21世纪高等院校经济与管理核心课经典系列教材

21SHIJIGAODENGYUANXIAO
JINGJIYUGUANLIHEXINKE
JINGDIANXILIEJIAOCAI

统计学概论

TONGJIXUE GAILUN

◎曾五一 主编

首都经济贸易大学出版社

21世纪高等院校经济与管理核心课经典系列教材

21SHIJIGAODENGYUANXIAO
JINGJIYUGUANLIHEXINKE
JINGDIANXILIEJIAOCAI

统计学概论

曾五一 主编

责任编辑 牛志伟
封面设计 小 尘

ISBN 7-5638-1022-6

9 787563 810222 >

ISBN 7-5638-1022-6/F·577 定价：33.00元

21世纪高等院校经济与管理核心课经典系列教材

统计学概论

曾五一 主 编

首都经济贸易大学出版社

·北 京·

图书在版编目(CIP)数据

统计学概论/曾五一主编.—北京:首都经济贸易大学出版社,2003.1
(21世纪高等院校经济与管理核心课经典系列教材)
ISBN 7-5638-1022-6

Ⅰ.统… Ⅱ.曾… Ⅲ.统计学-高等学校-教材 Ⅳ.C8

中国版本图书馆CIP数据核字(2002)第094976号

统计学概论
曾五一 主编

出版发行 首都经济贸易大学出版社
地　　址 北京市朝阳区红庙(邮编100026)
电　　话 (010)65976483 65065761 65071505(传真)
E-mail publish@cueb.edu.cn
经　　销 全国新华书店
照　　排 首都经济贸易大学出版社激光照排服务部
印　　刷 河北三河腾飞印刷厂
开　　本 787毫米×980毫米 1/16
字　　数 526千字
印　　张 26.25
版　　次 2003年1月第1版第1次印刷
印　　数 1~4 000
书　　号 ISBN 7-5638-1022-6/F·577
定　　价 33.00元

图书印装若有质量问题,本社负责调换
版权所有 侵权必究

出版总序

2001年11月10日。

这是一个历史性的庄严时刻。

世界贸易组织第四届部长级会议以全体协商一致的方式，审议并通过了我国加入世贸组织的决定。

它不仅标志着我国成为当今全球最大、最具代表性的国际经济组织的成员，而且标志着我国在融入经济全球化、参与国际经济竞争方面又迈出了决定性的一步，使我国的改革开放和经济发展从此步入了一个崭新的阶段。

入世，机遇与挑战并存。

可以预见到的是，经济领域中的竞争会更加激烈。

经济领域竞争的实质，是人才的竞争；而人才的培养，有赖于教育，尤其是培养高素质专业人才的高等教育。我们还缺乏一大批熟悉现代市场经济运行规律和世贸组织规则，精通专业知识，适应国际竞争需要的高级管理人才和专业人才。

教育是当代科技生产力发展的基础，是科学技术转化为现实生产力的条件，是培养高素质人才和劳动者的根本途径，也是实现管理思想、管理模式、管理手段现代化的重要因素。

《中共中央国务院关于深化教育改革全面推进素质教育的决定》指出："当今世界，科学技术突飞猛进，知识经济已见端倪，国际竞争日趋激烈。教育在综合国力的形成中处于基础地位，国力的强弱越来越取决于劳动者的素质，取决于各类人才的质量和数量，这对于培养和造就我国21世纪的一代新人提出了更加迫切的要求。"

中共中央和国务院的决定为高等教育的改革与发展确定了基本目标和方向。

教材是体现教学内容的知识载体，是进行教学的基本工具，更是培养人才的重要保证。

教材质量直接关系到教育质量，教育质量又直接关系到人才质量。因而，教材质量与人才质量密切相关。

正是由于教材质量在实施科教兴国的发展战略中具有十分重要的作用，我们在策划与组织编写这套教材的过程中倾注了大量的人力、物力和财力。

我们希望奉献给广大教师、学生、读者的是一套经得起专家论证和实践检验的经济与管理专业系列精品教材。

在策划和编写本套教材的过程中，我们贯彻了精品战略的指导思想，使之具有如下特点：

第一，以全面推进素质教育为着眼点，以教育部《普通高等教育教材建设与改革的意见》为指导，面向现代化，面向未来，面向经济全球一体化，充分考虑学科体系的完备性、系统性和科学性，以适应教学和教材改革的需要，满足培养高素质、创造型、复合型人才的需要，并力求教材在内容质量方面具有体系新、内容新、资料新、方法新的特点。

第二，在广泛调查研究的基础上，通过多所高等院校一批有着丰富教学经验的专家教授论证和推荐，优化选题，优选编者。参加本套教材论证和编写的专家教授分别来自北京大学、清华大学、中国人民大学、中国政法大学、对外经济贸易大学、首都经济贸易大学、东北财经大学、西南财经大学、中南财经政法大学、上海财经大学、武汉大学、北京工商大学、南开大学、天津财经大学、天津商学院、南京大学、华中科技大学、北京科技大学、厦门大学、复旦大学、四川大学、中央财经大学等二十余所国内著名或知名高等院校。

第三，在选择教材内容以及确定知识体系和编写体例时，注意素质教育和创新能力、实践能力的综合培养，为学生在基础理论、专业知识、业务能力以及综合素质的协调发展方面创造条件。在确定选题时，一方面根据教育部的指导性意见收入了各相关学科的专业主干课教材，以利于学生掌握各学科及各专业的基础理论、基本知识；另一方面又在充分学习和借鉴国外经典教材的基础上，编选了部分带有前沿性、创新性的专业教材，以利于中外高等教育在课程设置方面的接轨。

第四，考虑到培养复合型人才的实际需要，本套丛书突破了原有的较为狭隘的专业界限和学科界限，在经济学和管理学两大一级学科的统领下，广纳多个分支学科的基础课、专业基础课、专业主干课教材。这些分支学科和专业包括工商管理、金融学、人力资源管理、物流学、广告学、会计学、市场营销、电子商务等。从纵向上看，各学科、各专业的教材自成体系，完整配套；从横向上看，各学科、各专业的教材体系又是开放式的，相互交叉，学科与专业之间没有明确的界限，以便于各院校、各专业根据自身的培养目标设置课程，交叉选用。

本套丛书自身也是开放式的。我们将根据学科发展的需要、教学改革的需要、专业设置和课程调整的需要，不断加以补充和完善。

本套教材不仅是一大批专家教授多年科研成果的总结和教学实践的总结，而且在编写体例上也有所突破和创新，希望它的问世能够对我国经济与管理人才的培养有所帮助。

出版者

2001年11月

前　言

随着我国社会主义市场经济体制的逐步建立与完善，无论是进行宏观的国民经济管理，还是进行微观的企业经营决策，都需要准确地把握有关经济运行的各类数量信息。统计作为认识客观世界数量规律的一个有力工具，在新的形势下，必将发挥更大的作用。因此，《统计学》一直被教育部列为经济和管理类大学本科教育的核心基础课程。

本书是应首都经济贸易大学出版社的邀请，为高等院校工商管理类和经济类专业编写的教材，由经济管理类统计学国家级重点学科——厦门大学计划统计系的教师集体完成。参加本书编写的有：曾五一（第一章、第七章、第九章）、陈珍珍（第二章、第三章、第十一章）、罗乐勤（第八章、第十章、第十二章）、朱平辉（第四章、第五章、第六章）、袁加军（附录一 Excel 概述以及各章中与 Excel 应用有关的内容）。由曾五一教授担任主编，陈珍珍、罗乐勤教授担任副主编，负责全书的设计、修改、总纂和定稿工作。

在本书的编写过程中，我们一方面借鉴了国内外已有的成果，另一方面也做了一些探索，力图使本教材有一些特色和新意，从而更加适合新时期经济、管理类专业的统计教学要求。

首先，我们根据经济、管理类专业的培养目标来设计本书的内容体系。经济、管理类专业统计学教育的目的是：使学生具备基本的统计思想，掌握基本的统计方法，培养应用统计方法分析和解决经济管理中实际问题的能力。由于总学时的限制，非统计学专业的经济、管

理类专业难以开设很多的统计学课程。因此,作为经济、管理类各专业的共同基础课,本书将以往分别开设的《社会经济统计学原理》和《数理统计学》这两门课程有机地结合起来,既介绍具有通用方法论性质的一般统计理论与方法及其在经济管理中的应用,又讨论社会经济领域所特有的一些统计方法问题,同时,还包括了一些经济管理类本科生应当掌握的最基本的国民经济统计学知识。

其次,在本教材的编写过程中,我们根据经济管理类专业的特点,努力贯彻"少而精"和"学以致用"的原则,不仅较大幅度地精简了描述统计学的内容,而且对推断统计学的内容也作了适当的取舍,尽可能做到结构合理,概念明确,条理分明,深入浅出。除十分必要外,本书一般不作过多的数学推导与证明,着重通过实例讲述统计思想,培养和提高学生应用统计方法的能力。

最后,为了便于师生教学互动,提高学生学习的兴趣和学习效率,我们在统计与计算机的结合、教材体例的选择等方面,也作了一些尝试。本书采用最常见的通用软件 Excel 作为实现统计计算和分析的工具。书后附有 Excel 概述,各有关章节中,又结合实例(这些实例大多数采用我国真实的统计数据)讲述 Excel 在各种统计分析中的具体应用。通过多媒体演示和上机操作,培养学生的实际动手能力。在教材体例上,本书各章开篇有内容要点和教学要求提示,章末附有本章小结,对有关教学内容和计算公式作扼要的总结。本书的《思考与练习》不设一般常见的简答题和论述题,而是通过判断题、选择题和计算题来帮助学生掌握有关概念和计算方法。为便于学生自学,同时又为其留有独立思考和独立完成作业的余地,书末附有各章编号为奇数的习题的参考答案。本书出版之后,根据情况,我们还将编制该书的多媒体演示教学软件,提供给采用本书的教师使用。

应当指出,尽管我们为提高教材的质量,作了不少努力,但由于水平有限,加之时间仓促,书中难免有疏漏或错误之处,恳请同行专家和读者不吝赐教,以便今后进一步修改与完善。

本书的编写与出版得到了首都经济贸易大学出版社的大力支持，刘红同志为本书的组稿、编辑做了大量工作，袁加军同志也为本书排版付出了辛勤劳动，在此，一并表示衷心的感谢。

曾五一

2002年6月于厦门大学北村

目 录

第一章 绪论 …… 1
第一节 什么是统计 …… 1
第二节 统计学的种类及其性质 …… 4
第三节 统计学的基本概念 …… 7

第二章 统计资料的搜集与整理 …… 16
第一节 统计资料的搜集 …… 16
第二节 数据整理 …… 25
第三节 频数分布 …… 30

第三章 统计描述 …… 36
第一节 统计图与统计表 …… 36
第二节 分布的集中趋势 …… 45
第三节 分布的离中趋势 …… 59
第四节 分布的偏度和峰度 …… 65
第五节 Excel 在统计描述中的运用 …… 67

第四章 抽样分布与参数估计 …… 80
第一节 频率、概率与概率分布 …… 80
第二节 抽样分布 …… 88
第三节 总体参数估计 …… 94
第四节 抽样设计 …… 103

第五节　Excel 在概率计算与参数估计中的运用 …… 113

第五章　假设检验 …… 120
第一节　假设检验概述 …… 120
第二节　总体参数检验 …… 124
第三节　非参数检验 …… 130
第四节　Excel 在假设检验中的运用 …… 137

第六章　方差分析 …… 143
第一节　方差分析的基本问题 …… 143
第二节　单因素方差分析 …… 146
第三节　双因素方差分析 …… 149
第四节　Excel 在方差分析中的运用 …… 154

第七章　相关与回归分析 …… 161
第一节　相关与回归分析的基本概念 …… 161
第二节　简单线性相关与回归分析 …… 165
第三节　多元线性相关与回归分析 …… 182
第四节　Excel 在相关与回归分析中的运用 …… 191
第五节　非线性相关与回归分析 …… 196

第八章　时间序列分析 …… 208
第一节　时间序列分析概述 …… 208
第二节　时间序列的水平分析 …… 212
第三节　时间序列的速度分析 …… 217
第四节　长期趋势分析 …… 223
第五节　季节变动与循环波动分析 …… 232
第六节　Excel 在时间序列分析中的运用 …… 236

第九章　统计决策…… 247
第一节　统计决策的基本概念…… 247
第二节　完全不确定型决策…… 251
第三节　一般风险型决策…… 256
第四节　贝叶斯决策…… 261

第十章　统计指数…… 272
第一节　统计指数概述…… 272
第二节　综合指数…… 274
第三节　平均指数…… 278
第四节　指数体系与因素分析…… 280
第五节　统计指数的应用…… 287

第十一章　统计综合评价…… 301
第一节　综合评价概述…… 301
第二节　评价指标的选择与数据处理的方法…… 303
第三节　权重的确定与评价结果的综合…… 311

第十二章　国民经济统计概述…… 324
第一节　国民经济核算的基本问题…… 324
第二节　国内生产总值核算…… 331
第三节　国民经济分析常用指标…… 340

附录一　Excel 概述 …… 353
第一节　Excel 的基本操作 …… 353
第二节　使用分析工具库与统计函数…… 367

附录二　常用统计表…… 373
附表 1　二项分布临界值表 …… 373
附表 2　正态分布概率表 …… 375

附表 3　t-分布临界值表 …… 378
附表 4　x^2 分布临界值表 …… 380
附表 5　F 分布临界值表($\alpha=0.05$) …… 382
附表 6　秩和检验表 …… 386
附表 7　游程检验 R 临界值表 …… 387
附表 8　累计法平均增长速度查对表 …… 389

部分习题参考答案…… 396

主要参考书目…… 405

第一章

绪　论

★ 本章学习要点与要求 ★

本章从整体上对统计学的基本轮廓和有关概念作简要介绍，为以后各章的学习奠定基础。具体要求：①正确理解统计的科学涵义和统计研究对象的特点，对统计研究全过程的基本环节及其内容有初步的认识；②了解统计学的种类及其学科性质；③掌握统计学的基本概念，包括统计总体、总体单位、样本、标志、指标、指标体系和统计数据等。

第一节　什么是统计

一、关于统计的涵义

在日常生活中，人们对于“统计”这一术语常常有不同的用法。例如，每年的高考结束后要“统计”考生的总分，这是将其作为一种工作来看待。了解股票的交易状况要看有关成交额和股票指数“统计”，这时又是将其作为数据来运用。而学生们所说的我们正在学习“统计”，则是指一门科学，即统计学。

那么究竟何为统计，这里有必要给出一个比较准确的科学的定义。所谓统计，是人们认识客观世界总体数量变动关系和变动规律的活动的总称，是人们认识客观世界的一种有力工具。与其他认识活动相比，统计的研究对象具有以下特点。

(一)数量性

这是统计研究对象的基本特点,常言说:"数字是统计的语言","数据是统计的原料",指的正是这个意思。但并不是任何一种数量都可以作为统计对象。统计不同于抽象的数学运算。统计数据是客观事物量的反映,统计定量认识必须建立在对客观事物定性认识的基础上,统计研究是密切联系现象的质来研究它的量,并通过量反映现象的质,这一点和数学研究抽象的数量关系是不同的。

(二)总体性

统计的数量研究是对现象总体中各单位普遍存在的事实进行大量观察和综合分析,进而得出反映现象总体的数量特征。例如,进行城镇居民家庭调查,需要对具体的居民家庭进行调查,但是其目的并不在于了解个别居民家庭的生活状况,而是要反映一个国家、一个城市的居民收入水平、收入分配、消费水平、消费结构等等。

(三)变异性

统计研究同类现象总体的数量特征,它的前提是总体各单位的特征表现存在着差异,而且这些差异并不是事先可以预知的。例如,各种股票的价格和成交量每天不同,这才需要对其进行统计,编制股票指数等指标。如果不存在差异,也就无需进行统计。如果说,总体各单位的变异表现出个别现象的特殊性和偶然性,而对现象总体的数量研究,则是通过大量观察,从各单位的变异中归纳概括出它们的共同特征,显示出现象的普遍性和必然性。

统计作为一种社会实践活动已有悠久的历史。据历史记载,我国在西周时就已建立了统计报告制度。在英文中,统计写作 statistics,它与"国家"为同一词根。可以说,自从有了国家,就有统计实践活动。最初,统计只是为统治者了解国家的情况、管理国家提供数量依据。随着社会经济和科技的进步以及统计学自身的发展,统计的应用领域不断扩大。现在,统计不仅被用于经济管理领域,而且在军事、医学、生物、物理和气象等领域也得到广泛的应用。

人们通过统计实践活动所得的成果即统计数据,统计实践活动与统计数据的关系是工作过程与工作成果的关系。工作过程的好坏直接关系到工作成果质量的高低。人们对统计数据的要求是:客观性,即它能反映客观事实而不受任何偏见的影响或任何势力的干扰;准确性,即统计数据的偏差不能超过根据统计研究目的而事先确定的允许误差范围;及时性,即统计数据应及时搜集、及时加工、及时公布。

统计学是在统计实践活动的基础上,自 17 世纪中叶产生并逐步发展起来的一门科学。它是研究如何测定、收集、整理、归纳和分析反映客观现象总体数量的数据,以便给出正确认识的方法论科学。统计学与统计实践活动的关系是理论与实践的关系,理论源于实践,又高于实践,反过来又指导实践。

二、统计研究的基本环节

作为人类认识客观世界的一种活动，统计研究的全过程包括以下几个基本环节。

（一）统计设计

根据所要研究问题的性质，在有关学科理论的指导下，制定统计指标、指标体系和统计分类，给出统一的定义、标准。同时提出收集、整理和分析数据的方案和工作进度等。统计设计是整个统计研究的前期工程，其完成质量直接关系到整个统计研究的质量。搞好统计设计不仅要以统计学的一般理论和方法为指导，而且还要求设计者对所研究的问题本身具有深刻的认识和相关的学科知识。例如，要设计一套较好地反映国民经济运行情况的统计体系与方案，仅有一般的统计方法知识是不够的，设计者还必须具备经济学的知识和理论素养。

（二）收集数据

经过统计设计，形成方案之后，就可以开始收集统计数据。统计数据的收集有两种基本方法。对于大多数自然科学和工程技术研究来说，有可能通过有控制的科学实验去取得数据，这时可以采用实验法。在统计学中有一个专门的分支——实验设计，它就是研究如何科学地设计实验方案，从而使得通过实验采集的数据能够符合分析的目的和要求。而对于社会经济现象来说，一般无法进行重复实验，要取得有关数据就必须到社会总体中去选取足够多的单位进行调查观察，并加以综合研究。如何科学地进行调查是统计学研究的重要内容。本书是为工商管理类专业编写的统计学教科书，由于篇幅的限制，本书将只介绍有关统计调查的理论与方法。

（三）整理与分析

原始的统计数据收集上来之后，还必须经过整理、加工和分析才能真正发挥其作用。在统计研究的这一阶段，所运用的方法包括两大类：描述统计和推断统计。

描述统计是指对采集的数据进行登记、审核、整理、归类，在此基础上进一步计算出各种能反映总体数量特征的综合指标，并用图表的形式表示经过归纳分析而得到的各种有用的统计信息。统计描述是统计研究的基础，它为统计推断、统计咨询、统计决策提供必要的事实依据。统计描述也是对客观事物认识的不断深化过程。它通过对分散无序的原始资料的整理归纳，运用分组法和综合指标法得到现象总体的数量特征，揭示客观事物内在数量规律性，达到认识的目的。

推断统计是在对样本数据进行描述的基础上，利用一定的方法根据样本数据去估计或检验总体的数量特征。在进行统计研究时，常常存在这种情况：我们所掌握的数据由于各种原因只是部分单位的数据或有限单位的数据，而我们所关心的却是整个总体的数量特征。例如，民意测验中某一候选人是否能够当选？全国婴儿的性别比例如何？

某种电子元件的寿命多长？居民的平均消费倾向如何等等。这时就必须利用统计推断的方法来解决。统计推断是逻辑归纳法在统计推理中的应用，所以又称为归纳推理的方法。推断统计是现代统计学的主要内容。

（四）统计资料的积累、开发与应用

通过统计整理和分析，可以得到有关的统计资料。但统计资料的提供并不意味着统计研究的终结。统计的目的在于认识客观世界的规律。对于已经公布的统计资料需要加以积累，还可以进一步地加工，同时结合相关的实质性学科的理论知识去进行分析和利用。如何更好地将统计数据和统计方法应用于各自的研究领域是应用统计学研究的一个重要方面。

以上所述的统计研究的全过程可以用图 1－1 表示。

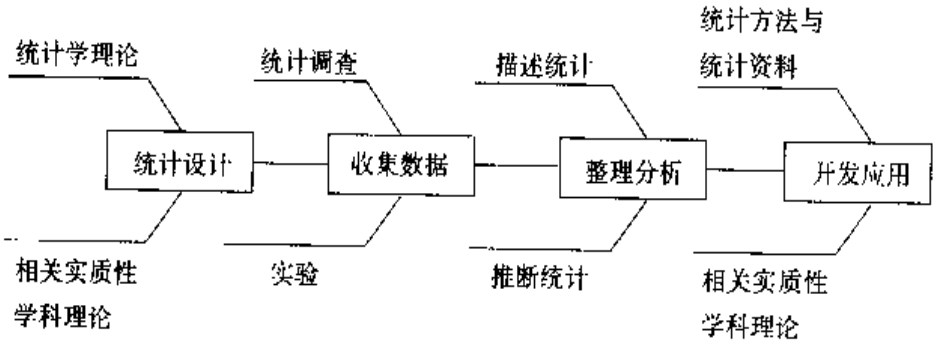


图 1－1　统计研究的全过程

第二节　统计学的种类及其性质

一、理论统计学和应用统计学

统计学历经了 300 多年的发展，目前已经成为跨社会科学领域和自然科学领域的多科性的科学。如前所述，统计学是有关如何测定、收集和分析反映客观观象总体数量的数据，以便给出正确认识的方法论科学。从横向看，各种统计学都具有上述共同点，因面能够形成一个学科“家族”。从纵向看，统计学方法应用于各种实质性科学，与其相结合，产生了一系列专门领域的统计学，如图 1－2 所示。

由此可见，统计学可以分为两大类：一类是以抽象的数量为研究对象，研究一般的

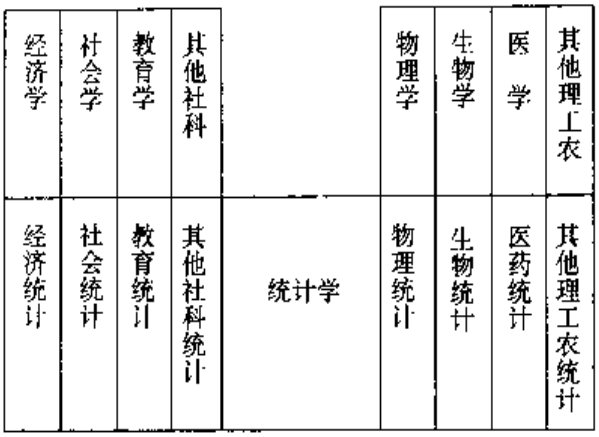


图 1－2 统计学学科体系

收集数据、整理数据和分析数据方法的理论统计学。另一类是以各个不同领域的具体数量为研究对象的应用统计学。

理论统计学(又称数理统计学)把研究对象一般化、抽象化,以数学中的概率论为基础,从纯理论的角度,对统计方法加以推导论证,其中心内容是以归纳方法研究随机变量的一般规律。例如,统计分布理论、统计估计与假设检验理论、相关与回归分析、方差分析、时间序列分析、随机过程理论等等。不论是自然现象还是社会现象,这些方法都是适用的。因此,理论统计学的特点是计量不计质,它具有通用方法论的理学性质。

