.16	7.44	1.12	7.87	95.19	69.3	1
33.06	6.34	1.08	6.77	94.08	69.7	1
36.26	9.24	1.04	8.97	97.3	68.8	1
40.17	13.45	1.43	13.88	101.2	66.2	1
50.06	23.03	2.83	23.74	112.52	63.3	1
33.24	6.24	1.18	22.9	160.01	65.4	2
32.22	4.22	1.06	20.7	124.7	68.7	2
41.15	10.08	2.32	32.84	172.06	65.85	2
53.04	25.74	4.06	34.87	152.03	63.5	2
38.03	11.2	6.07	27.84	146.32	66.8	2
34.03	5.41	0.07	5.2	90.1	69.5	3
32.11	3.02	0.09	3.14	85.15	70.8	3
44.12	15.02	1.08	15.15	103.12	64.8	3
54.17	25.03	2.11	25.15	110.14	63.7	3
28.07	2.01	0.07	3.02	81.22	68.3	3
50.22	6.66	1.08	22.54	170.6	65.2	待判
34.64	7.33	1.11	7.78	95.16	69.3	待判
33.42	6.22	1.12	22.95	160.31	68.3	待判
44.02	15.36	1.07	16.45	105.3	64.2	待判

第十二章 列联表和对应分析

我们前面介绍的相关分析可以用来分析定量变量之间的关系，但不能用于定性变量的分析。本章介绍的列联表检验和对应分析方法则可以用来分析定性变量之间的关系。

第一节 列联表与独立性检验

【例 12.1】美国的一般社会调查（General Social Survey）是由美国芝加哥大学的民意调查中心进行的一项随机抽样调查，调查对象为 18 岁以上的成年人。调查中获得了居民的婚姻状况和幸福状况方面的数据。下面我们根据 1996 年的调查结果来分析两个变量之间的关系（数据文件 gss96.sav）。在调查中，婚姻状况的取值为已婚、丧偶、离异、分居和未婚（分别用 1-5 表示）；幸福状况的取值为：非常幸福、比较幸福和不太幸福（分别用 1-3 表示）。在 SPSS 软件中打开数据文件，选择“分析”→“描述统计”→“交叉表”，把“婚姻状况”设为行变量，把“幸福状况”设为列变量，可以得到表 12-1 所示的列联表。从表中我们可以看出，从婚姻状况看，已婚人员的比重最高；从幸福状况看，比较幸福的人员比重最高。但从表中我们很难直观地看出两个变量之间的内在联系。

表 12-1 婚姻状况和幸福状况列联表

		幸福状况			
		非常幸福	比较幸福	不太幸福	合计
婚姻状况	已婚	574	726	82	1382
	丧偶	70	149	59	278
	离异	83	292	79	454
	分居	14	73	30	117
	未婚	136	419	99	654
合计		877	1659	349	2885

要研究二维列联表中的两个变量是否相互独立，可以使用我们在非参数检验中讲过χ^2检验。检验的零假设和备择假设为

H_0：婚姻状况和幸福状况这两个变量相互独立；H_1：婚姻状况和幸福状况不相互独立。

假定样本量为 n，列联表有 r 行、s 列，表中各行的合计值分别为$R_i, i=1,2,\Lambda\ ,r$，各列的合计值分别为$C_j, j=1,2\Lambda\ ,s$。每个单元格中的频数为$O_{i,j}$。在零假设成立，即行变量和列变量相互独立时，每个单元格频数的期望值可以按照式（12-1）计算：

$$E_{ij}=\frac{R_i}{n}\times\frac{C_j}{n}\times n=\frac{R_i\times C_j}{n} \qquad (12\text{-}1)$$

显然，如果期望频数E_{ij}和观测频数O_{ij}相差不大，则零假设可能是正确的；如果二者差别很大，则零假设可能不成立。按照式（12-2）构造检验统计量：

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad (12\text{-}2)$$

在零假设成立时，该统计量近似服从自由度为 $(r-1)(s-1)$ 的 χ^2 分布。当该统计量的值很大（或 p 值很小）时，就可以拒绝零假设，认为这两个变量不相互独立。

下面我们用 SPSS 软件来检验婚姻状况与幸福状况的独立性。首先重复绘制列联表的步骤，在 SPSS 软件中打开数据文件，选择“分析”→“描述统计”→“交叉表”，把“婚姻状况”设为行变量，把“幸福状况”设为列变量。接下来单击“统计量”，在弹出的对话框中选中“卡方”，单击“继续”；选择“单元格”，选中弹出对话框中的“期望值”，单击“继续”返回前一个是对话框，单击“确定”（相关操作参见图 12-1）。输出结果见表 12-2、表 12-3。