应用统计学则与各不同领域的实质性学科有着非常密切的联系,是有具体对象的方法论。所谓应用既包括一般统计方法的应用,又包括各自领域实质性科学理论的应用。应用统计学从所研究的领域或专门问题出发,视研究对象的性质采用适当的指标体系和统计方法,以解决所需研究的问题。应用统计学不仅要进行定量分析,还需要进行定性分析。它总是先从现象的质量分析中获得需要考察的指标,建立指标体系,然后采集数据,进行数据处理,并结合对现象的定性分析,得出符合客观现实的结论,作为行动的决策依据。所以应用统计学需要有关的专业实质性科学的理论作指导,它通常具有边缘交叉和复合型学科的性质。

在统计科学发展的过程中,理论统计学和应用统计学总是互相促进,共同提高的。理论统计的研究为应用统计的数量分析提供方法论基础,大大提高统计分析的认识能力,而应用统计学在对统计方法的实际应用中,又常常会对理论统计学提出新的问题,开拓理论统计学的研究领域。

作为工商管理类的学生,所要学习的统计学主要是社会经济统计学。这是一门以社会经济现象的数量方面为特定研究对象的应用统计学。要在社会经济领域应用统计方法,必须解决如何科学地测定社会经济现象,即如何科学地设置指标的问题,这就离

不开对有关社会经济现象的质的研究。要对社会经济问题进行统计分析，也必须以有关的经济与社会理论为指导。因此，社会经济统计学的特点是在质与量的紧密联系中，研究事物的数量特征和数量表现。不仅如此，由于社会经济现象所具有的复杂性和特殊性，社会经济统计学不仅要应用一般的统计方法，而且还需要研究自己独特的方法，如估算的方法、核算的方法、综合评价的方法等等。所以，从总体上看，我们认为社会经济统计学属于社会科学。通过社会经济统计，国家可以准确、及时、全面、系统地掌握国民经济和社会发展情况，对国民经济和社会运行监督和预警，为宏观调控和决策提供依据；企业也可以及时了解商品和要素市场运行的状况和企业自身的经营动态，为企业营销决策、投资理财提供参考。在知识经济和信息的时代里，社会经济统计学具有十分广阔的发展前景。

二、统计学和有关学科的联系与区别

数学是与统计学关系非常密切的一门科学。数学与统计学都是研究数量规律的，都要利用各种公式进行运算。现代统计学中运用了大量的数学理论与数学方法。数学中的概率论，研究随机现象的数量关系和变化规律，从数量方面体现了偶然与必然、个别与一般、局部与总体的辩证关系，为统计学提供了数量分析的理论基础。数学分析的方法适用于一切数量分析，当然也包括统计的数量分析。从某种意义上说，统计学中的理论统计学以抽象的数量为研究对象，计量不计质，其大部分内容也可以看做是数学的一个分支。

统计学虽然与数学有密切的联系，但两者之间也存在本质的区别。从研究对象看，数学撇开具体的对象，以最一般的形式研究数量的联系和空间形式。而统计学特别是应用统计学则总是与客观的对象联系在一起的。统计的过程就是从客观对象中抽出其数量表现，得到有关的数据，然后加以适当的运算，获得一定的结果。在此基础上，还要把这些结果返回到客观对象中去，寻求解释这些结果的意义，提供决策的事实依据。从研究方法看，数学的研究方法主要是逻辑推理和演绎论证的方法，从严格的定义、假设的命题和给定的条件出发，去推证有关的结论。而统计的方法，本质上是归纳的方法，根据实验或调查观察到大量情况，来归纳判断总体的情况。因此，数学家有可能凭借聪明的大脑，从一定的假设出发，冥思苦想开展研究。而统计学家特别是应用统计学家则需要深入实际，进行调查或实验去取得数据，研究时不仅要运用统计的方法，而且还要掌握某一专门领域的知识，才能得到有意义的成果。

统计学中的应用统计学与相关的实质性学科，如经济学等，有十分密切的联系。首先，统计学是开展经济研究不可或缺的重要工具。经济学对经济现象及其发展变化规律进行研究时，除了要做规范性的理论分析和定性分析外，还要进行实证的数量分析。

由于社会经济现象所具有的特殊性，对其数量规律的认识只能通过统计观测去进行。因此，无论是宏观经济研究还是微观经济研究，都需要运用大量的统计方法。通过统计的实证研究，可以帮助人们认识有关的数量规律，同时检验经济学理论的真实性和完善程度。统计归纳分析所获得的新知识常常为实质性学科的研究开辟新的领域，这在经济学的发展历史上是屡见不鲜的。已经有多位经济学家因为在将统计方法成功地运用于有关经济问题的分析方面作出贡献，而获得了诺贝尔经济学奖。其次，经济学等实质性学科对经济统计学等应用统计学起着十分重要的指导作用。不仅统计指标的设定离不开实质性学科的指导，而且应用统计方法也在很大程度上受所研究对象性质的影响。通常是实质性的学科提出了问题，统计学才提出相应的方法，并且才有其用武之地。当然，统计学与相关实质性学科也有着明显的区别。实质性学科研究该领域现象的本质关系并对有关规律作出合理的解释和论证；而统计学只是为实质性学科研究和认识数量规律提供专门的方法和工具，并不直接对规律产生的原因和机理作进一步的分析。例如，利用统计方法对吸烟和不吸烟者患肺癌的数据进行分析，可以得出吸烟是导致患肺癌的重要原因的结论，而吸烟为什么会患癌症，仅仅依靠统计学是无法说明的，而必须由医学作出解释。

第三节 统计学的基本概念

一、总体与总体单位

如前所述，统计学是研究客观观象总体的数量特征和数量关系。因此，首先对统计总体要有一个明确的认识。所谓统计总体就是根据一定目的确定的所要研究的事物的全体。它是由客观存在的、具有某种共同性质的许多个别事物构成的整体。例如，我们要研究全国城镇居民的收支情况，就以全国城镇居民作为一个总体。成千上万不同的城镇居民家庭可以结合在一起构成一个总体，这是因为它们具有共同的性质，即它们都是我国的城镇居民，都有一定的收入和支出，都要消费一定的商品和服务。有了这个总体，我们就可以研究全国城镇居民的各种数量特征，例如人均收入、人均消费等等。

同质性是确定统计总体的基本标准，它是根据统计的研究目的而定的。研究目的不同，则所确定的总体也不同，其同质性的意义也随之变化。例如，研究城镇居民户的生活状况，所有城镇居民户构成了统计总体，凡是城镇居民户都是同质的。如果研究的是城镇居民贫困户的生活状况，那么，贫困线下的城镇居民户就构成了统计总体，贫困

线下的城镇居民户都是同质的，而贫困线上的城镇居民户就是非同质的了。

统计总体还应具备大量性。统计对总体数量特征的研究，其目的是为了探索、揭示现象的规律，而现象的规律只有通过大量观察才能显示出来。因此，统计总体应该由足够数量的同质性单位构成。

总体单位（简称单位）是组成总体的各个个体。根据研究目的的不同，单位可以是人、物、机构等实物单位，也可以是一种现象或活动过程等非实物单位。

总体和总体单位的概念是相对而言的，随着研究目的和总体范围的不同而相互转化。同一个研究对象，在一种情况下为总体，但在另一种情况下又可能变成总体单位。例如，研究全国各省的人口情况时，全国为总体，各省为总体单位；而当要研究某省各县人口状况时，则该省就变成了总体，各县又成了总体单位。

根据所包含的单位的数量不同，总体可以分为有限总体和无限总体两类。有限总体是由有限量的单位构成的总体。例如，全国人口普查，尽管其包含的单位数量很大，但仍然是有限的，所以是有限总体。而当总体单位数难以确定，其数量有可能是无限多的时，便构成了无限总体。例如，要检验某种新工艺是否真正能够改善产品的性能时，由于该新工艺的应用有可能一直延续下去，利用该工艺制造的产品包括已经生产和将要生产的产品，其数量也难以具体确定，因此就属于无限总体。

二、样本

统计研究的目的是要确定总体的数量特征。但是，当总体单位数量很多甚至无限时，不必要或不可能对构成总体的所有单位都进行调查。这时，需要采用一定的方式，从由作为研究对象的事物全体构成的总体（又称母体）中，抽取一部分单位，作为总体的代表加以研究。这种由总体的部分单位组成的集合称为样本（又称子样）。样本也是由一定数量的单位构成的，样本所包含的总体单位数称为样本容量。

样本是推断统计中非常重要的概念。抽取样本时应注意以下问题。

第一，因为抽取样本的目的是为了推断总体，所以样本的单位必须取自同一总体。

第二，从一个总体可以抽取多个样本，样本个数的多少和样本所包含的单位数量与抽样方法有关。

第三，样本的抽取必须排除主观因素的影响，以确保样本的客观性与代表性。

第四，样本特征值是总体特征值的代表，由样本去推断总体的特征总是存在一定的误差。如何科学地从总体中抽取样本，怎样控制样本的代表性误差，是推断统计中研究的主要问题。

三、标志

总体各单位普遍具有的属性或特征称为标志。每个总体单位从不同方面考察都具有许多属性和特征，例如每个工人都具有性别、工种、文化程度、技术等级、年龄、工龄、工资等属性和特征，这些就是工人作为总体单位的标志。统计研究是从登记标志状况开始的，并通过标志综合反映出总体的数量特征。

标志分为品质标志和数量标志两种。品质标志表明单位属性方面的特征，品质标志的表现只能用文字、语言来描述。例如，工人的性别是品质标志，其标志具体表现为男、女。数量标志表明单位数量方面的特征，可以用数值来表现。例如，职工的工龄是数量标志，其标志具体表现为年数。

尽管标志是总体各单位都具有的普遍属性，但各单位有关标志的具体表现却未必相同。而总体各单位在特定时间、地点条件下的具体表现正是统计所关心的问题。如果一个总体中各单位有关标志的具体表现都相同，称之为不变标志。例如，在工人这一总体中，职业这一标志的具体表现都是工人，所以职业便是不变标志。在一个总体中，当一个标志在各单位的具体表现有可能不同时，这个标志便称为变异标志。例如，在工人总体中，各人的工龄可能表现不同，所以工龄便是可变标志。在统计总体中，不变标志和变异标志各自发挥着重要的作用。一个总体至少要有一个不变标志，才能够使各单位结合成一个总体。例如，工人总体中职业的标志是不变的，才能使全体工人构成一个总体。所以，不变标志是总体同质性的基础。如果没有不变标志，那么总体也就不存在。作为总体，同时必须存在变异标志，这表示所研究的现象在各单位之间存在着差异，这才需要进行统计研究。上例中工人的职业标志是不变的，但又存在工资等变异标志，这才需要开展调查统计工作，并计算平均工资等指标，如果各工人的工资水平都一样，就没有必要去统计工资，也不需要用统计方法测度平均工资水平了。

四、统计指标与指标体系

统计指标是反映统计总体数量特征的概念和数值。在《中国统计年鉴》中，人们可以查阅到一系列指标。例如，2000 年我国国内生产总值 89 403.5 亿元，进出口总值 4 475亿美元等，这些指标从某一侧面反映了我国国民经济的数量特征。由此可见，统计指标是由两项基本要素构成的，即指标的概念(名称)和指标的取值。指标的概念(名称)是对所研究现象本质的抽象概括，也是对总体数量特征的质的规定性。所以，确定统计指标必须有一定的理论依据，使之与社会经济或科学技术的范畴相吻合；同时，又必须对理论范畴加以具体化，以便达到量化的目的。例如，工资的涵义在经济学中是明确的，但在实际经济生活中，职工的奖金、津贴和劳保福利是不是应该纳入工资统计的

范围就必须加以具体规定。指标的数值反映所研究现象在具体时间、地点、条件下的规模和水平,不同时间、不同地点或不同条件下,指标的具体数值必然不同。所以,在观察指标数值时,必须了解其具体的时间状态、空间范围、计量单位、计量方法等的限定,同时也要注意由于上述条件的变化而引起的数值可比性的问题。总之,统计指标是统计研究对象的具体化,也是统计对客观事物认识过程的起点。

指标与标志之间存在密切的联系。标志反映总体单位的属性和特征,而指标则反映总体的数量特征。标志和指标的关系是个别和整体的关系。需要通过对各单位标志的具体表现进行汇总和计算才能得到相应的指标。由个体过渡到总体,由标志过渡到指标,是人们认识的深化和发展。因为各个个体的标志之间存在着变异,只有通过大量个体标志的综合,才能利用统计指标获得个体难以显现的信息,反映出现象本质的属性和特征。由于总体和单位的概念会随着研究目的不同而变化,因此指标与标志的概念也是相对而言的。例如,如果所要研究的是全国工业企业的情况,则各企业的职工人数、固定资产、工业增加值等都是总体单位(即各个企业)的标志,而如果研究目的变成某一企业的职工状况,则该企业变成一个总体,企业职工人数变成了统计指标,每个职工的文化程度、技术等级、性别、年龄等就成为标志。

统计指标按其所反映的数量特点不同,可以分为数量指标和质量指标。凡是反映现象总规模、总水平的统计指标都称为数量指标,例如人口总数、企业总数、职工总数、工资总额、国内生产总值、商品流转额、商品进出口总额等等,这些指标反映现象或过程的总规模和水平,所以数量指标也称为总量指标,用绝对数来表示。凡是反映现象相对水平和工作质量的统计指标都称为质量指标,例如职工平均工资、人口密度、工人出勤率等等。质量指标是总量指标的派生指标,用相对数或平均数来表示,以反映现象之间的内在联系和对比关系。

单个统计指标只反映总体的某一个数量特征,说明现象某一侧面的情况。客观现象是错综复杂的,要反映其全貌、描述现象发展的全过程,只靠单个统计指标是不够的,需要设立统计指标体系。统计指标体系是由一系列相互联系的统计指标所组成的有机整体,用以反映所研究现象各方面相互依存、相互制约的关系。例如,为了反映企业生产经营的全貌,需要设立产量、产值、品种、质量、职工人数、劳动生产率、工资总额、原材料、设备、财务成本等多项指标,来组成工业企业统计指标体系。指标体系的设置不但是客观现象的反映,而且也是人们客观认识的结果。随着客观形势的发展变化以及实践经验和理论研究的积累,指标体系也将不断改进更新,逐步完善。

五、统计数据

(一)数据的计量尺度

统计数据是总体单位标志或统计指标的具体数量表现。要对客观现象进行计量，首先必须弄清数据的计量尺度问题。根据对研究对象计量的精确程度不同，人们将计量尺度由低到高、由粗略到精确分为四个层次：定类尺度、定序尺度、定距尺度和定比尺度。

1.定类尺度。定类尺度是最粗略、计量层次最低的计量尺度，它是按照客观现象的某种属性对其进行分类。作为定类尺度的数值只是作为各种分类的代码，并不反映各类的优劣、量的大小或顺序。例如，人口按性别分为男女，用"1"表示男性，用"0"表示女性。定类尺度的主要数学特征是"＝"或"≠"。在统计处理中，对于不同的类别，虽然可以计算单位数，但它不能表明第一类的一个单位可以相当于第二类的几个单位。

2.定序尺度。定序尺度是对客观现象各类之间的等级差或顺序差的一种测度。利用定序尺度不仅可以将研究对象分成不同的类别，而且还可以反映各类的优劣、量的大小或顺序。例如，学生成绩可以分为优、良、中、及格和不及格等五类。在这里，定序尺度虽然无法表明一个优等于几个良，但却能确切地表明优高于良，良又高于中。定序尺度的主要数学特征是"＜"或"＞"。

3.定距尺度。定距尺度是对现象类别或次序之间间距的测度。定距尺度不但可以用数字表示现象各类别的不同和顺序大小的差异，而且可以用确切的数值反映现象之间在量方面的差异。定距尺度使用的计量单位一般为实物单位(自然或物理)或者价值单位。反映现象规模水平的数据必须以定距尺度计量，例如，产品产量、人口数、企业数、国内生产总值等都以定距尺度为计量尺度。定距尺度的主要数学特征是"＋"或"－"。定距尺度在统计数据中占据重要的地位，统计中的总量指标就是运用定距尺度计量的。

4.定比尺度。定比尺度是在定距尺度的基础上，确定相应的比较基数，然后将两种相关的数加以对比而形成的相对数(或平均数)，用于反映现象的结构、比重、速度、密度等数量关系。例如，将一个企业创造的增加值与该企业的职工人数对比，计算全员劳动生产率，以此反映该企业的生产效率。定比尺度的主要数学特征是"×"或"÷"。在统计的对比分析中，定比尺度的运用相当广泛。

(二)数据的类型

根据对客观现象观察的角度不同，统计数据可分为：横截面数据和时间序列数据。横截面数据又称为静态数据，它是指在同一时间对同一总体内不同单位的数量进行观察而获得的数据。时间序列数据又称为动态数据，它是指在不同时间对同一总体的数量表现进行观察而获得的数据。例如，2000年全国各省市自治区的国内生产总值就属于横截面数据，而"九五"期间我国历年的国内生产总值就属于时间序列数据。

在统计中，说明现象的某一数量特征的概念也被称为变量，变量的具体取值是变量

值,统计数据就是统计变量的具体表现。例如,每个企业所拥有的固定资产是不相同的,它是一个变量,各企业固定资产的具体数值可能是200万元,也可能是300万元,这些具体数值即变量值。为了区别,在本书中,凡是变量均用大写的英文字母表示,而变量值则用小写英文字母表示。

根据变量值连续出现与否,变量可分为连续型变量和离散型变量。连续型变量是指变量的取值在数轴上连续不断,无法一一列举,即在一个区间内可以取任意实数值,例如,气象上的温度、湿度,零件的尺寸,电子元件的使用寿命等。离散型变量是指变量的数值只能用计数的方法取得,其取值通常是整数值,可以一一列举,例如,企业数,职工人数等。根据变量的取值确定与否,变量又可分为确定性变量和随机变量。确定性变量是受确定性因素影响的变量,即影响变量值变化的因素是明确的,是可解释和可控制的。随机变量则是受许多微小的不确定因素(又称随机因素)影响的变量,变量的取值无法事先确定。社会经济现象既有确定性变量也有随机变量。统计学所研究的主要是随机变量。

(三)数据的表现形式

统计数据的表现形式主要有以下三种。

1.绝对数。现象的规模、水平一般以绝对数形式表现。绝对数的计量单位一般为实物单位或价值单位,有时也采用复合单位。实物单位可以是自然计量单位,也可以是物理计量单位,如人口数用"人"计量,机器数用"台"计量,对于一些化工产品和燃料,常常还折合成标准实物单位计量。价值单位是以货币形式进行计量,如国内生产总值、进出口总额等就是以价值单位为计量单位的。复合计量单位是由两种或两种以上计量单位复合而成的,如以"吨公里"为货物周转量的计量单位,以"千瓦时"为用电量的计量单位。

绝对数按其反映的时间状态不同,分为时期数据和时点数据。时期数据是反映现象在一段期间内发展过程的总量,它具有连续统计和可相加的特点,其数值大小与所属的时间长短有直接关系,如国内生产总值、进出口总额。时点数据是反映现象在某一特定时点所处的状态,它是采用间断登记方式取得资料的,不宜相加,其数值大小与时点间隔长短无直接关系,如期末人口数、期末在建工程投资额等。

2.相对数。相对数由2个互相联系的数值对比求得。常用的相对数包括:结构相对数、动态相对数、比较相对数、强度相对数、利用程度相对数、计划完成相对数等。结构相对数是在对资料进行分组的基础上,以总体的总量作为比较标准而求得的相对数,如三次产业的增加值占国内生产总值的比重。动态相对数是将不同时期的同类事物的水平加以比较求得的相对数,如国内生产总值的发展速度。比较相对数是将不同单位的同类现象对比而求得的相对数,如1999年北京市的人均消费水平为5 784元,相当

于上海市人均消费水平的56%。强度相对数由两个性质不同但又有联系的绝对数对比求得,如2000年我国的人均发电量为1 073千瓦小时。利用程度相对数是将实际完成水平与可能达到的水平(或设计能力)对比而求得的相对数,如出勤率、设备开工率等。计划完成相对数是将实际完成数与计划数对比用来反映计划完成情况。

3.平均数。平均数反映现象总体的一般水平或分布的集中趋势。关于这部分的内容,将在本书第三章作详细介绍。

本章小结

1.统计是人们认识客观世界总体数量变动关系和变动规律的活动的总称,统计的研究对象具有数量性、总体性与变异性。

2.统计学是研究如何测定、收集、整理、归纳和分析反映客观现象总体数量的数据,以便给出正确认识的方法论科学。统计学与统计实践活动的关系是理论与实践的关系。统计数据是统计实践活动的成果,人们对统计数据的要求是:客观性、准确性和及时性。

3.统计研究的全过程包括统计设计、收集数据、整理分析和积累、开发与应用等基本环节。统计设计不仅要有统计学的一般理论和方法为指导,而且还要求设计者对所要研究的问题本身具有深刻的认识和相关的学科知识。收集统计数据的基本方式包括科学实验和统计调查。统计整理分析的方法可分为描述统计和推断统计两大类。统计资料的积累、开发与应用必须将实质性学科的理论与统计方法相结合。

4.统计学可以分为两大类:一类是以抽象的数量为研究对象,研究一般的收集数据、整理数据和分析数据方法的理论统计学。另一类是以各个不同领域的具体数量为研究对象的应用统计学。前者是具有理论性质的通用方法论科学;后者是有具体对象的方法论,它具有边缘交叉学科和复合型学科的性质。

5.数学为统计学提供数量分析的理论基础,但两者之间在研究对象和研究方法上也存在着本质区别。

6.统计学为经济学等实质性学科提供数量分析工具,经济学等实质性学科对统计学在本领域的应用(包括统计指标的设定、统计方法的选择和统计分析等)起重要的理论指导作用。

7.统计总体是由客观存在的、具有某种共同性质的许多个别单位所构成的整体。总体应具备同质性和大量性。单位是组成总体的个体,总体和单位的概念是相对的。根据总体包括单位的数量不同,总体可分为有限总体与无限总体。

8.样本是由总体抽出的部分单位构成的集合,样本是总体的代表。

9.标志反映总体各单位的属性或特征。标志按所反映单位的特征不同,可分为品质标志和数量标志。标志按其在各单位的表现是否相同,分为变异标志和不变标志。

10.统计指标是通过对标志进行汇总计算而得到的反映总体数量特征的概念和数值,它由指标名称与指标数值两部分组成。统计指标体系是由一系列相互联系的统计指标组成的有机整体,用以反映所研究现象各方面相互依存、相互制约的关系。

11.统计数据是统计实践活动所取得的成果,也是开展统计分析的基础。它是总体单位标志或统计指标的具体数量表现。按对研究对象计量的精确程度不同,统计数据的计量尺度由低到高可分为:定类尺度、定序尺度、定距尺度和定比尺度。说明现象的某一数量特征的概念被称为变量,变量的具体取值是变量值,统计数据是统计变量的具体表现。根据变量取值的连续与否,变量可分为连续型变量和离散型变量。根据对客观现象的观察角度不同,统计数据可分为横截面数据与时间序列数据。统计数据的具体表现形式有:绝对数、相对数与平均数。

思考与练习题

一、判断题

1.统计学是数学的一个分支。(　　)

2.统计学是一门独立的社会科学。(　　)

3.统计学是一门实质性科学。(　　)

4.统计学是一门方法论科学。(　　)

5.描述统计是用文字和图表对客观世界进行描述。(　　)

6.对于有限总体不必应用推断统计方法。(　　)

7.经济社会统计问题都属于有限总体的问题。(　　)

8.理论统计学与应用统计学是两类性质不同的统计学。(　　)

二、分析题

试分析以下几种统计数据所采用的计量尺度属于何种计量尺度:人口、民族、信教人数、进出口总额、经济增长率。

三、选择题

1.社会经济统计学的研究对象是(　　)。

a.社会经济现象的数量方面　　b.统计工作

c.社会经济的内在规律　　d.统计方法

2.考察全国的工业企业的情况时,以下标志中属于不变标志的有(　　);属于数量标志的有(　　)。

a.产业分类　　b.职工人数

c.劳动生产率　　d.所有制

3.要考察全国居民的人均住房面积,其统计总体是:

a.全国所有居民户　　b.全国的住宅

c.各省市自治区　　d.某一居民户

四、问答题

1.请举一个实例说明品质标志、数量标志、质量指标、数量指标之间有怎样的区别与联系。

2.请举一个实例说明统计总体、样本、单位的含义,并说明三者之间的联系。

第二章

统计资料的搜集与整理

★ 本章学习要点与要求 ★

本章阐述统计数据搜集和整理的理论与方法，通过学习，要求：①了解资料的搜集方法；②了解普查、统计报表、抽样调查、重点调查等各种调查组织形式的特点及其适用场合；③掌握统计分组方法；④掌握频数分布数列、累计分布数列的编制方法。

第一节　统计资料的搜集

一、统计资料搜集概述

（一）统计数据搜集的概念

统计数据搜集是指根据统计研究预定的目的和任务，运用科学的调查方法与手段，有计划、有组织地从客观实际采集数据的过程。从统计工作过程的阶段性看，统计数据的搜集处于统计工作过程的基础阶段。

统计数据搜集的方式有两种：一种是直接向调查对象搜集反映调查单位的统计资料（一般称为原始资料）；另一种是根据研究的目的，搜集已经加工、整理过的，说明总体现象的资料（一般称为二手资料）。本节主要阐述统计调查（即搜集原始资料）的理论与方法，同时，也对二手资料的主要来源作简要的介绍。

(二)统计资料搜集的方案设计

在统计资料搜集工作正式开始之前,必须事先设计一个切实可行、周密细致的工作方案,依此方案开展统计调查。统计调查的方案主要包括如下几项内容:

1.确定调查目的。也就是明确统计调查要解决什么问题,只有明确调查目的,资料的搜集工作才能有序地进行。

2.确定调查对象和调查单位。所谓调查对象,是指需要调查的现象总体,该总体是由性质相同的许多调查单位组成的。所谓调查单位,是指所要调查的具体单位,它是进行调查登记的标志的承担者。例如,调查的目的是为了获取外商投资企业的财务状况,那么,所有的外商投资企业就是调查对象,而具体的每一个外商投资企业就是调查单位。明确调查单位,还必须把它与报告单位区别开。报告单位亦称填报单位,它是负责向上报告调查内容、提交统计资料的单位。报告单位一般是在行政上、经济上具有一定独立性的单位,而调查单位可以是个人、企事业单位,也可以是物。根据不同的调查目的,调查单位与报告单位,有时是一致的,有时则不一致。例如,进行工业企业普查,每个工业企业既是调查单位又是报告单位;进行工业企业职工基本状况普查,调查单位是工业企业的每一位职工,而报告单位是每个工业企业。

3.确定调查项目。调查项目就是调查中所要登记的调查单位的特征,即调查单位所承担的基本标志,它由一系列品质标志和数量标志构成。

4.调查表格和问卷的设计。将各个调查项目按照一定的顺序排列在一定的表格上,就构成了调查表。利用调查表,能够有条理地填写需要搜集的资料,还便于调查后对资料进行汇总整理。

调查表一般有两种形式,一种是一览表,另一种是单一表。一览表是把许多调查单位填写在一张表上。单一表是每个调查单位填写一份,可容纳较多标志,一般用于调查项目较多的场合。

问卷调查是一种特殊的调查形式,根据调查目的,在调查对象中随机选择或有意识地确定调查单位,以文字或表格形式了解被调查者的意见,调查者自愿、自由地回答问卷中所提出的问题。调查表格和问卷的设计应简明扼要,以保证所搜集资料的准确。

5.确定调查时间。统计调查时间包括两种涵义,即调查时间和调查期限。调查时间是指调查资料所属的时间,在统计调查中,如果所调查的是时期现象,就要明确规定调查资料所反映的起止日期。例如,调查2002年第二季度的零售商品销售额,则调查时间是从4月1日起至6月30日止的3个月。如果所要调查的是时点现象,调查时间就是规定的统一标准时点。例如,我国第五次人口普查的调查时点是2000年11月1日零时。调查期限是进行调查工作的时限,包括搜集资料和报送资料的工作所需的时间,应尽可能缩短。例如,2000年人口普查规定2000年11月1日至11月10日登记完

毕,则调查期限为11月1日至10日共10天。

6.确定调查的组织实施计划。调查组织工作包括确定调查机构,组织和培训调查人员,落实调查经费的来源和开支办法,确定调查资料的报送方法和公布调查结果的时间。

(三)实验设计的原则

如果想了解一种新技术的应用是否有助于提高劳动效率,或者一种新药物的使用对治疗某种疾病是否有效等问题,所进行的调查往往要运用实验设计法来采集数据。实验设计安排若不得当,所采集的数据就不能反映客观现象的真实情况。因此,运用实验设计采集数据应遵循如下原则:

1.重复性原则。在相同的条件下进行多次重复实验,可以获得更加精确的效应估计量。如果只做一两次实验,用所得的数据作为总体的估计量,其精确度就会较低。重复性实验有利于提高数据的准确性。

2.随机化原则。在进行实验设计时,实验对象的分配、实验次序的安排都是随机的。例如,一种新药的问世,如何设计一个实验测试它的疗效,常用的方法是,进行分组比较。参加实验的对象(一般是患者)随机地被分配到处理组或对照组中,处理组的人接受该药物的治疗,对照组的人服用的是安慰剂。通过对处理组与对照组的反应做比较,从而判断药物是否有效。实验中遵循随机化原则保证了实验的客观与公正。

3.双盲原则。实验设计是在双盲情况下进行的,即实验对象本身不知道自己是被分配在处理组或对照组中,测定反应的人也不知道谁在处理组,谁在对照组。进行双盲实验,是为了防止对反应的测定或评估带有偏见。

二、统计数据的搜集方法

搜集调查对象的原始资料,常用的方法有直接观察法、报告法(通讯法)、采访法、实验调查设计等。任何一种调查都必须采用一定的调查方法去搜集原始资料,即使调查的组织形式相同,其调查方法也可以是不同的。应根据调查目的与被调查对象的具体特点,选择合适的调查方法。

(一)直接观察法

直接观察法是指由调查人员到现场对调查对象进行观察点数和计量。

(二)报告法(通讯法)

报告法一般是由统计工作机构将调查表格分发或电传给被调查者,被调查者则根据填报的要求将填好的调查表格寄回。我国现行的统计报表制度采用的就是这种方法。

(三)采访法

采访法是根据被调查者的答复来搜集统计资料,这种方法又可分为口头询问法和被调查者自填法两种。口头询问法是由调查人员对被调查者逐一采访,当面填答。被调查者自填法,即调查人员把调查表交给被调查者,向被调查者说明填表的要求和方法,并对有关注意事项加以解释,由被调查者按实际情况一一填写,填好后交调查人员审核收回。

(四)登记法

登记法是由有关的组织机构发出通告,规定当事人在某事发生后到该机构进行登记,填写所需登记的材料。例如,人口的出生和死亡的统计及流动人口的统计,就是采用规定当事人或有利害关系的人到公安机关登记的方法。

(五)实验设计调查法

实验设计调查法是用于搜集测试某一新产品、新工艺或新方法使用效果的资料的方法。一般地,对于可以通过科学实验取得资料的,采用实验设计调查法,面对于无法通过科学实验取得资料的,如社会现象,则要大量应用观察法。

随着现代信息技术的发展,计算机、网络、光电技术、卫星遥感、地理信息系统等高新技术已经或正在被广泛地引入统计调查领域中,从而产生了一些新的调查方法。例如,上述各种方法都可以与网络相结合,形成网络调查。又如,可以利用卫星高度分辨辐射计所提供的地面农作物颜色的资料来估计农业产量等等。

三、统计调查的各种形式

统计调查的形式是指组织搜集数据信息资源的方式方法。

统计调查的形式多种多样,按调查的范围划分,可分为全面调查和非全面调查两大类。全面调查是对调查对象的所有单位一一进行调查;非全面调查是对调查对象其中的一部分单位进行调查,以取得调查对象的一部分资料,用来推断总体或反映总体的基本情况。

统计调查按时间标志可分为连续性(经常性)调查和非连续性(一次性)调查。连续性(经常性)调查是指随着研究现象的变化,连续不断地进行调查登记。非连续性(一次性)调查是指间隔一段较长的时间才对事物的变化进行一次性调查。

统计调查按组织形式可分为定期报表和专门调查。定期报表制度是按国家统一规定的表式和内容,定期地向各级领导机构报送统计资料的一种形式。专门调查是为某一专题研究而组织的专项调查。根据相关的标志分类,统计调查的形式如图 2-1 所示。

下面我们分别介绍具体的统计调查形式:

(一)普查

普查是专门组织的一种全面调查,它主要是用以搜集某些不能或不宜用定期报表搜集的统计资料。对国情国力的调查一般采用普查。

普查有两个主要特点:第一,它是非经常性的调查,一般间隔较长的时间才进行一次;第二,它是一种全面调查,比其他任何一种调查形式更能掌握大量、详细、全面的统计资料。

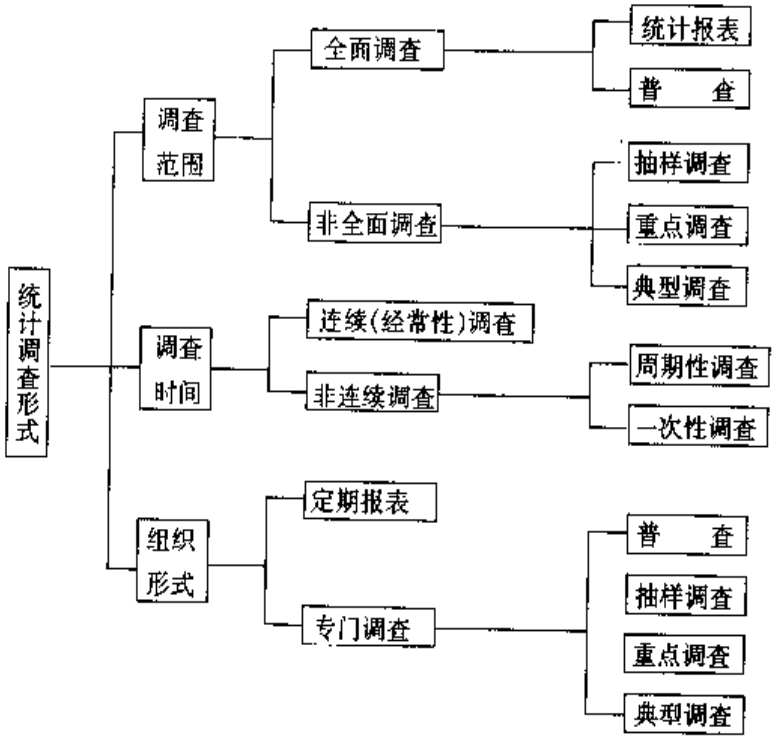


图 2-1 统计调查形式示意图

普查的组织形式有两种:一种是通过组织的普查机构,配备一定数量的普查人员,对调查单位直接进行登记,如我国人口普查就是采用这种形式;另一种是利用调查单位的原始记录和核算资料,结合清库盘点,由调查单位自行填报调查表格,如我国物资库存普查就是采用这种形式。

(二)统计报表制度

统计报表制度是依照国家有关法规,自上而下地统一布置,以一定的原始记录为依据,按照统一的表式、统一的指标项目、统一的报送时间和报送程序,自下而上逐级地定期提供统计资料的一种调查方式。

统计报表的主要特点有:第一,报表资料的来源建立在各个基层单位的原始记录的基础上,基层单位也可利用其资料对生产、经营活动进行监督管理;第二,由于统计报表是逐级上报和汇总的,各级领导部门能获得管辖范围内的报表资料,从而了解本地区、

本部门的经济和社会发展情况;第三,由于统计报表是属于经常性(连续性)调查,调查项目相对稳定,有利于积累资料,并进行动态对比分析。

统计报表包括全面报表与非全面报表。全面统计报表的实施范围是调查对象的全部单位。非全面统计报表的实施范围只要求调查对象中的部分单位填报。

(三)抽样调查

抽样调查是一种非全面调查。它是按随机原则从调查对象中抽取一部分单位作为样本进行观察,然后根据样本数据去推算调查对象的总体特征。

抽样调查有如下几个特点:第一,样本单位按随机原则抽取,排除了主观因素对选样的影响。第二,根据部分调查的实际资料对调查对象总体的数量特征作出估计。根据数理统计的原理,抽样调查中样本指标和相对应的总体指标之间存在着内在联系,而且两者的误差分布也是有规律可循的,因而抽样调查提供了一种用实际调查所得的部分信息来推断总体数量特征的科学方法。第三,抽样误差可以事先计算并加以控制。以样本资料推算总体数量特征,不可避免地会产生误差。但这种误差与其他统计估算所产生的误差不同,它可以根据有关资料事先加以计算,并且通过一定的途径来控制误差的范围,保证抽样推断结果达到预期的可靠程度。

抽样调查的适用范围主要有:第一,对一些不可能或不必要进行全面调查的社会现象,采用抽样调查。例如,电视显像管的耐用时数、轮胎的里程试验等,不可能毁去所有的产品面鉴定其质量,只能采用抽样调查。再如,对居民手存现金情况的调查,不可能也没必要对所有居民逐一观察、经常登记,只能按随机原则选定若干家庭加以调查以获取统计所需的资料。第二,对普查资料进行必要的修正。由于普查涉及面广,工作量大,容易产生登记误差,即出现重复登记或遗漏。通常,可以在普查开始之后,作一次小规模的抽样调查,将抽样调查的结果同原来的普查资料进行核对,计算出差错(重复或遗漏)比率,然后以此作为修订系数,对普查资料进行必要的修正。在复查工作完毕之后,还可利用抽样法对普查质量进行检查。

抽样调查必须遵循以下原则:首先是随机原则。所谓随机原则就是要使所有调查单位都有同样的被抽取的概率。只有按随机原则抽取调查单位,才能保证抽样法符合概率论和数理统计有关定理的要求,从而才能运用这些定理去进行推断。其次是最大抽样效果原则。所谓最大的抽样效果,就是在既定的调查费用下使抽样估计误差最小,或者是在给定的精确度下,使调查费用最少。调查费用是从人力、物力、财力等方面保证调查工作顺利进行的物质基础。提高抽样调查结果的精确度与节省调查费用的要求往往是矛盾的,抽样误差要求愈小,调查费用要求就愈大。从经济角度看,并非任何一种抽样误差最小的方案都是最优的方案。因此,为遵循上述原则,一般要求在给定的误差条件下,选择费用最少的抽样设计方案。

（四）重点调查

重点调查是指在调查对象中，只选择一部分重点单位进行的非全面调查。所谓重点单位，是着眼于现象的量的方面而言，尽管这些单位在全部单位中只是一部分，但它们在所研究现象的标志总量中却占有绝大的比重，在总体中具有举足轻重的作用。对这些单位进行调查，能够从数量上反映整个总体在该标志总量方面的基本情况。重点单位的确定，是组织重点调查的一个重要问题，重点单位的选择主要着眼于它在所研究现象的标志总量中所占的比重。重点单位可能是一些企业，也可能是一些地区、城市。重点调查由于选择的单位较少，因此调查项目就允许多一些，所了解的情况也可以详细一些。通常，当调查任务只要求掌握基本情况，而部分单位又能比较集中地反映研究的项目时，采用重点调查比较适宜。

（五）典型调查

典型调查是一种专门组织的非全面调查。它是根据调查的目的，在对所研究的对象进行初步分析的基础上，有意识地选取若干具有代表性的单位进行调查和研究，借以认识事物发展变化的规律。

四、统计调查体系

在传统的计划经济体制下，我国的统计调查方式以统计报表制度为主，是一种以全面调查、层层汇总为特征的统计调查体系。改革开放以来，随着社会主义市场经济的发展，一方面，社会经济现象空前复杂化，给准确把握统计口径带来困难；另一方面，各类经济组织发展迅速，统计调查单位急剧膨胀，训练有素的基层统计人员数量不足的矛盾日益明显。另外，再加上决策主体和利益主体的多层次化，各方面对统计数字的客观性也存在着重要的影响。这些都使得以全面统计报表为主的调查体系越来越难以满足政府与社会公众对统计信息的需求。为适应社会主义市场经济的要求，目前，我国建立起了以必要的周期性的普查为基础，经常性的抽样调查为主体，同时辅之以重点调查、科学推算和部分全面报表综合运用的统计调查方法体系。

在这一新的统计调查体系中，普查是基础，这是因为只有通过普查，才能收集到全面和详细的数据，同时为开展抽样调查和统计推断提供必要的基础资料。但是，由于普查要耗费大量的人力、物力、财力和时间，无法及时反映社会经济现象日新月异的变化状况，因此，对大量的社会经济现象，必须采用抽样调查方式才能及时地获得各类信息。抽样调查的调查单位少，可以由经过专门训练的人员去完成，同时也便于对某一社会经济现象进行更深入的研究，这样既可以节省调查费用又可以满足对统计时效和统计数据质量的要求。所以，新的统计调查体系要以经常性的抽样调查为主体。重点调查、典型调查和统计报表是我国过去统计实践中常用的方式，在新形势下也仍发挥一定的作

用。

在新的统计调查体系中,还要采用科学的推算方法。所谓统计推算,是在不可能或不必直接通过调查取得资料的情况下,根据已掌握的资料,运用各种统计方法进行科学的估计推算,以间接方式取得所需的资料。统计推算若是对同一时期内的未知项所作的推算,属于静态推算;若是针对未来的时期所作的推算,属于动态推算。统计推算的内容主要包括:从一个现象推算另一现象,从局部推算总体,从现在推算未来。统计推算的方法主要有:比例推算法,因素估算法,平衡估算法,线性插值法,拉格朗日插值法及各种动态数列的预测方法等。

总而言之,在统计调查中,应根据调查的目的和调查对象的特点,灵活地选用不同的调查方式,以及时、准确地获得各种不同的信息。

五、现有统计资料的主要来源

对于应用统计的分析人员来说,相当一部分统计数据不必亲自进行统计调查,可取自有关统计部门和机构发布的统计资料。这方面的资料,可通过两个途径获得:一是从相关的年鉴、期刊和有关出版物上获取;二是从有关网站搜寻。

(一)统计年鉴

1.《中国统计年鉴》。该书由国家统计局编辑、中国统计出版社出版。它是一部全面反映我国国民经济和社会发展情况的资料性书籍,每年9月份左右出版。该书收录了上年全国和各省、自治区和直辖市经济和社会各方面大量的统计数据,以及历史主要年份和近20年的全国统计数据。全书内容分为26个部分:①行政区划和自然资源;②综合;③国民经济核算;④人口;⑤从业人员和职工工资;⑥面定资产投资;⑦能源生产和消费;⑧财政;⑨物价指数;⑩人民生活;⑪城市概况;⑫农业;⑬工业;⑭建筑业;⑮运输和邮电;⑯国内贸易;⑰对外贸易;⑱旅游;⑲金融和保险;⑳教育和科技;㉑文化体育和卫生;㉒其他社会活动及环境保护;㉓㉔㉕分别为港、澳特别行政区及台湾省的主要社会经济指标;㉖我国经济社会统计指标同世界主要国家的比较。该书还附有光盘,提供中文与英文两种文字的统计年鉴电子版。为使用户能方便浏览和使用年鉴,光盘还提供了超文本(HTML)表格和Execl表格两种浏览方式。

2.《国际统计年鉴》(中国统计出版社出版)。它是一部综合性的国际经济、社会统计资料,收录了全球多达160个国家和地区的统计资料,尤其是对其中的40多个主要国家的经济和社会发展状况以及世界主要企业的基本情况作了更为详细的介绍。该书分为17个部分:①世界概况;②自然资源和环境保护;③人口和人均主要经济指标;④国民经济核算;⑤农业;⑥工业和建筑业;⑦运输和通讯;⑧就业、工资;⑨财政金融;⑩国际收支、债务;⑪价格指数;⑫国际贸易和旅游;⑬社会和军事;⑭性别差异;⑮家庭收

支和居民消费;⑯部分国家分地区的主要指标;⑰重要企业指标。

3.《地方统计年鉴》(中国统计出版社出版)。由各省、自治区和直辖市以及经济特区的统计局编纂,比较详细地反映各省、自治区和直辖市以及经济特区的社会、经济和科技等发展变化的情况。

4.《中国县(市)社会经济统计年鉴》(中国统计出版社出版)。由国家统计局农村社会调查总队编纂,主要内容包括:区域分析统计图,各县(市)的主要经济指标,分区域社会发展基本情况,按主要经济指标分组的社会经济基本情况。

5.《中国金融年鉴》(中国金融协会编)。该书记述了金融发展的基本情况,提供有关金融统计的资料,包括:各层货币供应量,银行概述,特定存款机构的资产负债表。

6.《中国人口统计年鉴》(中国统计出版社出版)。该书是人口状况资料性年刊,书中搜集了全国各省、自治区和直辖市大量的人口数据,以及世界各国的人口数据。

7.《中国统计摘要》(中国统计出版社出版)。收录了反映我国经济社会发展的主要统计数据,一般比统计年鉴早若干月出版。

(二)有关期刊

上述各种年鉴所提供的资料较为详细、全面、系统,但时效性较差。反映我国经济社会动态的数据,则可由以下刊物取得。

1.《中国经济数据分析》(中国信息中心经济预测部编写)。该期刊提供了当季(或月)我国的GDP增长率,工业生产指数,企业效益指标,固定资产投资、外贸出口和市场销售的规模和速度指标,居民消费水平等数据。

2.《经济预测分析》(国家信息中心编写)。它提供了有关国民经济运行状况的资料。

(三)有关网站

在计算机与网络技术飞速发展的今天,互联网成为获取统计数据的重要途径。目前可获取反映中国经济社会状况统计数据的网站,主要有:

1.中国统计信息网,网址为 www.stats.gov.cn,由国家统计局主办。主要内容有:统计公报,统计数据,统计分析,统计法规,统计管理和数据直报等。在该网站也可搜寻有关统计年鉴的数据资料。

2.国研网,网址为 www.drcnet.com,由国务院发展研究中心主办。提供的主要信息有:宏观经济、区域经济、金融市场、行业经济及企业经济的相关数据资料。

3.中国经济信息网,网址为 www.cei.gov.cn,由国家计委、国务院西部办等部门主办。从该网站可搜寻我国经济发展及各地区经济发展的数据资料。

4.中国经济时报网,网址为 www.cet.com.cn/2002/default.com,由国务院发展研究中心主办,从该网站可搜寻有关经济生产、资本市场的信息。

第二节 数据整理

一、数据整理的内容与程序

数据整理从广义上讲，包括两种整理。第一种是对统计调查所搜集到的各种数据进行分类和汇总，称为汇总性整理；第二种是对现成的综合统计资料进行整理。本节所说的数据整理，指的是第一种整理。