表 12-2 列出了婚姻状况和幸福状况列联表中各单元格的实际值和期望值。从表中可以看出，部分单元格的实际值和期望值差异较大。从表 12-3 的第一行可以看出，在这个例子中 χ^2 统计量的值为 225.274，相应的 p 值为 0.000。由于 p 值远远小于通常使用的显著性水平，因此检验的结论是拒绝原假设，不能认为婚姻状况和幸福状况相互独立。

图 12-1 SPSS 列联表检验的相关操作

当每个单元格的期望频数都大于 5 时检验统计量近似服从 χ^2 分布。在不满足这一条件时，需要把部分单元格合并，或者使用精确检验：在图 12-1 的对话框中选择“精确…”，进行相应的设置后可以得出精确检验的结果。在精确检验中所涉及的不是 χ^2 分布，而是超几何分布。由于样本很大时超几何分布计算比较慢甚至无法计算，因此在大样本时通常使用 χ^2 统计量。

表12-2　婚姻状况和幸福状况列联表中各单元格的实际值和期望值

			幸福状况			
			非常幸福	比较幸福	不太幸福	合计
婚姻状况	已婚	计数	574	726	82	1382
		期望的计数	420.1	794.7	167.2	1382.0
	丧偶	计数	70	149	59	278

		期望的计数	84.5	159.9	33.6	278.0
	离异	计数	83	292	79	454
		期望的计数	138.0	261.1	54.9	454.0
	分居	计数	14	73	30	117
		期望的计数	35.6	67.3	14.2	117.0
	未婚	计数	136	419	99	654
		期望的计数	198.8	376.1	79.1	654.0
合计		计数	877	1659	349	2885
		期望的计数	877.0	1659.0	349.0	2885.0

表 12-3 婚姻状况和幸福状况间独立性的χ^2检验结果

	值	df	渐进 Sig. (双侧)
Pearson 卡方	225.274[a]	8	.000
似然比	230.166	8	.000
线性和线性组合	137.494	1	.000
有效案例中的 N	2885		

a. 0单元格(.0%)的期望计数少于5。最小期望计数为14.15。

第二节 对应分析

一、对应分析的基本思想

对应分析是一种描述性、探索性的数据分析方法，通常用于列联表的分析，以便用图形的方法观察行变量和列变量取值之间的对应关系。

与因子分析类似，对应分析也是一种数据降维技术。在因子分析中我们可以对变量进行降维，根据因子分析的结果分析各个变量之间的接近程度，也可以对样品（观测）进行降维，根据因子分析的结果分析样品之间的接近程度。而对应分析则可以按照相同的刻度同时对列联表中的行变量和列变量进行降维，用较少的维度（一般选用二维或三维）来代表数据表中的行变量和列变量，从而在同一个空间中用图形方法显示行变量和列变量类别之间的关系。

在表 12-1 的列联表中，把 3 个幸福状况的取值看作 3 维空间中的坐标，我们可以把 5 个幸福状况在 3 维空间中表示出来。如果使用因子分析的方法对 3 个幸福状况进行降维（同时最大限度地保留原始信息），则我们能够在 2 维甚至 1 维空间上把 5 个点表示出来。类似的，把表中婚姻状况的取值看作 5 维空间的坐标值，使用因子分析的方法进行降维，也可以把 3 个幸福状况在低维空间中表示出来。这样，如果能够保证两个因子分析中采用相同的刻度，则可以在同一个坐标系中把幸福状况的 3 个点和婚姻状况的 5 个点绘制出来，通过图形观察两个变量取值之间的关系。按上述方法得到的图形称为对应分析图。在对应分析图中，如果同一变量的不同类别在某个方向上靠得较近，则说明这些类别在该维度上区别不大；落在图形中大致相同区域的不同变量的分类点彼此之间有联系。

当然，为了保证对行和列进行因子分析的结果之间的对应关系，在进行对应分析时并不是根据列联表中的频数直接进行因子分析的，而是先计算相应的频率，再进行必要的变量变

换，之后再用与因子分析类似的方法进行降维。

二、对应分析的软件操作和结果的解释

下面我们通过一个例子说明对应分析的软件操作和结果的解释。

【例 12-2】根据美国的一般社会调查（General Social Survey）1996 年的调查结果分析婚姻状况和幸福状况列两个变量之间的对应关系（数据文件 gss96.sav）。