（一）统计数据整理的内容

1.根据研究目的设计整理汇总方案。统计汇总方案的设计包括两方面：一是对于总体的处理方法，主要是如何进行统计分组；二是确定用哪些统计指标来说明总体。

2.根据汇总方案，对各个调查项目的资料进行汇总，通过汇总计算各项指标。

3.通过统计表或统计图的形式，描述整理的结果。

（二）数据整理的程序

1.统计资料的审核。在整理之前，必须对原始数据认真地审核，检查原始数据的完整性与准确性。

2.资料的分组和汇总。对全部调查数据资料，按其性质和特点，进行分组归类，综合汇总形成各项统计指标。统计分组和统计指标是整理的中心工作。

3.编制统计表或绘制统计图。

4.统计资料的积累、保管和公布。

二、统计分组

（一）统计分组的概念与种类

1.统计分组的概念。根据统计研究的目的和客观现象的内在特点，按某个标志（或几个标志）把被研究的总体划分为若干个不同性质的组，称为统计分组。统计分组的对象是总体。统计分组的标志可以是品质标志，也可以是数量标志。

从分组的性质来看，分组兼有分和合的双重含义。对于现象总体而言，是“分”，即把总体分为性质相异的若干部分；而对于单位而言，又是“合”，即把性质相同的许多单位结合为一组。对于分组标志而言，是“分”，即按分组标志将不同的标志表现分为若干组；而对于其他标志而言，是“合”，即在一个组内的各单位即使其他标志表现不相同也只能结合在一组。由此可见，选择一种分组方法，突出了一种差异，显示了一种矛盾，必

然同时掩盖了其他差异,忽略了其他矛盾。不同的分组方法,可能得出不同的结论。因此,统计分组必须先对所研究现象的本质作全面地、深刻的分析,确定所研究现象类型的属性及其内部差别,而后才能选择反映事物本质的正确的分组标志。

2.统计分组的原则。统计分组,必须遵循两个原则:穷尽原则和互斥原则。

所谓穷尽原则,就是使总体中的每一个单位都应有组可归,或者说各分组的空间足以容纳总体中所有的单位。

所谓互斥原则,就是在特定的分组标志下,总体中的任何一个单位只能归属于某一组,而不能同时或可能归属于几个组。

3.统计分组的种类。统计分组可以按照不同的标志进行分类。分组的标志是划分资料的标准和依据,分组的标志选择是否得当,关系到能否正确地反映总体数量特征及其变化规律。统计分组主要有如下几种。

(1)按分组标志的多少,可分为简单分组和复合分组。简单分组就是对研究现象按一个标志进行分组,它只能从某一方面说明和反映事物的分布状况和内部结构。许多简单分组从不同角度说明同一个总体,就构成一个平行的分组体系。

许多场合,要用两个或两个以上标志分组,即先按一个标志分组,在此基础上再按第二个标志分小组,再层叠地按第三个标志分成更小的组,这被称为复合分组。例如,对国内生产总值先按三次产业分组,其后对各产业按行业分组,再对各行业按经济类型分组,最终形成如下的复合分组体系:

国内生产总值
- 第一产业
- 第二产业
  - 工业
    - 国有工业
    - 非国有工业
  - 建筑业
- 第三产业

(2)按分组标志的性质不同,分为品质分组(或称属性分组)和数量分组(或称变量分组)。

品质分组就是按品质(或属性)标志进行分组。一般地,对于以定类尺度(列名尺度)或定序尺度(顺序尺度)计量的,采用品质分组。例如,职工按性别分组,企业按经济类型分组等。数量分组就是按数量标志分组,数量标志的变异性体现在它不断变动自身的数量上,故也称为变量分组。例如,企业按产值、工人数分组。品质分组所形成的数列称为品质数列,变量分组所形成的数列称为变量数列。

(3)按分组的作用和任务不同,分为类型分组、结构分组和分析分组。

把复杂的现象总体,划分为若干个不同性质的部分,就是类型分组。例如,我国全社会消费品零售额分为国有及国有控股商业零售额、集体商业零售额、私营及个体商业零售额和其他类型商业零售额。

在对总体分组的基础上计算出各组对总体的比重,藉此研究总体各部分的结构,即结构分组。类型分组与结构分组往往紧密地联系在一起。

为研究现象之间依存关系而进行的统计分组即分析分组。例如,为研究工人的劳动生产率与产值之间、商品流通费用率与商品销售额之间的依存关系,就要按分析分组法来研究它们之间的关系。

(二)统计分组的方法

要进行科学的分组,必须选择适当的分组标志。分组标志选择的正确与否,关系到能否正确地反映总体的性质特征、完成统计研究的目的任务。选择不科学的分组标志,不但无法显示现象的本质特征,而且会把不同性质的事物混淆在一起,从而歪曲了客观现象的实际情况。

分组标志确定之后,可以具体进行分组。如前所述,按分组标志的性质不同,分为品质分组和数量分组,这两种分组的具体处理方法存在不同的特点。

1.品质分组。有些品质分组比较简单,分组标志一经确定,组的名称和组数也随之确定,例如人口按性别分为男女两组;有些品质分组还取决于统计分析对分组层次的不同要求,例如我国把社会经济各部门划分为第一产业、第二产业和第三产业,第一产业还可细分为农业、林业、畜牧业和渔业等,对于这种类别繁多的分组又称为分类。对于这类问题,统计工作中采用统一的分类标准。这样的具体规定分类(组)标准,为统计整理提供了统一的依据。

2.数量分组。按数量标志分组,应注意如下两个问题:首先,分组时各组数量界限的确定必须能反映事物质的差别;其次,应根据被研究的现象总体的数量特征,采用适当的分组形式,确定适宜的组距、组限。

(1)单项式分组与组距式分组。单项式分组是按数量标志分组,数量标志的表现就是变量的取值,即标志值,又称变量值。单项式分组,就是用一个变量值作为一组,形成单项式变量数列。例如,育龄妇女按其生育子女的存活数分组,可分为0个、1个、2个、3个、4个、5个等6组。单项式分组一般适用于离散型变量且变量变动范围不大的场合。离散型变量是指所描述对象的标志值可以按一定次序一一列举(通常取整数值)的数量标志。

组距式分组,就是将变量依次划分为几段区间,一段区间表现为从"……到……"距离,把一段区间内的所有变量值归为一组,形成组距式变量数列。区间的距离就是组

距。例如,居民居住水平情况按人均居住面积分组分为:6平方米以下,6~10平方米,10~14平方米,14平方米以上等4组。对于连续型变量或者变动范围较大的离散型变量,适宜采用组距式分组。

(2)间断组距式分组和连续组距式分组。在组距式分组中,每组包含许多变量值,每一组变量值中,其最小值为下限,最大值为上限。组距是上下限之间的距离,相邻两组的界限,称为组限。凡是组限不相连的,称为间断组距式分组。例如,工人按其看管机器台数分组,分为1台、2台、3台、4台等4组。凡是组限相连(或称相重叠的),即以同一数值作为相邻两组的共同界限,称为连续组距式分组。例如,工人按工时定额完成程度分组,分为90%~100%,100%~110%,110%~120%等组。

连续型变量,只能采用连续组距式分组。在连续组距式分组中,存在以同一个数值作为相邻两组共同的界限,根据统计分组必须遵循的原则之一:"互斥原则"。因此,统计上规定,凡是总体某一个单位的变量值是相邻两组的界限值,这一个单位归入作为下限值的那一组内,即所谓"上限不在内"原则。例如学生按成绩分组,把70分的学生归入70~80分组内,把80分的学生归入80~90分组内。

(3)等距分组与异距分组。按数量标志进行组距式分组,还可分为等距分组和不等距(或称异距)分组。等距分组就是标志值在各组保持相等的组距,即各组的标志值变动都限于相同的范围。凡是在标志值变动比较均匀的情况下,都可采用等距分组。等距分组有很多好处,它便于计算,便于绘制统计图。

异距分组即各组的组距不相等。一般地,异距分组适用于如下几种场合:第一,标志值分布很不均匀的场合。第二,标志值相等的量具有不同意义的场合。例如,生命的每一个月对于新生婴儿和对于成年人是大不一样的,此时,若按年龄分组进行人口疾病研究,应采用异距式分组。第三,标志值按一定比例发展变化的场合。

(三)组距式分组中相关指标的计算

1.组距。在组距式分组中,组距是上下限之间的距离。连续组距分组的组距计算公式是:

$$组距=本组上限-本组下限 \tag{2.1}$$

对于间断式分组的组距大小的计算,采用如下公式:

$$组距=本组上限-前组上限 \tag{2.2}$$

或

$$组距=本组下限-前组下限 \tag{2.3}$$

或

$$组距=本组上限-本组下限+1 \tag{2.4}$$

连续组距式分组的组距大小,也可根据公式(2.2)或(2.3)求得。

2.组数。组距的大小直接关系到组数的多少,组距大,组数就少;组距小,组数就多。这里,向大家介绍一种确定组数和组距的经验公式,这一公式是美国学者斯特杰斯

(H. A. Sturges)创用的,称为斯特杰斯经验公式,即:

$$n = 1 + 3.3\log N \tag{2.5}$$

$$d = \frac{R}{n} = \frac{x_{max} - x_{min}}{1 + 3.3\log N} \tag{2.6}$$

式中,n 为组数;N 为总体单位数;d 为组距;R 为全距,即最大变量值 x_{max}与最小变量值 x_{min}之差。

根据这一公式,可以得出如下(见表 2-1)的组数参考标准。

表 2-1 分组组数参考表

N	15~24	25~44	45~89	90~179	180~359
n	5	6	7	8	9

上述公式及表中数据仅供参考,不能生搬硬套。实际分组时采用组数的多少应依据所研究资料的性质而定。

3. 组中值。上下限之间的中点数值称为组中值,组中值的计算公式为:

$$组中值 = \frac{上限 + 下限}{2} \tag{2.7}$$

在计算平均指标或进行其他统计分析时,常以组中值来代表各组标志值的平均水平。当各组标志值均匀分布时,组中值就能较强地代表各组标志值的水平,因此,分组时,应尽可能使组内各单位标志值分布均匀。

有时候,连续型变量按离散型变量表示,组距数列的编制采取相邻组限不重叠的形式,组中值的确定应考虑到连续型变量自身的特点。年龄就是比较典型的例子,它实质上是连续型变量,习惯上用整数表示。例如,一群大学生分为 17~19 岁、20~22 岁两组,则组距为 3 岁,组中值分别为 18.5 岁和 21.5 岁。因为第一组应包括 19 岁又不到 20 岁的大学生,上限应视为 20 岁。同样道理,第二组上限应视为 23 岁。

4. 开口组的组距与组中值。在编制组距式变量数列时,使用"××以上"或"××以下"这样不确定组距的组,称为开口组。例如,反映××工业企业工人生产定额完成情况,按生产定额完成程度分组,分为 90% 以下,90% ~100% ,100% ~110% ,110% 以上。开口组的组距是以相邻组的组距为本组的组距,如上例,90% 以下的组,因相邻组的组距为 10% (=100% -90%),故第一组可视为 80% ~90% ,其组中值为(80+90)÷2=85,即 85% ;110% 以上的组距以邻组的组距 10% 为本组组距,视为 110% ~120% ,组中值为(110+120)÷2=115,即 115% 。

第三节　频数分布

一、频数分布的基本概念

(一)频数分布

在统计分组的基础上,将总体所有的单位按某一标志进行归类排列,称为频数分布,或次数分布。频数分布是统计整理的一种重要形式,通过对零乱的、分散的原始资料进行有次序的整理,形成一系列反映总体各组之间单位分布状况的数列,即分布数列。根据分组标志特征的不同,分布数列可分为两类:按品质标志分组所形成的数列即品质分布数列,亦称品质数列;按数量标志分组所形成的数列叫变量分布数列,亦称变量数列。

(二)分布数列的两个要素

分布数列由两个要素构成:一个是总体按某标志所分的组;另一个是各组所出现的单位数,即频数,亦称次数。就变量数列而言,总体按数量标志分组,分组标志在各组有不同的数量表现,形成标志值数列,亦称变量,一般用 x 表示;频数(次)用 f 表示。

二、变量数列的编制

(一)变量数列的编制方法

【例2　1】 对产品生产过程进行质量控制,抽取 50 个电子元件调查其耐用时数(单位:小时),具体资料如下:

1 180	1 010	1 230	1 100	1 180	1 580	1 210	1 460	1 170	1 080
1 050	1 100	1 070	1 370	1 200	1 630	1 250	1 360	1 270	1 420
800	1 030	870	1 150	1 410	1 170	1 230	1 260	1 380	1 510
1 010	860	810	1 130	1 140	1 190	1 260	1 350	930	1 420
1 080	880	1 050	1 250	1 160	1 320	1 380	1 310	1 250	1 270

对上述资料采用等距分组,分为 8 组,组距为 100,以 800 为第一组下限。经过整理,得出计算结果如表 2 - 2 所示。

表 2-2 50 个电子元件耐用时数资料表

按耐用时数分组(x)	频数(f_i)(小时)	频率($f_i/\sum_i f_i$)
800～900	5	0.10
900～1 000	1	0.02
1 000～1 100	8	0.16
1 100～1 200	11	0.22
1 200～1 300	11	0.22
1 300～1 400	7	0.14
1 400～1 500	4	0.08
1 500 以上	3	0.06
合计	50	1.00

表 2-2 中第 1 列是变量,第 2 列是各组出现的次数,即频数,各组频数之和等于总体单位数,第 3 列是频率。

(二)相关指标的计算

1.频率,即各组频数与总体单位总和之比,它反映了各组频数的大小对总体所起的作用的相对强度。计算公式如下:

$$频率=f_i/\sum_i f_i \quad (f_i 表示第 i 组频数) \tag{2.8}$$

通过对总体各单位分组而形成变量数列,显示了各单位标志值在各组间的分布状况,从而使杂乱无章的原始数据显示出一定的规律性。从表 2-2 中可以看出,耐用时数在 1 000～1 300 的电子元件占总数的 60%,而耐用时数低于 1 000 的和高于 1 400 的所占比重较小,呈现出一种近似"两头小,中间大"的钟型分布特征。

2.频率具有如下两个性质:

(1)任何频率都是界于 0 和 1 之间的一个分数,即:$0\leqslant f_i/\sum_i f_i\leqslant 1$

(2)各组频率之和等于 1,即:$\sum_i f_i/\sum_i f_i=1$

3.频数密度与频率密度。对于异距分组,由于各组次数的多少还受到组距不同的影响,各组的频数可能会随着组距的扩大而增加,随着组距的缩小而减少。为消除异距分组所造成的这种影响须计算频数密度(或称次数密度),频数密度的计算公式如下:

$$频数密度=频数/组距 \tag{2.9}$$

$$频率密度=频率/组距 \tag{2.10}$$

各组频数密度与各组组距乘积之和等于总体单位数,各组频率密度与各组组距乘积之和等于 1。

三、累计频数与累计频率

累计频数(或频率)可以是向上累计频数(或频率),也可以是向下累计频数(或频率)。向上累计频数(或频率)分布,其方法是先列出各组的上限,然后由标志值低的组向标志值高的组依次累计。向上累计频数表明某组上限以下的各组单位数之和是多少,向上累计频率表明某组上限以下的各组单位数之和占总体单位数的比重。向下累计频数(或频率)分布,其方法是先列出各组的下限,然后由标志值高的组向标志值低的组依次累计。向下累计频数表明某组下限以上的各组单位数之和是多少,向下累计频率表明某组下限以上的各组单位数之和占总体单位数的比重。

累计频数分布具有如下两个特点:①第一组的累计频数等于第一组本身的频数;②最后一组累计频数等于总体单位数。累计频率同样也具有两个特点:①第一组的累计频率等于第一组本身的频率;②最后一组的累计频率等于1。

现仍以50个电子元件耐用时数的资料为例,进行向上和向下累计,其结果如表2-3所示。

表2-3 50个电子元件耐用时数累计表

向上累计					向下累计				
电子元件耐用时数分组上限	频数	累计频数	频率(%)	累计频率(%)	电子元件耐用时数分组下限	频数	累计频数	频率(%)	累计频率(%)
900	5	5	10	10	800	5	50	10	100
1 000	1	6	2	12	900	1	45	2	90
1 100	8	14	16	28	1 000	8	44	16	88
1 200	11	25	22	50	1 100	11	36	22	72
1 300	11	36	22	72	1 200	11	25	22	50
1 400	7	43	14	86	1 300	7	14	14	28
1 500	4	47	8	94	1 400	4	7	8	14
1 600	2	49	4	98	1 500	2	3	4	6
1 700	1	50	2	100	1 600	1	1	2	2
合计	50	—	100	—	合计	50	—	100	—

由表2-3可知,电子元件耐用时数在1 000小时以下的有6个,占总数12%,电子元件耐用时数在1 200小时以下的有25个,占总数50%;电子元件耐用时数在1 000小时以上的有44个,占总数88%,电子元件耐用时数在1 200小时以上的有25个,占

总数 50% 等。

本章小结

1.统计数据搜集是根据统计研究预定的目的和任务,运用相应科学的调查方法与手段,有计划、有组织地从客观实际搜集资料的过程。

2.调查对象是指需要调查的现象总体,该总体是由性质相同的调查单位组成的。调查单位是进行调查登记的标志的承担者。报告单位是负责向上报告调查内容、提交统计资料的单位。

3.调查时间是指调查资料所属的时间。调查期限是进行调查工作的时限,包括搜集资料和报送资料的工作所需的时间。

4.统计调查按调查的范围划分,可分为全面调查和非全面调查两大类。

5.普查是一种非经常性的全面调查,通过普查可以掌握大量、详细、全面的资料。

6.统计报表制度是依照国家有关法规,自上而下地统一布置,按照统一的表式、统一的项目、统一的报送时间和程序,自下而上逐级地定期提供统计资料的一种调查方式。

7.抽样调查是按随机原则从调查对象中抽取一部分单位作为样本进行观察,然后根据所获得的样本数据,对调查对象总体特征作出具有一定可靠程度的推算。

抽样调查的特点:第一,按随机原则抽取样本;第二,根据部分调查的实际资料对调查对象总体的数量特征作出估计;第三,抽样误差可以事先计算并加以控制。

抽样调查的适用范围主要有:第一,对一些不可能或不必要进行全面调查的社会现象进行调查;第二,对普查资料进行检查和必要的修正。

抽样调查必须遵循随机原则与最大抽样效果原则。

8.重点调查是在调查对象中,只选择一部分重点单位进行的非全面调查。

9.我国新的统计调查体系目标模式是建立以必要的周期性的普查为基础,经常性的抽样调查为主体,同时辅之以重点调查、科学推算和部分全面报表综合运用的统计调查方法体系。

10.统计分组与统计指标是统计整理的中心。统计分组应遵循穷尽与互斥两个原则。数量分组有单项式分组和组距式分组,组距式分组又分为间断组距式与连续组距式分组、等距与异距分组。对于组距式分组,须计算组距、组数、组中值。开口组的组距是以相邻组的组距为本组的组距。

11.分布数列由两个要素构成,一个是总体按某标志所分的组,另一个是各组所出

现的单位数，即频数。

对于变量数列，总体按数量标志分组，分组标志在各组有不同的数量表现，形成标志值数列。

频率的性质：频率是界于0和1之间的一个分数，各组频率之和等于1。

12.频数密度与频率密度：为消除异距分组所造成的影响须计算频数密度。频数密度＝频数/组距，频率密度＝频率/组距。

13.向上累计频数(或频率)分布，其方法是先列出各组的上限，然后由标志值低的组向标志值高的组依次累计。向下累计频数(或频率)分布，其方法是先列出各组的下限，然后由标志值高的组向标志值低的组依次累计。

思考与练习题

一、选择题

1.统计调查对象是(　　)。

a.总体各单位标志值　　b.总体单位

c.现象总体　　d.统计指标

2.统计调查方法体系中，作为"主体"的是(　　)。

a.经常性抽样调查　　b.必要的统计报表

c.重点调查及估计推算等　　d.周期性普查

3.要对某企业生产设备的实际生产能力进行调查，则该企业的"生产设备"是(　　)。

a. 调查对象　　b. 调查单位

c. 调查项目　　d. 报告单位

4.下面哪些现象适宜采用非全面调查？(　　)

a.企业经营管理中出现的新问题　　b.某型号日光灯耐用时数检查

c.一个地区新增加的人口数　　d.某地区森林的木材积蓄量

二、问答题

1.有人说抽样调查"以样本资料推断总体数量特征"缺乏科学依据，你认为呢？

2.对足球赛观众按男、女、老、少分为四组以分析观众的结构，这种分组方法合适吗？

3.以一实例说明统计分组应遵循的原则。

4.抽样调查某地区50户居民的月消费品支出额(单位：元)数据资料如下：

886	928	999	946	950	864	1 050	927	949	852
1 027	928	978	816	1 000	918	1 040	854	1 100	900
866	905	954	890	1 006	926	900	999	886	1 120
893	900	800	938	864	919	863	981	916	818
946	926	895	967	921	978	821	924	651	850

要求：

(1)根据上述资料编制次(频)数分布和频率分布数列。

(2)编制向上和向下累计频数、频率数列。

第三章

统计描述

★ 本章学习要点与要求 ★

本章主要阐述如何利用统计表或统计图的形式描述统计整理的结果，直观地显示数据分布的特征。通过学习，要求：①熟悉统计表和统计图的结构与绘制方法；②掌握算术平均数、调和平均数、几何平均数、众数、中位数等反映总体的一般水平和分布的集中趋势的指标的含义与计算方法；③掌握极差、方差、标准差、变异系数等反映分布离中趋势的指标的含义与计算方法；④掌握偏度和峰度等指标的含义与计算方法；⑤能利用 Excel 编制各种分布数列，绘制直方图、折线图等统计图，计算反映分布的集中、离散趋势的特征值和偏度、峰度等指标。

第一节　统计表与统计图

一、统计表

（一）统计表的定义和结构

对统计调查所获得的原始资料进行整理，得到说明社会现象及其发展过程的数据，把这些数据按一定的顺序排列在表格上，就形成了统计表。广义的统计表包括统计工作各个阶段中所用的一切表格。狭义的统计表专指分析表和容纳各种统计资料的表格，也就是通常所说的统计表，它清楚地、有条理地显示统计资料，直观地反映统计分布

特征，是统计分析的一种重要工具。

统计表的结构，可以从表式和内容两个方面来认识。

1.从表式上看，统计表是由纵横交叉的线条组成的一种表格，表格包括总标题、横行标题、纵栏标题和指标数值四个部分。总标题是统计表的名称，它扼要地说明该表的基本内容，并指明时间和范围。它置于统计表格的正上方。横行标题是横行的名称，一般放在表格的左方。纵栏标题是纵栏的名称，一般放在表格的上方。横行标题和纵栏标题共同说明填入表格中的统计数字所指的内容。指标数值列在横行和纵栏的交叉处，用来说明总体及其组成部分的数量特征，它是统计表格的核心部分。具体表式如表3-1所示。

2.从内容上看，统计表由主词栏和宾词栏两个部分组成。主词栏是统计表所要说明的总体及其组成部分；宾词栏是统计表用来说明总体数量特征的各个统计指标，如表3-1所示。主词一般列在表的左方，宾词一般列在表的右方。必要时，主宾词可以变换位置或合并排列。此外，统计表还有补充资料、注解、资料来源、填表单位、填表人等。

表3-1 我国各项税收收入情况表 ←总标题

单位：亿元

税收按税种分组	1999年总额	2000年总额	速度(%)(以上年为100)
工商税	8 885.44	10 366.09	116.7
关 税	562.23	750.48	133.5
农业各税	423.50	465 31	109.9
国有企业所得税	639.00	827.41	129.5
集体企业所得税	172.41	172.22	99.9
合 计	10 682.58	12 581.51	117.8