在 SPSS 软件中打开数据文件，选择“分析”→“降维”→“对应分析”，把“婚姻状况”设为行变量；在弹出的对话框中单击“定义范围”，最小值设为 1，最大值设为 5，单击“更新”、“继续”；然后把“幸福状况”设为列变量，再通过“定义范围”对话框定义其取值范围为 1-3；最后单击“确定”，就可以得到相关输出结果了（相关软件操作参见图 12-2）。

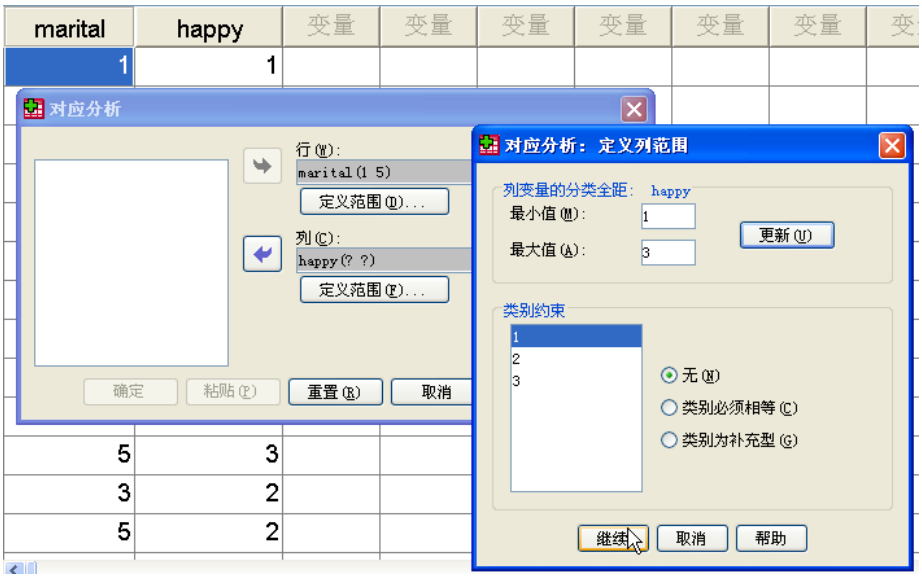


图 12-2 SPSS 对应分析的相关操作

表 12-4 是 SPSS 对应分析的摘要表。在表 12-4 中，“惯量”类似于因子分析中特征值对应的方差；“惯量比例”一栏中，“解释”的惯量比例类似于因子分析中的方差贡献率，“累积”的惯量比例类似于因子分析中的累积方差贡献率，这几个指标反映了每个维度的因子重要性和解释能力。表中的“卡方”是关于列联表行列独立性检验的χ^2 统计量的值，和前面表中的相同。其后面的 Sig 为在行列独立的零假设下的 p 值，注释表明统计量对应的自由度为(5-1)×(3-1)=8。p 值很小说明列联表的行与列之间有较强的相关性。

表 12-4 对应分析的摘要表

维数	奇异值	惯量	卡方	Sig.	惯量比例		置信奇异值	
					解释	累积	标准差	相关 2
1	.272	.074			.944	.944	.017	.064
2	.066	.004			.056	1.000	.021	
总计		.078	225.274	.000[a]	1.000	1.000		

a. 8 自由度

在 SPSS 的输出中还有 12-5 和 12-6 两个表。这两个表分别给出了绘制对应分析图所需要的两套坐标。表 12-5 中的“质量”对应的英文为“Mass”，译为“密度”更为确切，反映

的是每种婚姻状况的人数比例。表中给出了婚姻状况 5 个水平的坐标：已婚 (-0.531, -0.016)，丧偶(0.437, -0.660)，离异 (0.535, 0.168)，丧偶 (0.041, 0.979)，未婚 (0.389, 0.257)。表 12-6 给出了幸福状况三个水平的坐标：非常幸福(-0.684, -0.193)，比较幸福(0.157, 0.207)，不太幸福(0.975, -0.498)。

表12-5 行变量（婚姻状况）各水平的坐标

		维中的得分		
婚姻状况	质量	1	2	惯量
已婚	.479	-.531	-.016	.037
丧偶	.096	.437	-.660	.008
离异	.157	.535	.168	.013
分居	.041	.979	-.331	.011
未婚	.227	.389	.257	.010
有效总计	1.000			.078

表12-6 列变量（幸福状况）的坐标

		维中的得分		
婚姻状况	质量	1	2	惯量
非常幸福	.304	-.684	-.193	.039
比较幸福	.575	.157	.207	.005
不太幸福	.121	.975	-.498	.033
有效总计	1.000			.078