纵栏标题

横行标题

指标数值

主词栏

宾词栏

资料来源：国家统计局.中国统计年鉴2001年.北京：中国统计出版社，2002

（二）统计表的分类

1.按主词的结构分类，根据主词是否分组和分组的程度，分为简单表、分组表和复合表。

(1)简单表：主词未经任何分组的统计表称为简单表，也称一览表。主词罗列各单位的名称。

(2)分组表：主词只按一个标志进行分组形成的统计表，也称简单分组表，如表3-1就是税收按税种进行简单分组所形成的统计表。

(3)复合表:主词按两个或两个以上标志进行分组的统计表,也称复合分组表,如表3-2,表中的国内生产总值在按产业分组的基础上再按国民经济行业分组。在复合分组表中设计横行标题时,应在第一次分组的各组组别下退一或二字填第二次分组的组别,此时第一次分组的组别就成为第二次分组的各组小计。若需再进行第三、四次分组,均可按此类推。如表3-2,对每天创造的国内生产总值先按产业进行第一次分组,分为第一、第二和第三产业。对第二产业又进行第二次分组,按行业分为工业和建筑业。

表3-2 2000年平均每天创造国内生产总值

国内生产总值按产业和行业分组	国内生产总值(亿元)	比重(%)
第一产业	38.9	15.9
第二产业	124.6	50.9
工业	108.4	87.0
建筑业	16.2	13.0
第三产业	81.4	33.2
交运仓储邮电通信业	13.5	16.6
批发和零售贸易餐饮业	20.0	24.6
其他	47.9	58.8
合　计	244.9	100.0

资料来源:国家统计局.中国统计年鉴2001年.北京:中国统计出版社,2002

2.按宾词设计分类,可分为宾词简单排列、分组平行排列和分组层叠排列等三种。

(1)宾词简单排列。宾词不进行任何分组,按一定顺序排列在统计表上。

(2)宾词分组平行排列。宾词栏中各分组标志彼此分开,平行排列,如表3-3就是宾词分组平行排列表。

表3-3 各地区社会商品零售总额　　单位:亿元

按地区分组	按商品性质和用途分组		按城乡分组		按经济类型分组			
	消费品零售总额	农业生产资料销售额	城镇	乡村	国有	集体	个体	其他
北京								
天津								
河北								
⋮								
合计								

(3)宾词分组层叠排列。统计指标同时有层次地按两个或两个以上标志分组,各种

分组层叠在一起，宾词的栏数等于各种分组的组数连乘积。例如表3－4中，各地区从业人员先按三次产业分为3组，再按性别分为2组，另加小计栏，则复合分组设计的宾词栏数共有3×3＝9栏(不包括总计栏)。

表3－4 各地区从业人员分布表(2001年年末数) 单位：万人

地区	合计	三次产业								
		第一产业			第二产业			第三产业		
		小计	男	女	小计	男	女	小计	男	女
北京										
天津										
河北										
⋮										
总计										

统计表的主词分组与宾词分组是有区别的：主词分组的结果使总体分成许多组成部分，它们需要用统计指标(宾词)来描述。宾词分组的结果并不增加统计总体的各组成部分，仅仅是比较详细地描述总体已有的各个组成部分。由此可见，主词分组具有独立的意义，而宾词分组从属于主词的要求，是为更详细地描述主词的数量特征而设计的。

(三)统计表的设计

统计表设计总的要求是：简练、明确、实用、美观，便于比较。统计表的设计应注意如下事项：

(1)线条的绘制。表的上下端应以粗线绘制，表内纵横线以细线绘制。表格的左右两端一般不划线，采用“开口式”。

(2)合计栏的设置。统计表各纵列若需合计时，一般应将合计列在最后一行，各横行若需要合计时，可将合计列在最前一栏或最后一栏。

(3)标题设计。统计表的总标题，横栏、纵栏标题应简明扼要，以简练而又准确的文字表述统计资料的内容、资料所属的空间和时间范围。

(4)指标数值。表中数字应该填写整齐，对准位数。当数字小且可略而不计时，可写上“0”；当缺某项数字资料时，可用符号“…”表示；不应有数字时用符号“—”表示。

(5)计量单位。统计表必须注明数字资料的计量单位。当全表只有一种计量单位时，可以把它写在表头的右上方。如果表中各格的指标数值计量单位不同，可在横行标题后添一列计量单位。

(6)注解或资料来源。必要时，在统计表下应加注解或说明，以便查考。

二、统计图

运用统计图形可以直观地描述次数分布的类型特征，常用的图形有三种。

（一）直方图

直方图是用直方形的宽度和高度来表示次数分布的图形。绘制直方图时，横轴表示各组组限，纵轴表示次数（一般标在左方）和比率（或频率，一般标在右方），若没有比率，直方图只保留左侧的次数。依据各组组距的宽度与次数的高度绘成直方形。

（二）折线图

折线图是以线段的起伏表示其数量分布的特征。绘图时，可以在直方图的基础上，用折线将各组次数高度的坐标连接而成，也可以用组中值与次数求坐标点连接而成。

（三）曲线图

当变量数列的组数无限多时，折线便表现为一条平滑曲线。曲线图的绘制方法与折线图基本相同，只是在连接各组次数坐标点时应当用平滑曲线。

【例 3－1】 某企业加工车间 40 名工人日产量（件）资料如表 3－5 所示，用图示法描述工人日产量的分布状况。

表 3－5　40 名工人日产量次数分布表

按日产量分组（件）	次数 f_i	频率 $\frac{f_i}{\sum f_i}$
50～60	2	5.0
60～70	7	17.5
70～80	11	27.5
80～90	12	30.0
90～100	8	20.0
合 计	40	100.0

解：

根据表 3－5 的资料绘制的直方图如图 3－1 所示。对于不等组距式变量数列，则通常按次数密度（频数密度）绘制直方图以表示其分布。

根据表 3－5 的资料绘制的折线图，如图 3－2 所示。

根据表 3－5 的资料直接绘制或对折线图（3－2）的折线平滑，所得的曲线图，如图 3－3 所示。

（四）累计曲线图

1. 累计曲线图的绘制。累计频数（频率）分布图分为向上累计频数（频率）分布图和向下累计频数（频率）分布图。不论是向上累计或向下累计，均以分组变量为横轴，以累

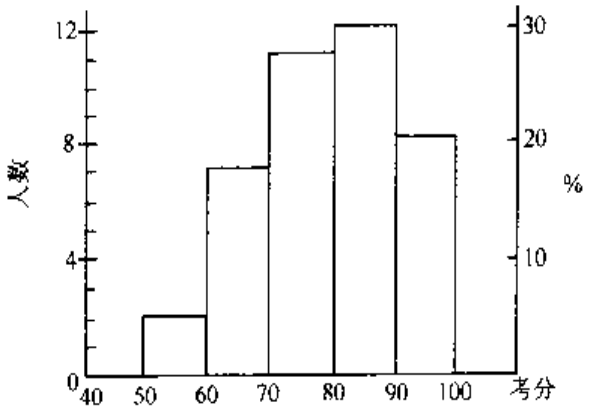


图 3-1 工人日产量次数分布直方图

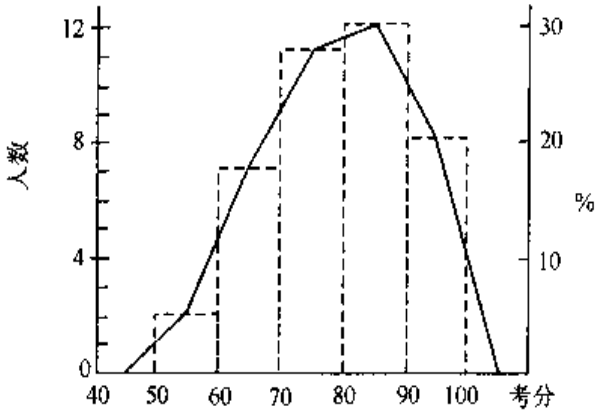


图 3-2 工人日产量次数分布折线图

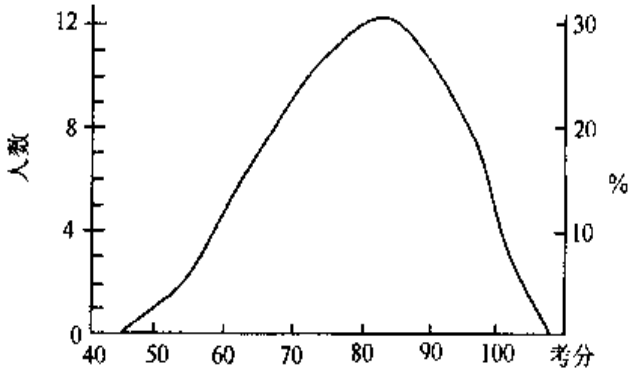


图 3-3 工人日产量次数分布曲线图

计频数(频率)为纵轴。在直角坐标系上将各组组距的上限与其相应的累计频数(频率)构成坐标点,依次用折线(或光滑曲线)相连,即是向上累计曲线。对于向下累计频数分布图,在直角坐标系上将各组组距下限与其相应累计频数(频率)构成坐标点,依次用折

线(或光滑曲线)相连,即是向下累计分布曲线图。根据第二章的[例2-1]表2-3的资料绘制累计曲线图,如图3-4所示。

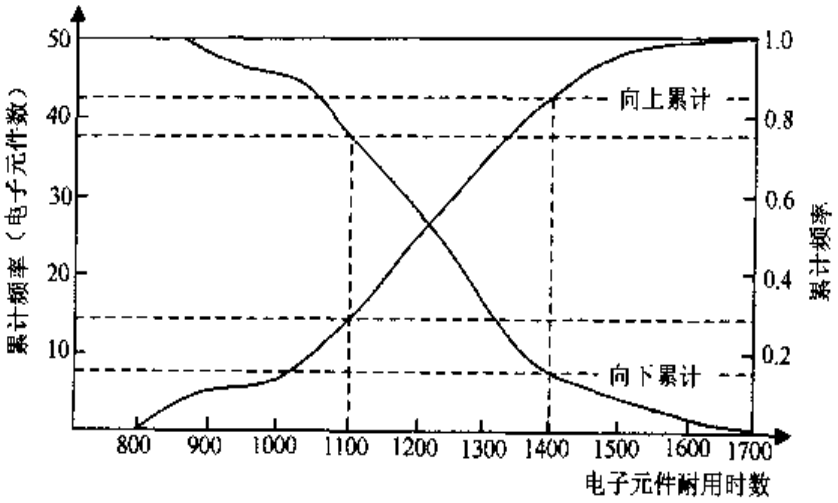


图3-4 累计频数(频率)分布示意图

从图3-4可以看出,电子元件耐用时数在1 100小时以下的有14个,占28%,在1 100小时以上的有36个,占72%;耐用时数在1 400小时以下的有43个,占86%,以上的有7个,占14%等。由此可见,累计频数和累计频率可以简要地概括反映总体各单位的分布特征。

向上累计分布曲线呈上升状,向下累计分布曲线呈下降状。组的次数(或频率)越少,曲线显得越平缓;组的次数(或频率)越多,曲线显得越陡峭。

2.洛伦茨曲线图。美国洛伦茨博士(Dr. M.O.lorenz)把累计频数(或频率)分布曲线运用于研究社会财富、土地和工资收入的分配是否公平。这种累计曲线又称为洛伦茨曲线图。其绘制方法如下:

(1)将分配对象和接受分配者的数量化成结构相对数并进行向上累计。

(2)纵轴和横轴均为百分比尺度,纵轴自下而上,用以测定分配的对象(如一国的财富、土地或收入等的分配状况),横轴由左向右用以测定接受分配者(如一个地区人口)。

(3)根据计算所得的分配对象和接受分配者的累计百分数,在图中标出相应的绘示点,连接各点并使之平滑,所得曲线即所要求的洛伦茨曲线。

【例3-2】某地区某年家庭收入资料如表3-6所示,据此绘制洛伦茨曲线。

表 3-6 某国家收入所得的分配情况

按收入所得水平分组	人口			收入			累计收入的(%)	
	人口数(万人)	结构(%)	累计的(%)	月收入额(亿美元)	结构(%)	实际情况	绝对平等	绝对不平等
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
最低	128.5	12.85	12.85	1.57	5	5	12.85	0
中下等	348.0	34.80	47.65	4.08	13	18	47.65	0
中等	466.9	46.69	94.34	16.33	52	70	94.34	0
较高	45.6	4.56	98.90	7.54	24	94	98.90	0
最高	11.0	1.10	100.00	1.88	6	100	100.00	100
合计	1 000.0	100.00	—	31.40	100	—	—	—

解：

先将人口数及其收入额[第(1)(4)列]化为结构相对数[第(2)(5)列]，再求出其累计的百分比[第(3)(6)列]，然后在制好的比率图上依累计百分比标出绘示点，平滑地连接各绘示点即可，如图 3-5 所示。

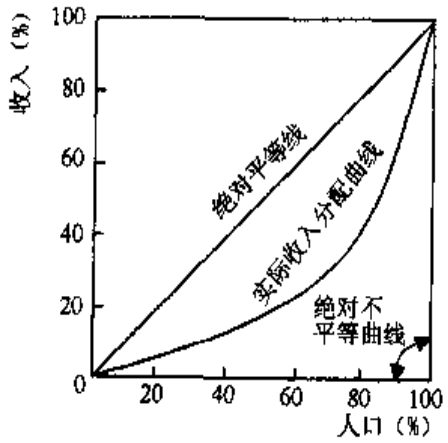


图 3-5 洛伦茨曲线示意图

图 3-5 中的曲线为实际收入分配曲线，对角线为绝对平等线，横轴与右纵轴构成绝对不平等线。用实际收入分配曲线与绝对平等线所包围的面积对比总面积，计算基尼系数，以它衡量收入分配的平等与否。基尼系数值越小，即实际收入分配曲线越靠近绝对平等线，则收入分配越平等；反之，基尼系数值越大，即实际收入分配曲线越靠近绝对不平等线，则收入分配越不平等。

三、频(次)数分布图的类型

利用统计图的形式对社会现象的数量分布特征进行描述，通过这些图形，可以直观

地显示不同类型现象的分布特征。

各种不同性质的社会经济现象的次数分布的类型，概括起来，大致有三种类型：钟型分布、U 型分布和 J 型分布。

（一）钟型分布

钟型分布的特征是“两头小，中间大”，即中间的变量值分布的次数多，靠近两边的变量值分布的次数少，其曲线图宛如一口古钟，如图 3－6 所示。

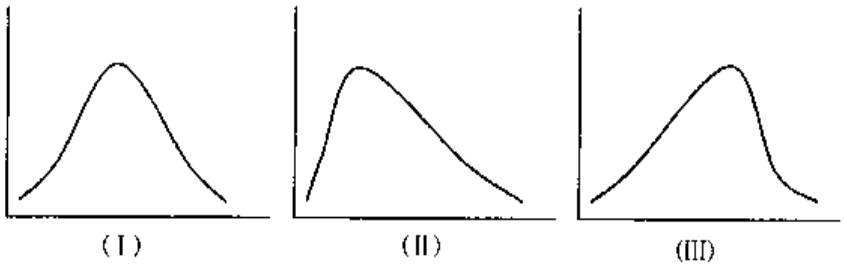


图 3－6　钟型分布示意图

如图 3－6(Ⅰ)所示，其分布特征是以变量的平均数为对称轴，左右两侧对称，两侧变量值分布的次数随着与其平均值距离的增大而渐次减少。在统计学中，称这种分布为正态分布。图 3－6 中的(Ⅱ)(Ⅲ)为非对称分布，它们各有不同方向的偏态。图(Ⅱ)曲线是正偏(右偏)分配，图(Ⅲ)曲线是负偏(左偏)分配。在客观实际中，许多社会现象总体的分布都趋于正态分布。例如，农作物单位面积产量的分布、零件公差的分布、商品市场价格的分布等。

（二）U 型分布

U 型分布的形状与钟型分布相反，靠近中间的变量值分布次数少，靠近两端的变量值分布次数多，形成“两头大，中间小”的 U 型分布。例如，人口按年龄死亡率的分布，人口总体中，幼儿和老年人死亡率高，而中青年死亡率低。图 3－7 是 U 型分布图。

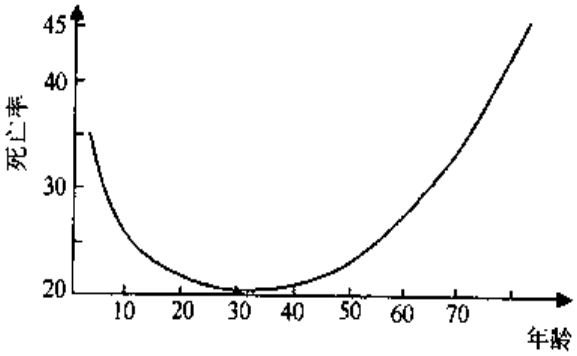


图 3－7　U 型分布示意图

(三)J 型分布

J 型分布有两种类型,一种是次数随着变量的增大而增多,另一种呈反 J 型分布,即次数随着变量的增大而减少,如图 3-8 所示。

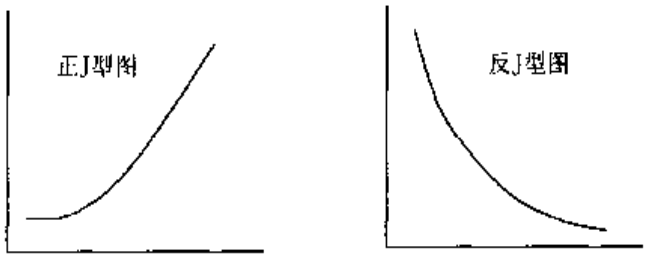


图 3-8 J 型分布示意图

第二节 分布的集中趋势

一、描述分布集中趋势的主要指标及其分类

(一)平均指标的含义

研究统计分布集中趋势的主要指标是各类统计平均数,或称平均指标。所谓统计平均数,就是概括地描述统计分布的一般水平或集中趋势的数值。统计平均数将总体各单位标志值的差异抽象化,它与各单位的标志值不一定相等,但又反映总体各单位标志值的一般水平或表明各单位标志值的集中趋势。

(二)平均指标的作用

在统计分析中,统计平均数具有重要的作用,主要体现在:

1.反映总体各单位变量分布的集中趋势和一般水平。大部分的社会现象,总体各单位某一变量从小到大形成一定的分布,标志值很小或很大的单位比较少,而靠近平均数的单位数比较多,即标志值围绕平均数周围的单位数在总体单位数占有最大的比重,这就显示了总体各单位向平均数集中的趋势,也表明了总体的一般水平。

2.便于比较同类现象在不同单位间的发展水平。在说明不同企业的生产水平、经济效益或工作质量以及对不同投资项目的评估等许多场合都广泛地应用平均指标。

3.能够比较同类现象在不同时期的发展变化趋势或规律。社会经济现象的变化易受偶然因素和现象规模的影响。用平均指标来分析,既可以消除偶然因素的作用,又能

够避免受现象规模的影响，比较确切地反映总体现象变化的基本趋势。

4.分析现象之间的依存关系时也常借助于平均指标。例如，将工业企业按照规模的大小进行分组，并计算各不同规模工业企业的平均劳动生产率、利润率等指标，可以较深入的分析企业规模的大小与劳动生产率或利润率之间的关系。

（三）统计平均数的分类

根据各种平均数的具体代表意义和计算方式的不同，一般地，统计平均数可分为两类：即数值平均数和位置平均数。所谓数值平均数就是以统计数列的所有各项数据来计算的平均数，用以反映统计数列的所有各项数值的平均水平。这类平均数的特点是，统计数列中任何一项数据的变动，都会在一定程度上影响到数值平均数的计算结果。换句话说，数值平均数是由分配数列的所有变量数值来确定的。数值平均数包括了算术平均数、调和平均数、几何平均数和幂平均数。所谓位置平均数是根据标志值的某一特定位置来确定的，它不是对统计数列中所有各项数据进行计算所得的结果，而是根据数列中处于特殊位置上的个别单位或部分单位的标志值来确定的，常用的位置平均数有众数和中位数两种。

二、数值平均数

数值平均数是对总体各单位某一标志值的平均，表明总体单位标志值的平均水平。它的基本计算公式是：

$$\text{平均数}=\frac{\text{总体标志总量}}{\text{总体单位数}} \tag{3.1}$$

使用上述公式应注意：平均数是对同质总体中各单位的标志值进行平均，它要求总体标志总量和总体单位数严格的相对应，即总体标志总量必须是总体各单位标志值的总和，标志值和单位之间存在一一对应关系。例如，全国人均能源消费量指标，是全国能源消费总量与全国人口数的比率，因为每个人都有能源消费这个标志，所以人均能源消费量是个平均指标。又如，人均能源生产量指标，是全国能源总产量与全国人口数之比，但是能源生产量并不是每个人都具有的标志，所以人均能源生产量就不是平均指标，在统计上称之为强度相对指标。

（一）算术平均数

计算算术平均数有两种形式：简单算术平均数和加权算术平均数。

1.算术平均数的计算

(1)简单算术平均数。简单算术平均数适用于未分组的资料，用总体各单位标志值简单加总得到的标志总量除以总体单位总量而得。计算公式如下：

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{3.2}$$

式中，$\bar{x}$ 代表算术平均数；x_i 表各单位标志值；n 代表总体单位数。

(2)加权算术平均数。加权算术平均数适用于原始资料已经经过分组，并得出次数分布的场合。计算各组的标志总量时，必须先将各组标志值乘以相应的次数，求得各组的标志总量。加权算术平均数的计算公式为：

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}x_i\cdot f_i}{\sum_{i=1}^{n}f_i} \tag{3.3}$$

或

$$\bar{x}=\sum_{i=1}^{n}x_i\frac{f_i}{\sum_{i=1}^{n}f_i} \tag{3.4}$$

式中，f_i 为各组标志值出现的频数(次数)，$\frac{f_i}{\sum_{i=1}^{n}f_i}$ 为频率。

【例 3－3】现以第二章[例 2－1]产品生产过程质量控制的资料为例，说明算术平均数的计算。

解：

对未分组的原始资料，采用简单算术平均法计算 50 个电子元件的平均耐用时数。具体计算过程如下：

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{810+880+1\ 230+1\ 100+\cdots+1\ 270+1\ 250}{50}=1\ 195.6$$

对[例 2－1]的资料进行等距分组并计算各组的组中值，用加权平均法计算 50 个电子元件的平均耐用时数。具体计算过程如表 3－7 所示。

$$\bar{x}=\frac{\sum_{i=1}^{n}x_if_i}{\sum_{i=1}^{n}f_i}=\frac{59\ 900}{50}=1\ 198$$

表 3－7　50 个电子元件耐用时数资料表　　单位：小时

按耐用时数分组组中值 (x)	频数 (f_i)	x_if_i	频率 ($f_i\Big/\sum_{i=1}^{n}f_i$)	$x_i\frac{f_i}{\sum_{i=1}^{n}f_i}$
850	5	4 250	0.10	85
950	1	950	0.02	19
1 050	8	8 400	0.16	168
1 150	11	12 650	0.22	253
1 250	11	13 750	0.22	275
1 350	7	9 450	0.14	189

续表

按耐用时数分组组中值 (x)	频数 (f_i)	xf_i	频率 $(f_i/\sum_{i=1}^{n}f_i)$	$x_i\frac{f_i}{\sum_{i=1}^{n}f_i}$
1 450	4	5 800	0.08	116
1 550	3	4 650	0.06	93
合计	50	59 900	1.00	1 198

如果用频率加权计算，也可以得到相同的结果：$\bar{x}=\sum_{i=1}^{n}x_i\cdot(f_i/\sum_{i=1}^{n}f_i)=1\ 198$

从上面的计算结果可以看到，用简单算术平均法计算的结果和加权算术平均法计算的结果是不同的，这是因为在计算加权算术平均数时应用的是组距式分组的资料，以组中值代替各组标志值的平均水平来计算的，其前提条件是假定各标志值在组内均匀分布。由此可见，分组时，应尽量使各组标志值在组内分布均匀，以避免出现误差。当以简单算术平均数公式和以加权算术平均数公式计算的结果不一致时，以简单算术平均数公式计算的结果为准确值。

2. 权数的作用与选择

(1)权数的作用。在进行加权算术平均时，我们利用了权数进行计算。权数有两种表现形式：绝对数形式和相对数形式。用各组的次数做为权数，即为绝对数形式；用各组的频率做为权数，即为相对数形式。权数在平均数中的权衡轻重的作用，是直接通过各组单位数占总体单位数的比重，也就是各组的频率——相对权数的大小体现出来的。从例 3－3 可以明显地看出，频率的大小就直接表明了该组标志值在平均数中占据的地位，如电子元件耐用时数在 1 150 小时组的频率是 0.22，即以 22% 记入总平均数，1 250 小时组也是以 22% 计入总平均数……频率越大，该标志值记入平均数的份额也越大，对平均数的影响就越大；反之，频率就越小，该标志值记入平均数的份额也越小，对平均数的影响就越小。这就是权数权衡轻重作用的实质。

(2)权数不起作用的场合。当各组的次数都相同时，各标志值对平均数的影响都相同时，那就无所谓权数的“权衡轻重”了。在这种情况下(又假定各组标志值在组内均匀分布)，各组权数对平均数的作用都一样，加权算术平均数就等于简单算术平均数。

即当 $f_1=f_2=f_3=\cdots=f_n$ 时：

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i\cdot f_i}{\sum_{i=1}^{n}f_i}=\frac{f\sum_{i=1}^{n}x}{nf}=\frac{\sum_{i=1}^{n}x}{n}$$

可以说，简单算术平均数实际上是加权算术平均数的特例。

(3)权数的选择。在分配数列的条件下，一般说来，次数就是权数，但在计算相对指

标的平均数时，经常遇到次数是不合适的权数的情况。在求相对指标的平均数时，应根据相对指标的含义，选择适当的权数。

【例3-4】某公司所属15个企业资金利润率分组资料如表3-8，要求计算该公司15个企业的平均利润率。

表3-8 某公司所属15个企业资本金利润率资料

资本金利润率 x_i(%)	企业数	资本金总额 f_i(万元)	利润总额 x_if_i(万元)
5	8	40	2
10	6	100	10
15	6	160	24
合计	20	300	36

解：

$$平均利润率=\frac{\sum_{i=1}^{n}x_if_i}{\sum_{i=1}^{n}f_i}=\frac{40\times5\%+100\times10\%+160\times15\%}{40+100+160}=\frac{36}{300}=12\%$$

该例的平均对象是各企业的资本金利润率，表中的企业数虽然是次数，但却不是合适的权数。要正确计算公司15个企业的平均资本金利润率，因为资本金利润率=利润总额/资本金总额，所以计算平均资本金利润率需要以资本金额为权数，才能符合该指标的性质。

3.是非标志的平均数。在第一章中，我们曾提到数据根据其计量尺度的高低可分为四个层次，对于以定类尺度和定序尺度计量的数据资料，统计处理往往是把它们的类别差异过渡到数量上的变异，然后再计算平均数。

例如，从业人员按性别分为男、女两组；高中生的升学考试将考生分为文科和理科两组；企业生产的产品分为合格和不合格两组等等。它们都是以品质标志把总体单位划分成两类，且均可用“是”与“非”来区分，故称之为“是非标志”，或“交替标志”。

在一批产品中，若将合格品看做具有“是”的属性，那么不合格品就具有“非”的属性。设合格率为 p，不合格率为 q，以1作为“是”(合格品)的单位的标志值，以0作为“非”(不合格品)的单位的标志值，这样就把以文字表示的品质标志转化为数量标志。按加权算术平均数公式计算，得到：

$$\bar{x}=\frac{\sum_{i=1}^{n}x_if_i}{\sum_{i=1}^{n}f_i}=\frac{1\times p+0\times q}{p+q}=p \tag{3.5}$$

p 也被称为总体中具有某种属性的单位成数，所以成数是一种特殊的平均数，即是

非标志的平均数。

【例 3－5】对某批产品进行质量抽检，抽取 400 件，其中不合格品 8 件，求合格品成数。

解：

设合格率为 p

$$\bar{x}=p=\frac{400-8}{400}=98\%$$

则该批产品合格品成数为 98%。

4.算术平均数的数学性质。算术平均数有如下四个主要数学性质。

(1)算术平均数与标志值个数的乘积等于各标志值的总和。

简单算术平均数：
$$n\bar{x}=\sum_{i=1}^{n}x_i \tag{3.6}$$

加权算术平均数：
$$\sum_{i=1}^{n}f_i\bar{x}=\sum_{i=1}^{n}x_if_i \tag{3.7}$$

(2)各个标志值与算术平均数离差之和等于零。

简单算术平均数：
$$\sum_{i=1}^{n}(x_i-\bar{x})=0$$

因为 $n\bar{x}=\sum_{i=1}^{n}x_i$，所以 $\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$。

加权算术平均数：

因为 $\sum_{i=1}^{n}f_i\bar{x}=\sum_{i=1}^{n}x_if_i$，所以 $\sum_{i=1}^{n}(x_i-\bar{x})f_i=\sum_{i=1}^{n}x_if_i-\sum_{i=1}^{n}f_i\bar{x}=0$。

(3)各标志值与算术平均数离差的平方和为最小值。

设 x_0 为任意常数，$x_0\neq\bar{x}$ 时，

$$\begin{aligned}\sum_{i=1}^{n}(x_i-x_0)^2&=\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-x_0)^2\\&=\sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-x_0)]^2\\&=\sum_{i=1}^{n}[(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-x_0)+(\bar{x}-x_0)^2]\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-x_0)^2\end{aligned}$$

因为 $n(\bar{x}-x_0)^2>0$，所以 $\sum_{i=1}^{n}(x_i-x_0)^2>\sum_{i=1}^{n}(x_i-\bar{x})^2$

因此，$\sum_{i=1}^{n}(x_i-\bar{x})^2=\min$

(4)对被平均的变量实施某些线性变换后，新变量的算术平均数等于对原变量的算术平均数实施同样的线性变换的结果。

$$\overline{(a+bx)}=a+b\bar{x}$$

(二)调和平均数

调和平均数也称倒数平均数，它与算术平均数一样，有简单调和平均数和加权调和平均数两种。

1.简单调和平均数。简单调和平均数是各个标志值 x_i 的倒数的算术平均数的倒数。计算公式如下：

$$H=\frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}{n}}=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{3.8}$$

式中，H 代表调和平均数。

简单调和平均数的应用场合是，各标志值对应的标志总量为 1 个单位(或相等)。当各标志值对应的标志总量不为 1 个单位(或不相等)时，则要用加权调和平均数。

2.加权调和平均数。加权调和平均数的计算公式：

$$H=\frac{m_1+m_2+\cdots+m_n}{\frac{m_1}{x_1}+\frac{m_2}{x_2}+\cdots+\frac{m_n}{x_n}}=\frac{\sum_{i=1}^{n}m_i}{\sum_{i=1}^{n}\frac{m_i}{x_i}} \tag{3.9}$$

式中，m 表示各单位或各组的标志值对应的标志总量。

【例 3－6】某企业工人的月工资资料如表 3－9 所示，试计算其平均工资。

在本例中，已知各组工人的月工资和工资总额，即被平均的各组的标志值和对应标志总量，但不知各组的工人数，因此要应用调和平均数的计算公式来计算。

表 3－9 某公司职工月工资资料

工资等级	月工资(元) x_i	工资总额(元) m_i	工人数 $f_i=m_i/x_i$
1	600	30 000	50
2	800	40 000	50
3	1 000	50 000	50
4	1 200	48 000	40
5	1 600	16 000	10
合计		184 000	200

解：

$$H=\frac{\sum_{i=1}^{n}m_i}{\sum_{i=1}^{n}\frac{m_i}{x_i}}=\frac{30\,000+40\,000+50\,000+48\,000+16\,000}{\frac{30\,000}{600}+\frac{40\,000}{800}+\frac{50\,000}{1\,000}+\frac{48\,000}{1\,200}+\frac{16\,000}{1\,600}}$$

$$=\frac{184\,000}{200}=920(\text{元})$$

在这里，加权调和平均数实际上是加权算术平均数的变形：

$$\frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{x_i}} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} \frac{x_i f_i}{x_i}} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

当各组标志总量相等，即 $m_1 = m_2 = m_3 = \cdots = m_n$ 时，加权调和平均数可化简成为简单调和平均数形式：

$$\frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{x_i}} = \frac{nm}{m\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x}}$$

所以，简单调和平均数是加权调和平均数的特例。

（三）几何平均数

几何平均数是 n 项标志值连乘积的 n 次方根。几何平均数一般应用于计算时间上相互衔接的比率的平均数。几何平均数也有简单几何平均数和加权几何平均数两种。

1. 简单几何平均数。简单几何平均数就是 n 个标志值 x_i 连乘积的 n 次方根。计算公式为：

$$G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \prod_{i=1}^{n} x_i \tag{3.10}$$

式中，G 表示几何平均数，x_i 表示各项标志值。

【例 3－7】某产品需经三个车间加工，已知第一个车间加工合格率为 95%，第二个车间加工合格率为 90%，第三个车间加工合格率为 98%，求三个车间平均加工合格率。

解：

由于产品是由三个车间连续加工完成的，第二个车间加工的是第一个车间完工的合格品，第三个车间加工的又是第二个车间完工的合格品，因此，三个车间总合格品率是三个车间相应合格品率的连乘，求平均加工合格品率就不能采用算术平均法：(95% + 90% + 98%)/3 = 94.333%，而应当用几何平均法求三个车间的平均加工合格率。

$$G = \sqrt[3]{x_1 \cdot x_2 \cdot x_3}$$
$$= \sqrt[3]{95\% \times 90\% \times 98\%} = 94.275\%$$

2. 加权几何平均数。当计算几何平均数的各个标志值的次数不相同时，应采用加权几何平均数。加权几何平均数是各标志值 f_i 次方的连乘积的 $\sum_{i=1}^{n} f_i$ 次方根，计算公式为：

$$G = \sum_{i=1}^{n} f_i \sqrt{x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}} = \sum_{i=1}^{n} f_1 \sqrt{\prod_{i=1}^{n} x_i^{f_i}} \tag{3.11}$$

（四）幂平均数

设有一组变量 $x_1, x_2, x_3, \cdots, x_n$，求各变量 k 次方的和：

$$x_1^k + x_2^k + x_3^k + \cdots + x_n^k = \sum_{i=1}^{n} x_i^k$$

因为平均数是各变量一般水平的代表值，以幂平均 $\bar{x}_k$ 代替各具体变量 x_i，其数值总和不变，因而有：

$$\sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} (\bar{x}_k)^k = n(\bar{x}_k)^k$$

$$\bar{x}_k = \left(\frac{\sum_{i=1}^{n} x_i^k}{n}\right)^{\frac{1}{k}} \tag{3.12}$$

上式就是幂平均数的计算公式。$\bar{x}_k$ 称为 k 阶幂平均数，它是以变量 x 的 k 次方的算术平均数的 k 次方根来计算的。当 k 取不同的整数值时，幂平均数就给出不同的数值平均数计算公式。

当 $k=1$ 时，幂平均数 $\bar{x}_k = \bar{x}_1 = [(\sum x^1)/n]^{\frac{1}{1}} = (\sum x)/n$，为算术平均数计算公式。

当 $k=-1$ 时，幂平均数 $\bar{x}_k = \bar{x}_{-1} = [(\sum x^{-1})/n]^{\frac{1}{-1}} = [(\sum x^{-1})/n]^{-1}$，为调和平均数计算公式。

当 $k \to 0$ 时，$\lim\limits_{k \to 0} \bar{x}_k = \lim\limits_{k \to 0} [(\sum x^k)/n]^{1/k} \approx \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$，为几何平均数计算公式。

由于幂平均数 $\bar{x}_k = [(\sum x^k)/n]^{1/k}$ 是关于阶数 k 的递增函数，即幂平均数是随着 k 的增大而增大，随着 k 的减少而减少。当 $k_1 < k_2$ 时，就有：

$$\left(\frac{\sum x^{k_1}}{n}\right)^{\frac{1}{k_1}} \leqslant \left(\frac{\sum x^{k_2}}{n}\right)^{\frac{1}{k_2}} \tag{3.13}$$

因为算术平均数($\bar{x}$)、几何平均数(G)、调和平均数(H)都是幂平均数的阶数 k 由 1 递减为 0 又减为 -1 的特例，由此可知它们的一般数量关系为：$H \leqslant G \leqslant \bar{x}$；当各变量相等 $x_1 = x_2 = x_3 = \cdots = x_n$ 时，调和平均数等于几何平均数等于算术平均数。

三、位置平均数

(一)众数

1.众数的含义。众数是指统计总体或分布数列中出现的频数最多、频率最高的标志值，用符号 M_0 表示。在分配曲线图上，众数就是曲线的最高峰所对应的标志值，如图3－9所示。

众数具有非常直观的代表性。可利用众数来表明社会现象的一般水平，也可以将它作为某些生产决策的参考依据。例如，为了掌握农贸市场上某一农产品的价格水平，

不一定要全面登记该农产品每一次成交的价格，只要调查其最普遍的成交价格即可。再如制鞋厂在制定各种尺码鞋子的生产计划时，计划产量最多的尺码就应是在市场上销售量最大(众数)的尺码。

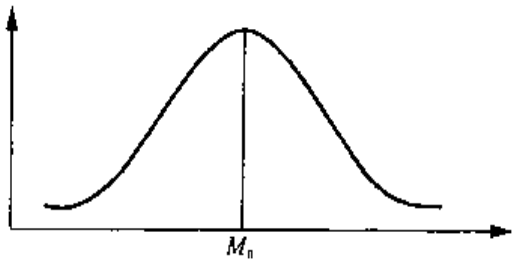


图3-9

2.确定众数的方法，确定众数，必须先对资料进行整理，编制分配数列。由于分组有单项式分组和组距式分组，而组距式分组又有等距分组和不等距分组之分，因而对于各种不同的资料条件，确定众数的方法又有所不同。

(1)由单项式分配数列确定众数。由单项式分配数列确定众数，方法比较简单。即出现次数最多的标志值就是众数。

【例3-8】某企业职工工资资料如表3-10所示。

表3-10 某企业职工工资资料 单位：元

工资 x	500	600	900	1 100	1 200	合计
工人数 f	20	30	80	15	5	150

解：

表3-10中，职工工资水平为900元的次数最多，在150个工人中，有80个工人集中在这一组，所以900就是众数。

(2)由组距式分配数列确定众数。由组距数列确定众数，应首先确定众数组，然后再通过一定的公式计算众数的近似值。在等距分组条件下，众数组就是次数最多的那一组；在不等距分组的条件下，众数组则是频数密度或频率密度最高的那一组。众数值是依据众数组的次数与众数组相邻的两组次数的关系来近似计算的，如图3-10所示。

根据图3-10可以得到两个公式：

下限公式：

$$M_0 = L_{M_0} + \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})} \times d_{M_0} \tag{3.14}$$

上限公式：

$$M_0 = U_{M_0} - \frac{f_{M_0} - f_{M_0+1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})} \times d_{M_0} \tag{3.15}$$

公式中，M_0 代表众数；L_{M_0}代表众数组的下限；U_{M_0}代表众数组的上限；f_{M_0}代表众数组的次数；f_{M_0-1}代表众数组前一组的次数；f_{M_0+1}代表众数组后一组的次数；d_{M_0}代表众数组的组距。

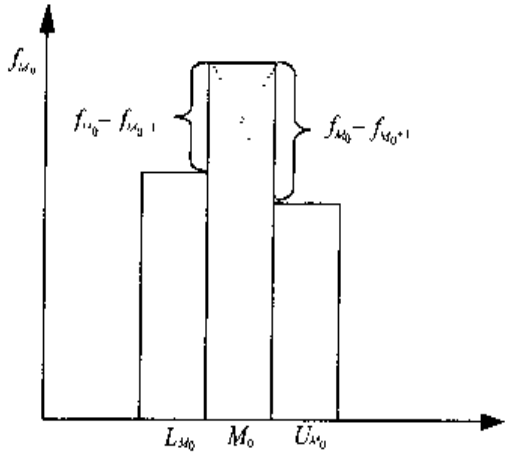


图 3－10

【例 3－9】某城市 A 社区 300 户居民月人均可支配收入的资料如表 3－11 所示，要求：确定居民月人均可支配收入的众数。

表 3－11 居民月人均可支配收入资料表

月人均可支配收入(元)	居民户数
500 以下	16
500～700	32
700～900	60
900～1 100	96
1 100～1 300	42
1 300－1 500	34
1 500 以上	20
合 计	300

解：

$$M_0 = L_{M_0} + \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})} \times d_{M_0}$$

$$= 900 + \frac{96 - 60}{(96 - 60) + (96 - 42)} \times 200 = 980(\text{元})$$

或

$$M_0 = U_{M_0} - \frac{f_{M_0} - f_{M_0-1}}{(f_{M_0} - f_{M_0-1}) + (f_{M_0} - f_{M_0+1})} \times d_{M_0}$$

$$= 1\,100 - \frac{96 - 42}{(96 - 60) + (96 - 42)} \times 200 = 980(\text{元})$$

值得一提的是，如果一个分配数列没有明显的集中趋势，就不存在众数。

（二）中位数

1. 中位数的含义。将总体各个单位按其标志值的大小顺序排列，处于数列位次中点的单位的标志值即中位数，用符号 M_e 表示。从中位数的概念可以看出，在数列中，标志值小于中位数的单位占一半，标志值大于中位数的单位也占一半。如果数列出现极端值，常用中位数来代表总体标志值的一般水平，以避免数列的平均水平受极端值的影响。例如，人口的平均年龄会受到一些特别长寿人年龄的影响，使计算结果偏大，而年龄中位数往往能够更好地体现人口年龄的平均水平。

2. 确定中位数的方法。所掌握的资料不同，采用的方法也不同。

(1)未分组资料确定中位数。在资料未分组时，确定中位数的方法是：

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & (\text{当 } n \text{ 为奇数}) \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & (\text{当 } n \text{ 为偶数}) \end{cases} \tag{3.16}$$

(2)单项式分组资料确定中位数。单项式分组已经将资料的标志值序列化，这时总体单位数 $n = \sum f$，确定中位数位置的方法要通过累计次数计算。具体确定中位数的方法是：

$$M_e = \begin{cases} x_{\frac{\sum f+1}{2}} & (\text{当} \sum f \text{ 为奇数}) \\ \dfrac{x_{\frac{\sum f}{2}} + x_{\frac{\sum f}{2}+1}}{2} & (\text{当} \sum f \text{ 为偶数}) \end{cases} \tag{3.17}$$

【例 3－10】某村居民户按子女数分组资料如表 3－12 所示，求家庭子女数的中位数。

表 3－12　某村居民户按子女数分组资料

子女数 x_i	0	1	2	3	4	合计
家庭户 f_i	50	122	155	306	18	651
累计家庭户	50	172	327	633	651	

解：

中位数位置 $= \dfrac{\sum f + 1}{2} = \dfrac{651 + 1}{2} = 326$，所以中位数为 2 个子女。

(3)组距式分组资料确定中位数。先根据组距式数列的累计资料确定中位数所在

的组，再按公式3.18，3.19计算中位数的近似值。计算公式如下：

下限公式为：

$$M_e = L_{M_e} + \frac{\frac{\sum f}{2} - S_{M_e-1}}{f_{M_e}} \times d_{M_e} \tag{3.18}$$

上限公式为

$$M_e = U_{M_e} - \frac{\frac{\sum f}{2} - S_{M_e+1}}{f_{M_e}} \times d_{M_e} \tag{3.19}$$

式中，M_e 表示中位数；L_{M_e} 表示中位数所在组的下限；U_{M_e} 表示中位数所在组的上限；S_{M_e-1} 表示向上累计至中位数所在组前一组的次数；S_{M_e+1} 表示向下累计至中位数所在组后一组的次数；f_{M_e} 表示中位数所在组的次数；d_{M_e} 表示中位数所在组的组距。

【例3－11】根据表3－13资料计算农户年均收入中位数。

表3－13 农户年均收入中位数计算表

农户年均纯收入(元)	户数(f)	向上累计	向下累计
4 000以下	18	18	100
4 000～5 000	22	40	82
5 000～7 000	40	80	60
7 000～10 000	16	96	20
10 000以上	4	100	4
合　计	100		

解：

确定中位数所在组。因为$(\sum f)/2 = 100/2 = 50$，所以中位数在5 000～7 000这一组。根据公式计算中位数近似值：

$$M_e = L_{M_e} + \frac{\frac{\sum f}{2} - S_{M_e-1}}{f_{M_e}} \times d_{M_e}$$

$$= 5\,000 + \frac{50-40}{40} \times 2\,000 = 5\,500(\text{元})$$

或

$$M_e = U_{M_e} - \frac{\frac{\sum f}{2} - S_{M_e+1}}{f_{M_e}} \times d_{M_e} = 7\,000 - \frac{50-20}{40} \times 2\,000 = 5\,500(\text{元})$$

(三)众数、中位数和算术平均数的关系

众数、中位数和算术平均数都是用于反映总体的一般水平或分布的集中趋势的代

表值，但因为它们的计算方法不同，具体的含义有异，故它们也有各自的特点。第一，众数和中位数是由所处的特殊位置确定的，而算术平均数是由数列所有变量值计算的，所以算术平均数对数据的概括能力比众数、中位数强。第二，算术平均数易受数列中极端值的影响，中位数次之，众数几乎不受极端值的影响。第三，它们对数据的量化尺度要求不同，算术平均数要求最高，它只适用于定距尺度和定比尺度的数据；中位数次之，它还适用于定序尺度的数据；众数对数据的计量尺度没有严格的限制，除上述的三种计量尺度外，众数甚至还适用于定类尺度的数据。

众数、中位数和算术平均数彼此间存在着一定的数量关系：在对称的正态分布条件下，$\bar{x}=M_0=M_e$；在非对称分布的情况下，众数、中位数和算术平均数三者的差别取决于分布偏斜的程度，分布偏斜的程度越大，它们之间的差别越大。当次数分配呈右偏（正偏）时，算术平均数受极大值的影响，就有 $\bar{x}>M_e>M_0$。当次数分配呈左偏（负偏）时，算术平均数受极小值的影响，有 $M_0>M_e>\bar{x}$，中位数则总是介于众数和平均数之间，如图 3－11 所示。

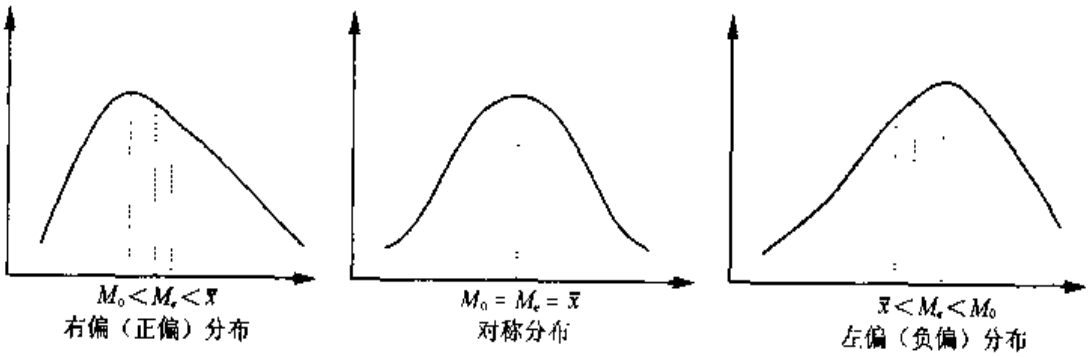


图 3－11

英国统计学家皮尔生的研究提出，在存在轻微偏斜的情况下，众数、中位数和算术平均数数量关系的经验公式为：算术平均数和众数的距离约等于算术平均数与中位数距离的三倍。即：

$$\bar{x}-M_0\approx 3(\bar{x}-M_e) \tag{3.20}$$

利用这个关系式，可以从已知的两个平均指标来推算另一个平均指标。

【例 3－12】某车间生产的一批零件中，直径大于 402 厘米的占一半，众数为 400 厘米，试估计其平均数，并判定其偏斜方向。

解：

已知 $M_e=402, M_0=400$

$$\bar{x}-400\approx 3(\bar{x}-402)$$

$$\bar{x}=403$$

因为 $M_0 < M_e < \bar{x}$，所以，该零件的直径分布为右偏。

第三节 分布的离中趋势

一、描述分布离中趋势的主要变异指标及其作用

(一)变异指标的含义

统计平均指标是把同质总体中各个单位的数量差异抽象化，以数值的形式综合地反映总体各单位的集中趋势。变异指标则是从另外一个侧面，概括地反映总体中各单位的离中趋势或变异状况。或者说，变异指标是用来刻画总体分布的离散程度或变异状况，变异指标值越大，表明总体各单位标志值的变异程度越大。