根据表 12-5、12-6 可以得到图 12-3 所示的对应分析图。这个图是在 SPSS 默认输出结果的基础上进行编辑得到的。从图中我们可以很直观地看出“婚姻状况”和“幸福状况”取值之间的对应关系：“已婚”和“非常幸福”最接近；“离异”、“未婚”和“比较幸福”比较接近；而“丧偶”、“分居”和“不太幸福”比较接近。两个变量取值之间的这种对应关系是单独用χ^2检验所无法看出的。

当然，如前所述，对应分析是一种探索性的描述统计方法，并不能保证所有的对应分析结果中两个变量的取值之间都有如此明确的对应关系。

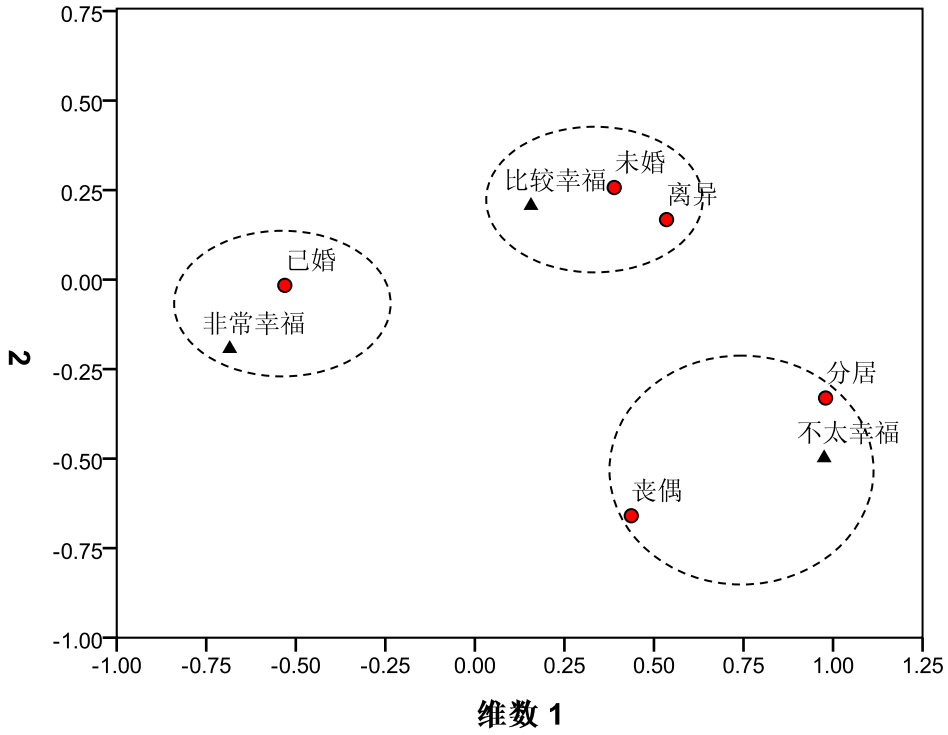


图 12-3 对于婚姻状况和幸福状况关系的对应分析图

小 结

本章首先介绍了列联表中两个定性变量独立性的检验方法（χ^2 检验），然后介绍了对应分析方法。对应分析方法可以让我们通过图形直观地观察两个定性变量取值之间的对应关系。

在基于列联表的独立性检验中，如果行变量和列变量相互独立，则每个单元格中观测到的频数与其期望频数应该比较接近。如果二者之间悬殊较大，则说明零假设是不成立的。基于这种思想可以构造χ^2 检验统计量进行相应的假设检验。

对应分析是一种描述性、探索性的数据分析方法，通常用于列联表的分析，以便用图形的方法观察行变量和列变量取值之间的对应关系。与因子分析类似，对应分析也是一种数据降维技术。对应分析则可以按照相同的刻度同时对列联表中的行变量和列变量进行降维，用较少的维度（一般选用二维或三维）来代表数据表中的行变量和列变量，从而在同一个空间中用图形方法显示行变量和列变量类别之间的关系。

思考与练习

1. 表12-7的数据是592个人头发和眼睛颜色的列联表，请检验着两个变量是否相互独立。

提示：在SPSS软件中录入数据时，数据文件中应该有三个变量：头发颜色、眼睛颜色和频数，并在SPSS中选择“数据”→“加权个案”，把“频数”变量指定为权数；合计值不需要录入到数

据表中。

表12-7 592个人头发和眼睛颜色的列联表

	眼睛颜色				
头发颜色	蓝色	棕色	绿色	淡褐色	合计
黑色	20	68	5	15	108
金色	94	7	16	10	127
棕色	84	119	29	54	286
红色	17	26	14	14	71
合计	215	220	64	93	592

2. 用对应分析方法第1题数据中两个定性变量取值之间的关系。