由于分布的离散程度可以从不同角度、用不同方法去考察，故描述分布离中趋势的变异指标也有多种。常见的变异指标有：极差、分位差、标准差、方差和变异系数。

(二)变异指标的作用

在统计分析中，对平均指标和变异指标是互相补充，相互结合地加以运用的。变异指标的作用主要有以下几种。

1.用于衡量平均指标的代表性。平均指标作为总体数量标志的代表值，其代表性取决于总体各单位标志值的差异程度。总体各单位标志值的变异程度越大，平均指标的代表性就越小；总体各单位标志值的变异程度越小，平均指标的代表性就越大。

【例 3－13】某公司所属的三个企业三年来的年销售额资料如表 3－14 所示。

表 3－14 某公司三个企业销售额资料 单位：万元

年份	企业 A	企业 B	企业 C
	年销售额	年销售额	年销售额
1	1 100	200	1 600
2	1 300	3 000	1 200
3	1 200	400	800

该公司 A,B,C 三个企业，年平均销售额都是 1 200 万元。但对这三个企业各年销售额的变动情况进行分析，显然，年平均销售额 1 200 万元对各个企业的代表性是不同的，对 A 企业而言，其代表性最大，C 企业次之，对于 B 企业，其代表性最小。

2.反映社会经济活动的均衡性。变异指标可反映生产过程的节奏性和经济活动的

均衡性,因此可作为企业产品质量控制和评价经济管理工作的依据。

【例 3－14】某公司下属两个企业销售额计划完成情况如表 3－15 所示。

表 3－15　某公司下属两个企业销售额计划完成情况　　单位:万元

企业	计划数	实际数	第一季度		第二季度		第三季度		第四季度	
			绝对数	比重%	绝对数	比重%	绝对数	比重%	绝对数	比重%
甲	1 000	1 000	140	14.0	100	10.0	460	46.0	300	30.0
乙	1 200	1 200	300	25.0	280	23.3	310	25.8	310	25.8

表 3－15 资料可以看出两个企业的销售计划虽然都已完成,但两者计划执行过程的情况则不相同:乙企业各季度的销售情况较为均衡,而甲企业则存在前松后紧的情况,各季度销售变动幅度较大。如果不存在季节变动等其他因素的影响,那么可以说乙企业的销售情况比甲企业好。

3.研究总体标志值分布偏离正态的情况。一般地说,标志值分布越集中,从分布图看,其分布曲线越陡峭;标志值分布越分散,其分布曲线越平坦。

4.用于衡量统计推断效果。有关这部分内容,本书将在第四、五章中加以介绍。

二、极差与四分位差

(一)极差

极差也称全距,用 R 表示,它是最简单的变异指标,以变量数列中的两个极端的标志值之差表示,反映数列中标志值变动的范围。其计算公式是:

$$R = x_{max} - x_{min} \tag{3.21}$$

【例 3－15】根据例 3－13 的数据计算极差。

解:

$$R_A = x_{max} - x_{min} = 1\ 300 - 1\ 100 = 200$$

$$R_B = x_{max} - x_{min} = 3\ 000 - 200 = 2\ 800$$

$$R_C = x_{max} - x_{min} = 1\ 600 - 800 = 800$$

计算结果表明,A 企业销售额的变异程度最小,B 企业的变异程度最大。

如果资料经过统计整理,并形成组距分配数列,则全距的近似值为最高组的上限与最低组的下限之差。用公式表示即:

$$R = U_{max} - L_{min} \tag{3.22}$$

式中,U_{max}代表最高组的上限;L_{min}代表最低组的下限。

极差的优点是计算简便,直观,容易理解。不足之处是它只以两个极端的标志值计算,而不考虑总体内部的分配状况,不能充分利用数列的全部信息,因此,它无法反映标志值变动的一般程度。

（二）四分位差

四分位差是从变量数列中剔除最大和最小各 1/4 的单位，用 3/4 位次与 1/4 位次的标志值之差除以 2 来表示。计算公式如下：

$$Q \cdot D = \frac{Q_3 - Q_1}{2} \tag{3.23}$$

公式中，Q_1 表示分布数列中最小的 1/4 的单位，Q_3 表示分布数列中最大的 1/4 的单位。四分位差的实质是对变量数列剔除了最小和最大的 1/4 后求全距。

四分位差的特点是避免了极差受数列中极端值影响的弱点，可以说，是对极差指标的一种改进。同样地，它和极差的计算方法一样，也是只由两个标志值确定的，不能充分利用数列的全部信息，因此，也无法反映标志值变动的一般程度。

三、方差与标准差

方差（σ^2）与标准差（σ）是测定标志变异程度最常用的指标。标准差是方差的平方根，也称均方差。方差与标准差是利用了 $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min$ 的数学性质，因此它们是测定标志变异程度最灵敏的指标。

（一）数量标志的方差与标准差

1. 数量标志方差与标准差的计算。其计算公式为：

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{3.24}$$

或

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}} \tag{3.25}$$

上面公式中 σ^2 为方差，σ 为标准差。标准差的计量单位同平均数。

【例 3－16】管理学院某班学生统计学考试成绩分组资料如表 3－16 所示，试计算方差和标准差。

表 3－16 学生成绩的方差和标准差计算表

按成绩分组	学生数 f_i	组中值 x_i	xf	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
60 以下	7	55	385	400	2 800
60～70	21	65	1 365	100	2 100
70～80	25	75	1 875	0	0

续表

按成绩分组	学生数 f_i	组中值 x_i	xf	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 f_i$
80～90	19	85	1 615	100	1 900
90～100	8	95	760	400	3 200
合 计	80	—	6 000	1 000	10 000

解：

$$\bar{x}=\frac{\sum xf}{\sum f}=\frac{6\ 000}{80}=75$$

$$\sigma^2=\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2 f_i}{\sum f_i}=\frac{10\ 000}{80}=125$$

$$\sigma=\sqrt{\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2 f_i}{\sum f_i}}=\sqrt{\frac{10\ 000}{80}}=\sqrt{125}=11.18$$

2.总方差、组间方差和组内方差。在资料分组的条件下，总体各标志值对平均数的方差可以分解为组内方差和组间方差。组内方差反映组内标志值对组平均数的方差，组间方差反映组平均数对总平均数的方差。它们之间的关系是总方差等于组间方差与组内方差平均数之和。即：

$$\sigma^2=\delta^2+\overline{\sigma_i^2} \tag{3.26}$$

式中，σ^2 代表总体方差；$\overline{\sigma_i^2}$代表组内方差的平均数；δ^2 代表组间方差。

3.方差与标准差的数学性质

(1)变量的方差等于变量平方的平均数减去变量平均数的平方。即：

$$\sigma^2=\overline{x^2}-(\bar{x})^2 \tag{3.27}$$

由于

$$\begin{aligned}\sigma^2&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\\&=\frac{\sum_{i=1}^{n}[x_i^2-2x_i\bar{x}+(\bar{x})^2]}{n}\\&=\frac{\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n(\bar{x})^2}{n}\\&=\overline{x^2}-2(\bar{x})^2+(\bar{x})^2\\&=\overline{x^2}-(\bar{x})^2\end{aligned}$$

根据这个关系，可以进行方差或标准差的简化计算。

(2)变量对算术平均数的方差小于对任意常数的方差。这是因为$\sum_{i=1}^{n}(x_i-\bar{x})^2=\min\{\sum_{i=1}^{n}(x_i-x_0)^2\}$,所以,当$\bar{x}\neq x_0$($x_0$为任意常数)时,$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\leqslant\frac{\sum_{i=1}^{n}(x_i-x_0)^2}{n}$。

(3)n个同性质独立变量和的方差等于各个变量方差的和。

设 $$x=x_1+x_2+\cdots+x_n,$$

则 $$\sigma_x^2=\sigma_{x_1}^2+\sigma_{x_2}^2+\sigma_{x_3}^2+\cdots+\sigma_{x_n}^2 \tag{3.28}$$

(4)n个同性质独立变量平均数的方差等于各变量方差平均数的$\frac{1}{n}$。

设 $$\bar{x}=\frac{x_1+x_2+x_3+\cdots+x_n}{n}$$

则 $$\sigma_{\bar{x}}^2=\frac{\sigma_1^2+\sigma_2^2+\sigma_3^2+\cdots+\sigma_n^2}{n^2}=\frac{1}{n}\overline{\sigma_i^2} \tag{3.29}$$

(5)变量线性变换的方差等于变量的方差乘以变量系数的平方。

设 $$y=a+bx$$

则 $$\sigma_y^2=b^2\sigma_x^2 \tag{3.30}$$

(二)是非标志的方差与标准差

是非标志的平均数为p,我们可以通过表3-17,按照定义求它的方差与标准差。

表3-17 是非标志的方差与标准差计算表

品质标志	标志值(x_i)	次数f_i	$(x_i-\bar{x})^2f_i$
是	1	N_1	$(1-p)^2N_1$
非	0	N_0	$(0-p)^2N_0$
合计	—	N	$q^2N_1+P^2N_0$

$$\begin{aligned}\sigma^2&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})f_i}{\sum_{i=1}^{n}f_i}\\&=\frac{(1-p)^2N_1+(0-p)^2N_0}{N}\\&=q^2p+p^2q\\&=pq(p+q)\\&=pq=p(1-p)\end{aligned} \tag{3.31}$$

$$\sigma=\sqrt{p(1-p)} \tag{3.32}$$

【例3-17】已知某产品的合格率为98%,求其合格率的方差和标准差。

解:

$$\sigma^2=p(1-p)=98\%\times2\%=1.96\%$$

$$\sigma = \sqrt{p(1-p)} = \sqrt{1.96\%} = 14\%$$

确定是非标志的平均数与方差，就是计算 0－1 分布的数值特征。当 $p = q = 0.5$ 时，0－1 分布的方差达到最大可能值，即 0.25。这是 0－1 分布特有的一个重要性质。

四、变异系数

极差、四分位差、标准差等各种变异指标，都与它们相应的平均指标有着相同的计量单位。这些变异指标的大小不仅取决于总体的变异程度，而且还与标志值绝对水平的高低、计量单位的不同有关。所以，不宜直接用上述变异指标对不同水平、不同计量单位的现象进行比较，应当先做无量纲化处理，而后再比较。对上述变异指标进行无量纲化处理，就是用它们相应的平均数去除。

变异系数也称离散系数，是各变异指标与其算术平均数的比值。例如，将极差与其平均数对比，得到极差系数；将标准差与其平均数对比，得到标准差系数。最常用的变异系数是标准差系数：

$$V = \frac{\sigma}{\bar{x}} \times 100\% \tag{3.33}$$

【例 3－18】某学校男子体操队 5 名队员的体重分别为 55，54，52，52，51 公斤；女子体操队 6 名队员的体重分别为 46，45，44，44，43，42 公斤。试比较哪个队的队员体重更均匀。

解：

$$\sigma_{男} = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{2\ 790 - 2\ 788} = 1.47(公斤)$$

$$\sigma_{女} = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{1\ 938 - 1\ 936} = 1.29(公斤)$$

$$V_{男} = \frac{\sigma}{\bar{x}} \times 100\% = \frac{1.47}{52.8} = 2.78\%$$

$$V_{女} = \frac{\sigma}{\bar{x}} \times 100\% = \frac{1.29}{44} = 2.93\%$$

从标准差看，男队体重的标准差比女队的大，但男队的体重水平比女队高，所以不能直接根据标准差来判断哪个队队员的体重更均匀，必须以标准差系数来判断。根据标准差系数的计算结果表明，男队的标准差系数比女队的标准差系数更小，正确的结论应当是男队队员的体重比较均匀。

第四节 分布的偏度和峰度

一、统计动差

描述统计总体分布的变异状况，除了用第二节所介绍的各种变异指标外，在许多场合还利用统计 k 阶动差(动差也称矩，是物理学的概念在统计分析中的运用)来更进一步地刻划分布的形态特征。

k 阶动差的一般形式为$\sum_{i=1}^{n}x^k f_i \Big/ \sum_{i=1}^{n}f_i$，称之为原点矩，通常用 μ_k 表示。

显然，当 k 等于 1 时，即 1 阶的原点动差就是算术平均数；当 k 等于 2 时，2 阶的原点动差就是平方平均数。

如果把原点移到算术平均数的位置，就可以得到一个以频数分配各组标志值 x_i 对平均数 $\bar{x}$ 的 K 阶中心动差，或称中心矩，通常用 v_k 表示。

$$v_k = \frac{\sum_{i=1}^{n}(x_i - x)^k f_i}{\sum_{i=1}^{n}f_i} \tag{3.34}$$

当 $k=0$ 时，即零阶中心动差 $v_0=1$；

当 $k=1$ 时，即一阶中心动差 $v_1=0$；

当 $k=2$ 时，即二阶中心动差 $v_2=\sigma^2$。

二、偏度

偏度是用于衡量分布的不对称程度或偏斜程度的指标。如果用矩的方式测定，偏度指标 α 是变量的三阶中心动差除以标准差的三次方，用公式表示为：

$$\alpha = \frac{v_3}{\sigma^3} = \frac{v_3}{(v_2)^{\frac{3}{2}}} \tag{3.35}$$

当分布对称时，变量的所有奇数阶中心矩均为 0，要判断分布是否对称，可考虑用奇数阶中心矩测定。一阶中心矩恒为 0，五阶以上的中心矩计算较为繁琐，偏度指标 α 就是以三阶中心动差来测定的。由于三阶中心矩含有计量单位，为消除计量单位的影

响，以 σ^3 除之。

正态分布曲线左右完全对称，三阶中心动差 v_3 等于 0，即 $\alpha=0$。当分布不对称时，则三阶中心动差不为 0，其分布的偏斜程度使 α 大于 0 或小于 0。如图 3－12 所示，当 $\alpha=0$ 时为正态分布；当 $\alpha>0$ 时为正偏斜；当 $\alpha<0$ 时为负偏斜。

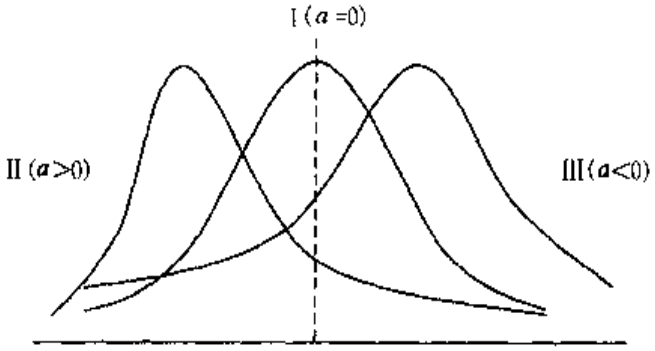


图 3－12

三、峰度

峰度是用于衡量分布的集中程度或分布曲线的尖峭程度的指标。峰度指标 β 的计算公式如下：

$$\beta=\frac{v_4}{\sigma^4}-3=\frac{\sum(x-\bar{x})^4 f}{\sigma^4\sum f}-3 \tag{3.36}$$

分布曲线的尖峭程度与偶数阶中心矩的数值大小有直接的关系，v_2 是方差，于是就以四阶中心动差 v_4 度量分布曲线的尖峭程度。v_4 含有计量单位，其计量单位同 σ^4。为消除计量单位的影响，将 v_4 除以 σ^4，就得到无量纲的相对数。因为衡量分布的集中程度或分布曲线的尖峭程度是以正态分布的峰度作为比较标准的，在正态分布条件下，

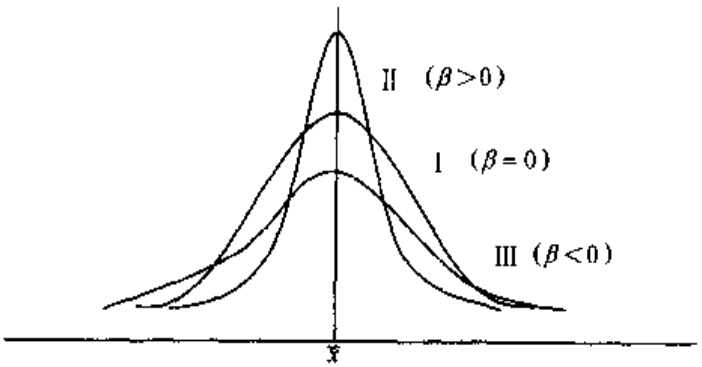


图 3－13

$v_4/\sigma^4=3$,将各种不同分布的尖峭程度与正态分布比较,即 v_4/σ^4 减 3,就得峰度指标 β 的测定公式。

当峰度指标 $\beta>0$ 时,表示分布比正态分布更集中在平均数周围,分布呈尖峰状态;$\beta=0$ 时,分布为正态分布;$\beta<0$ 时,表示分布比正态分布更分散,分布呈低峰态,如图 3-13 所示。

第五节 Excel 在统计描述中的运用

一、编制分布数列

【例 3-19】某地区进行农产量抽样调查,抽取 50 个单位构成样本。样本资料如表 3-18 所示。

表 3-18 亩*产量数据 单位:千克

240	440	400	375	500	528	412	425	320	400	520	600	485
600	588	412	444	430	342	364	450	455	465	500	580	345
340	320	300	403	445	360	517	390	410	360	422	428	520
400	425	385	440	380	548	304	448	605	325	500		

* 亩非标准单位,但鉴于习惯本书中依然采用

要求:①编制次数分布数列(组距为 100,进行等距分组),并绘制次数分布图;②分别向上累计和向下累计并绘制累计分布图。

解:

本例可用"直方图"分析工具或 FREQUENCY 等函数来完成。但"直方图"工具不能直接进行向下累计,需要借助 Excel 公式才可进行。

(一)用"直方图"工具进行分析

直方图工具,用于在工作表中数据单元格区域和接收区间给定的情况下,计算数据的个别和累积频率,还可以统计和有限集中某个数值元素的出现次数。

主要步骤如下:

1.输入数据。如图 3-15 所示,A 列输入亩产量,A1 为列标志"亩产量",C 列为分组标志,B2:B7 为分组组限,B1 为列标志"组限"。在确定组限时,需要注意:①组限指的是某一组的上限,如 299.9 是 300 以下(也即 200~300)这一组的上限,399.9 是 300~400这一组的上限;②对未排序的亩产量数据,可先用 MAX 与 MIN 函数确定亩

产量的最大与最小值，以方便确定组限；③“直方图”工具分组时采用的是“上限在内”原则，为符合统计分组的“上限不在内”的原则，每个组限均不应与亩产量相同，如本例中组限 399.9 表示将统计出亩产量大于 299.9 而小于等于 399.9 的数据点个数，如果亩产量中有 399.9 这个数，则可将 300～400 这一组的上限上为 399.99。

2. 调出“直方图”对话框，其主要选项的含义如下所述。

输入区域：在此输入待分析数据区域的单元格范围。本例输入区域为“$A$1：$A$51”。

接收区域(可选)：在此输入接收区域的单元格范围，该区域应包含一组可选的用来计算频数的边界值。这些值应当按升序排列。只要存在的话，Excel 将统计在当前边界点和相邻的高值边界点之间的数据点个数。如果某个数值等于或小于某个边界值，则该值将被归到以该边界值为上限的区间中(注意：与统计分组“上限不在内”原则有所区别)。如果省略此处的接收区域，Excel 将在数据组的最小值和最大值之间创建一组平滑分布的接收区间。本例接收区域为“$B$1：$B$6”。

标志：如果输入区域的第一行或第一列中包含标志项，则选中此复选框；如果输入区域没有标志项，则清除此该复选框，Excel 将在输出表中生成适宜的数据标志。

柏拉图：选中此复选框，可以在输出表中同时显示按降序排列频率数据。如果此复选框被清除，Excel 将只按升序来排列数据。

累积百分比：选中此复选框，可以在输出结果中添加一列累积百分比数值，并同时在直方图表中添加累积百分比折线。如果清除此选项，则会省略以上结果。

图表输出：选中此复选框，可以在输出表中同时生成一个嵌入式直方图表。

本例“直方图”对话框的填写如图 3－14 所示。

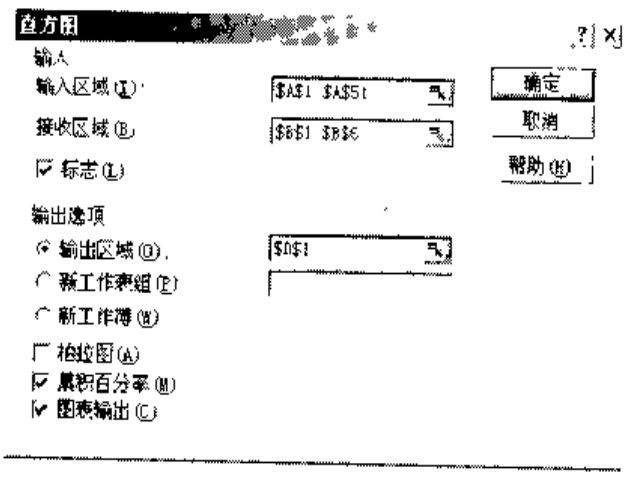


图 3－14

3. 单击“确定”按钮即可。完整的结果通常包括三列和一个频率分布图，第一列是

数值的区间范围，第二列是数值分布的频数，第三列是频数分布的累积百分比。在输出的结果中，“频率”指的是统计中所指的“频数”，“累积%”指的是“向上累计频率”。结果中原本还有一“其他”行的数据，该行是亩产量大于699.9的频数与累计频率(位于单元格区域D7:F7)，因本例中没有大于699.9的数据，所以可将该行删除，即选定D7:F7，然后执行菜单命令“编辑”→“删除”命令。由于该工具总会自动统计出大于最大组限的数据的频数，所以本例中组限也可以不使用699.9，即在第2步中接收区域为“$B$1:$B$6”，在输出的结果中，将“其他”更改为我们所需要的标志性文字，如改为“699.9”。经过修改的结果如图3-15所示。

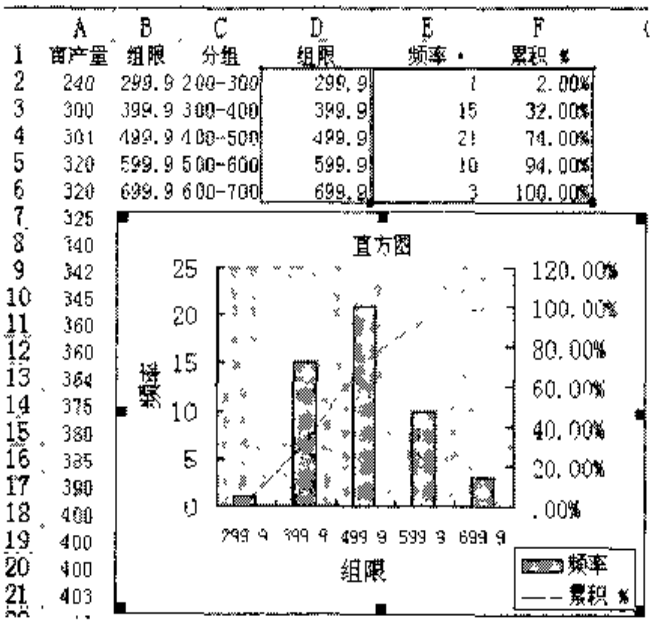


图3-15

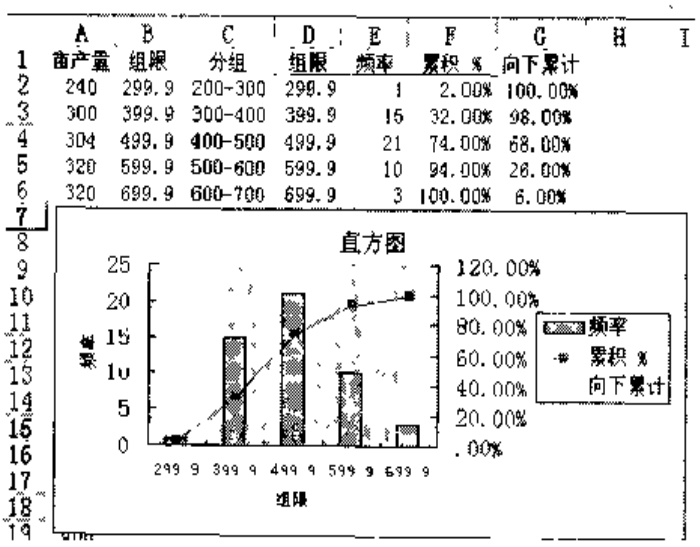


图3-16

4.计算向下累计频率。在G1单元格输入列标志“向下累计”。在G6单元格输入“=E6/SUM($E$2:$E$6)”,计算出最后一组的频率,在G5单元格中输入“=E5/SUM($E$2:$E$6)+G6”得到最后一组和倒数第二组的累计频率,将G5单元格的公式复制到G2:G4,计算出向下累计这一列。

5.在图表中加入向下累计曲线。选定图表区,此时菜单上出现“数据”按钮变为“图表”按钮,执行菜单命令“图表”→“数据源”,调出“源数据”对话框。在“源数据”对话框的“系列”页框中,单击“添加”按钮,输入所添加系列的名称与值。要输入系列名称,请单击“名称”文本框,再用鼠标单击工作表中的G1单元格,此时“名称”文本框显示“=分布数列!$G$1”(“分布数列”为工作表标签名);要输入系列值,请单击“值”文本框,再用鼠标在工作表上选定G2:G6单元格,此时“值”文本框中显示“=分布数列!$G$2:$G$6”。最后单击“确定”按钮,结果见图3-16。

(二)用FREQUENCY函数编制分布数列的主要步骤

1.输入数据。如图3-17所示,A,B,C列输入的数据同第一种方法,在D1:F1中依次输入列标“频数”、“向上累计频数”和“向下累计频数”。其中“组限”的含义同第一种方法。

2.选定D2:D6,输入公式“=FREQUENCY(A2:A51,B2:B6)”,然后按CRTL+SHIFT+ENTER组合键,即可计算出各组的频数。

3.计算向上累计频数。在E2单元格中输入“=D2”,在E3单元格中输入公式“=D3+E2”,再将公式复制到E4:E6。

4.计算向下累计频数。在F6单元格中输入“=D6”,在F5单元格中输入公式“=D5+F6”,再将公式复制到F2:F4。计算结果如图3-17所示。

	A	B	C	D	E	F
1	亩产量	组限	分组	频数	向上累计频数	向下累计频数
2	240	299.9	200-300	1	1	50
3	300	399.9	300-400	15	16	49
4	304	499.9	400-500	21	37	34
5	320	599.9	500-600	10	47	13
6	320	699.9	600-700	3	50	3
7	325					

图3-17

5.根据得到的计算结果,可用Excel的图表功能绘制出亩产量分布的直方图、折线图等。

二、计算有关的描述统计指标

(一)利用"描述统计"分析工具

【例 3-20】表 3-19 是 50 名工人完成某一装配工序所需时间:

表 3-19　　单位:分钟

35	38	44	33	44	43	48	40	45	30
45	32	42	39	49	37	45	37	36	42
35	41	32	46	34	30	43	37	44	49
36	46	45	36	37	37	45	36	46	42
38	43	34	38	47	35	29	41	40	41

要求对装配工时进行描述统计分析。

解:

主要操作步骤如下。

(1)输入数据。如图 3-19 所示,在 A2:A51 输入时间数据,在 A1 输入列标志"工时"。

(2)调出"描述统计"对话框,其主要选项的含义如下。

"输入区域":在此输入待分析数据区域的单元格范围。本例输入区域为"$A$1:$A$51"。

"分组方式":如果需要指出输入区域中的数据是按行还是按列排列,则单击"逐行"或"逐列"。本例分组方式为"逐列"。

标志位于第一行/列:如果输入区域的第一行中包含标志项(变量名),则选中"标志位于第一行"复选框,本例要选中该复选框;如果输入区域的第一列中包含标志项,则选中"标志位于第一列"复选框;如果输入区域没有标志项,则不选任何复选框,Excel 将在输出表中生成适宜的数据标志。

"输出区域":在此框中可填写输出结果表左上角单元格的地址,用于控制输出结果的存放位置。本例输出区域填"$C$1",整个输出结果分为两列,左边一列包含统计标志项,右边一列包含统计值。根据所选择的"分组方式"选项的不同,Excel 将为输入表中的每一行或每一列生成一个两列的统计表。

"新工作表":单击此选项,可在当前工作簿中插入新工作表,并由新工作表的 A1 单元格开始存放计算结果。如果需要给新工作表命名,则在右侧编辑框中键入名称。

"新工作簿":单击此选项,可创建一新工作簿,并在新工作簿的新工作表中存放计算结果。

"汇总统计":指定输出表生成下列统计结果,则选中此复选框。这些统计结果有:样本的平均值($\bar{X}$),抽样平均误差($S/\sqrt{n}$),组中值(Median),众数(Mode),样本标准差

(S),样本方差(S^2),峰度值,偏度值,极差(max－min),最小值(min),最大值(max),样本总和,样本容量(n)和一定显著水平下总体均值的置信区间。本例选中该复选框。

"平均数置信度":若需要输出由样本均值推断总体均值的置信区间,则选中此复选框,然后在右侧的编辑框中,输入所要使用的置信度。例如,置信度95%可计算出的总体样本均值置信区间为10,则表示:在5%的显著水平下总体均值的置信区间为($\bar{X}-10,\bar{X}+10$)。本例平均数置信度为95%。

第k个大/小值:如果需要在输出表的某一行中包含每个区域的数据的第k个最大/小值,则选中此复选框。然后在右侧的编辑框中,输入k的数值。

本例"描述统计"对话框的填写如图3－18所示。

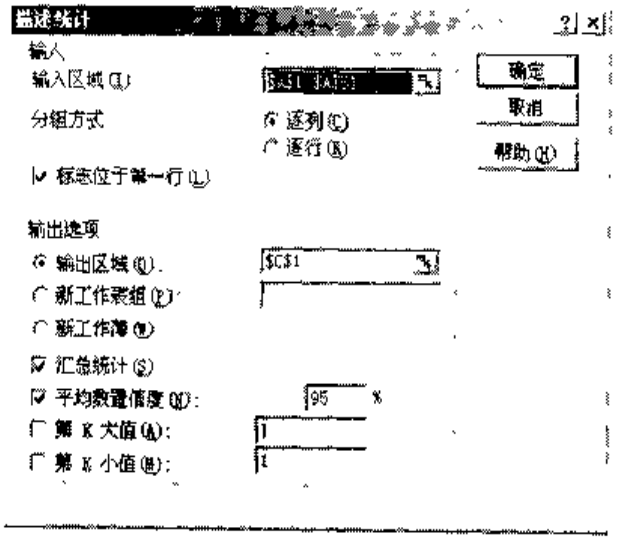
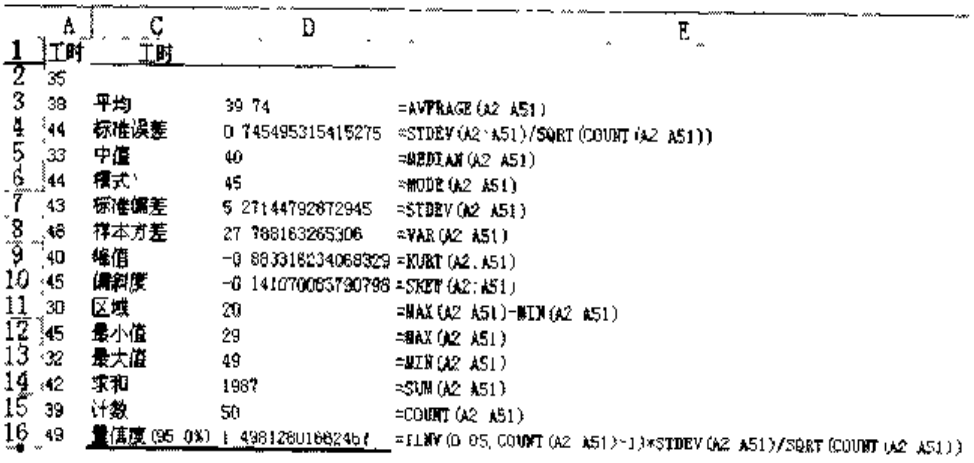


图3－18

(3)"描述统计"对话框填完后,单击"确定"按钮,结果如图3－19所示。



	A	C	D	E
1	工时	工时		
2	35			
3	38	平均	39.74	=AVERAGE(A2:A51)
4	44	标准误差	0.745495315415275	=STDEV(A2:A51)/SQRT(COUNT(A2:A51))
5	33	中值	40	=MEDIAN(A2:A51)
6	44	模式	45	=MODE(A2:A51)
7	43	标准偏差	5.271447928729 45	=STDEV(A2:A51)
8	48	样本方差	27.788163265306	=VAR(A2:A51)
9	40	峰值	-0.883318234068329	=KURT(A2:A51)
10	45	偏斜度	-0.141070063730798	=SKEW(A2:A51)
11	30	区域	20	=MAX(A2:A51)-MIN(A2:A51)
12	45	最小值	29	=MAX(A2:A51)
13	32	最大值	49	=MIN(A2:A51)
14	42	求和	1987	=SUM(A2:A51)
15	39	计数	50	=COUNT(A2:A51)
16	49	置信度(95.0%)	1.498128018824 6	=FINV(0.05,COUNT(A2:A51)-1)*STDEV(A2:A51)/SQRT(COUNT(A2:A51))

图3－19

有两点需要注意:①各种分析工具对话框的输出选项中"输出区域"、"新工作表组"和"新工作簿"的含义基本相同,以下不再赘述。②图 3－19 中的指标名称与统计中的习惯叫法不大一致,确切的指标名称如本例操作步骤第 2 步中所述。

(二)利用统计函数完成

对于分组数据,不能直接用"描述统计"分析工具来计算描述统计的有关指标,应综合应用 Excel 的公式与函数来实现。

【例 3－21】某粮食作物的产量和播种面积资料如表 3－20。

表 3－20

亩产量 x	400～500	500～600	600～700	700～800	800～900	900～1 000	合计
播种面积 f	6	30	50	60	40	14	200

根据表 3－20 资料计算加权平均亩产量、标准差以及偏度和峰度等。

解:

主要操作步骤如下。

(1)输入数据。如图 3－20 所示,A2:A7 输入亩产量的分组,这些数据在 Excel 中被认为是文本,不能直接参加数值运算,A1 输入本列标志"亩产量"。B2:B7 输入各组的组中值,B1 输入本列标志"组中值"。C2:C7 输入播种面积数据,C1 为本列标志"播种面积"。B9:B15 存放的是最终结果与一些中间变量值,对应的 A9:A15 则是它们的名称。

(2)定义变量名。先定义样本数据组中值及播种面积的变量名。选定 B1:C7,执行菜单命令"插入"→"名称"→"指定",单击"首行"选项,最后单击"确定"按钮。再定义最终结果与中间变量的名称,选定 A9:B15,执行菜单命令"插入"→"名称"→"指定",单击"最左列"选项,最后单击"确定"按钮。

(3)计算加权平均亩产量。在 B9 单元格输入如下公式,"＝SUMPRODUCT(组中值＊播种面积)/SUM(播种面积)"。

当然,如果没有定义变量名,则要使用单元格引用,公式如下,"＝SUMPRODUCT(B2:B7＊C2:C7)/SUM(C2:C7)"。

(4)计算其他指标。在 B10:B14 中,依次输入以下内容,"＝SQRT(SUMPRODUCT((组中值－加权平均亩产量)^2＊播种面积)/SUM(播种面积))";"＝SUMPRODUCT((组中值－加权平均亩产量)^3＊播种面积)/SUM(播种面积)";"＝SUMPRODUCT((组中值－加权平均亩产量)^4＊播种面积)/SUM(播种面积)";"＝三阶动差/标准差^3";"＝四阶动差/标准差^4－3"。

(5)计算结果如图 3－20 所示。

这里有两点需要说明：①可参照本例的方法计算分组情况下的其他有关描述统计

	A	B	C
1	亩产量	组中值	播种面积
2	400～500	450	6
3	500～600	550	30
4	600～700	650	50
5	700～800	750	60
6	800～900	850	40
7	900～1 000	950	14
8			
9	加权平均亩产量	720	
10	标准差	122.882057	
11	三阶动差	-1140000	
12	四阶动差	543970000	
13	偏度	-0.0614383	
14	峰度	-0.6142713	
15	众数	733.333333	

图 3－20

的指标，如方差、标准差等；②对于众数及中位数的计算，不需要特别的技巧，只要将统计中的计算公式在 Excel 中实现即可。这时 Excel 更像一个普通的计算器，但 Excel 可记下公式，便于检查与修改，同时在数据量很大，分组很多时，借助 Excel 的相关函数可很快地找出众数组和中位数所在组。如本例众数的计算，首先确定众数组，为“700～800”这一组。然后根据等距分组的上限公式，在 B15 单元格输入公式“＝700＋(C5－C4)＊100/((C5－C4)＋(C5－C6))”即可，其中 700 为众数组下限，100 为组距，结果如图 3－20 所示。

本章小结

1. 统计表的结构从格式上，可分为总标题、横行标题、纵栏标题和指标数值四个部分；从内容上看可分为主词栏和宾词栏两部分。

2. 常用的统计图有直方图、折线图和曲线图。洛伦茨曲线是一种累计曲线图，利用洛伦茨曲线计算出的基尼系数可以用来衡量收入分配是否公平。

3. 算术平均数有简单算术平均数和加权算术平均数两种。权数起权衡轻重的作用。当各组频数都相等时，权数不起作用。求相对指标的平均数应选择合适的权数。

4. 调和平均数是各标志值倒数平均数的倒数，有简单调和平均数和加权调和平均数两种。

5. 几何平均数用于计算时间上相互衔接的比率的平均数。

6.算术平均数、几何平均数和调和平均数是 k 阶幂平均数。当分别 $k=1, k\to 0, k=-1$ 时,可以得到各种不同数值平均数的形式。

7.众数和中位数都是位置平均数。众数是总体中最普遍出现的标志值。中位数是总体各单位按大小顺序序列化后,处在数列中点位置上那个单位的标志值。对未分组的资料和由单项式及组距式分组的变量数列,众数和中位数各有不同的确定方法。

8.众数、中位数和算术平均数各自的含义、计算方法不同,但三者之间存在一定的数量联系。对称分布时,三者相等;适度正偏斜时,$M_o < M_e < \bar{x}$;适度负偏斜时则相反。适度偏斜情形下,$|M_e - M_o| \approx 2|\bar{x} - M_e|$。

9.极差与四分位差。极差是总体中两个极端标志值的差。四分位差则是剔除了最大和最小各 1/4 单位后的极差。

10.方差与标准差。标准差是方差的平方根,它们都是测定总体标志变异程度最常用、最灵敏的指标。

11.极差、四分位差和标准差都是绝对量指标,它们的计量单位与标志值相同。它们的大小不仅取决于总体的变异情况,也和标志值水平的高低、计量水平的不同有关。

12.比较标志值水平不同、计量单位不同的总体之间的变异情况,不宜直接用绝对量变异指标,而应用相对数变异指标即变异系数。最常用的变异系数是标准差系数:$V=\frac{\sigma}{\bar{x}}\times 100\%$。

13.偏度是衡量分布的不对称程度或偏斜程度的指标。偏度指标是三阶中心动差除以标准差的三次方。峰度是测定分布的集中程度,即分布曲线的尖峭程度的指标。峰度指标是四阶中心动差除以标准差的四次方再减去 3。

14.在描述统计中,Excel 主要有以下几方面的应用:(1)"描述统计"分析工具可以计算分布的集中、离散趋势和偏度、峰度等有关指标,但该工具仅适用于没有分组的分布数列,对于分组的分布数列,要用相关统计和数学函数来计算;(2)"直方图"分析工具可以计算分布数列的个别和累积频率(频数),并绘制出直方图;(3)利用 FREQUENCY 等函数也可以方便地编制分布数列,如频数分布数列、累计频数数列等,并可根据编制的数列绘制直方图、折线图等统计图。

思考与练习题

一、选择题

1.有 5 辆汽车在同一距离的公路上行驶的速度资料,确定汽车平均每小时行驶速度的平均数的公式是(　　)。

a. $\frac{\sum x}{n}$　　b. $\frac{\sum xf}{\sum f}$　　c. $\frac{n}{\sum \frac{1}{x}}$　　d. $\frac{\sum m}{\sum \frac{m}{x}}$

2. 加权算术平均数的大小取决于:(甲)频数绝对量的大小;(乙)频数之间的比率;(丙)变量值的大小。(　　)

a. 甲丙　　b. 乙　　c. 甲乙　　d. 乙丙

二、问答题

1. 指出表 3-21,表 3-22 和表 3-23 中的主词与宾词,横行标题与纵栏标题,并根据主词的结构分类法,指出该统计表的类型。

表 3-21　××市 2000 年和 2001 年国内生产总值表

指标	单位	2001 年	2000 年	2001 年比 2000 年增加(%)
国内生产总值	万元	622 051	505 774	23.0
第一产业	万元	64 385	60 782	5.9
第二产业	万元	334 301	258 661	29.2
第三产业	万元	223 365	186 331	19.9
人均国内生产总值	元/人	5 411	4 426	22.3

表 3-22　国内商业主要经济统计指标

2001 年×月

指标	总额(亿元)	比上年同期增长(%)
社会商品零售总额	1 088	22.2
消费品零售总额	964	25.2
居民	848	24.5
社会集团	116	30.5
农业生产资料零售额	124	2.4

表 3-23　××市饮食业机构、人员基本情况

2001 年

按经济类型分组	合计		市		县		县以下	
	机构	人员	机构	人员	机构	人员	机构	人员
国有饮食业	35	324	26	245	9	79	—	—
集体饮食业	195	1 129	164	1 023	4	20	27	86
合营饮食业	17	1 144	17	1 144	—	—	—	—
个体有证饮食业	2 436	4 873	1 746	3 549	294	597	396	727

2.抽样调查某省50户城镇居民平均每人全年可支配收入资料如表3-24所示。

表3-24 居民年人均可支配收入 单位:百元

88	77	66	85	74	92	67	84	77	94
58	60	74	64	75	66	78	55	70	66
78	64	65	87	49	97	77	69	68	71
65	78	77	86	78	82	98	95	86	100
66	74	70	62	68	56	83	52	71	108

要求:

(1)根据上述资料编制次(频)数分布数列。

(2)编制向上和向下累计频数、频率数列。

(3)根据所编制的次数分布数列绘制直方图、折线图与曲线图,并说明其属于何种分布类型。

(4)根据所编制的向上(向下)累计频数(频率)数列绘制累计曲线图。

(5)根据频数分布曲线图说明居民年人均可支配收入的分布类型。

3.请搜集一实际资料,利用洛伦茨曲线对收入分配的公平度进行评价。

4.本章例3-6中,按调和平均法和按算术平均法计算的结果一致,根据幂平均数是参数 k 的单调不减函数的性质,算术平均数≥调和平均数,这两者是否存在矛盾?

5.甲乙两企业生产三种产品的单位成本和总成本资料如表3-25,试比较哪个企业的平均成本高,并分析其原因。

表3-25

产品	单位成本(元)	总成本(元)	
		甲企业	乙企业
A	15	2 100	3 255
B	20	3 000	1 500
C	30	1 500	1 500

6.甲、乙两市场农产品价格及成交量资料如表3-26,试比较哪个市场的平均价格高,并分析其原因。

表3-26

品种	价格(元/千克)	甲市场成交额(万元)	乙市场成交量(万千克)
甲	1.2	1.2	2
乙	1.4	2.8	1
丙	1.5	1.5	1
合计	—	5.5	4

7.某企业工人平均月工资为1 440元,月收入少于1 280元的占一半,试估计众数,并对该企业工人工资的分布情况做一简要说明。

8.某城市对3 000户居民人均月消费支出进行调查,得到表3-27的资料。

表3-27

居民户月均支出(元)	户数	比重(%)
200以下	30	1
200~300	180	6
300~400	450	15
400~500	600	20
500~600	1 050	35
600~700	300	10
700~800	180	6
800~900	120	4
900~1 000	60	2
1 000以上	30	1
	3 000	100.00

要求:

(1)计算居民总平均月支出。

(2)计算居民月均支出标准差。

(3)计算居民月均支出中位数和众数。

9.已知标志值总和为415,标志值平方和为1 775,总体单位数为100,求其方差和标准差。

10.某管理局下属8家企业的产品销售数据如表3-28所示。

表3-28

单位:万元

企业	产品销售额(X_1)	销售利润(X_2)
1	170	8.1
2	220	12.5
3	390	18.0
4	430	22.0
5	480	26.5
6	650	40.0
7	950	64.0
8	1 000	69.0

要求:比较其产品销售额和销售利润的离散程度。

11.某生产班组11个工人日生产零件数为:15,17,19,20,22,22,23,23,25,26,30。要求:

(1)计算平均数和方差。

(2)按照15～19,20～24,24以上分成三组,计算组内方差和组间方差。

(3)验证总方差等于组间方差与组内方差平均数之和。

12.某高校学生参加英语四级考试的优秀率和合格率分别为15%和90%,试计算优秀率和合格率分布的方差和标准差。

13.某粮食作物的产量和播种面积资料如表3-29。

表3-29

亩产量 x	400-500	500-600	600-700	700-800	800-900	900-1 000	合计
播种面积 f	6	30	50	60	40	14	200

要求:测定其偏度和峰度。

第四章

抽样分布与参数估计

★ 本章学习要点与要求 ★

本章介绍有关概率和抽样分布的基础知识，阐述参数估计的理论与方法。通过学习，要求：①理解概率的基本知识，掌握正态分布的概率计算；②理解有关抽样的基本概念以及不同抽样方式对抽样分布的影响；③熟悉在优良性标准下总体参数的点估计公式；④掌握总体均值与成数指标的区间估计方法；⑤了解抽样设计的基本内容；⑥能运用 Excel 求解概率、计算有关样本统计量，进行总体参数的区间估计。

第一节　频率、概率与概率分布

一、随机事件与概率

（一）随机试验与事件

随机现象在自然界和社会经济生活中都普遍存在，如产品抽样检验，抽到的产品可能是正品，也可能是次品；投一枚硬币可能出现正面，也可能出现反面；手机通话，可能一次拨通，也可能要二次、三次，甚至更多次才能拨通。随机现象的特点是：在基本条件不变的情况下，一系列的试验或观测会得到不同的结果，并且在试验或观测前不能预见何种结果将出现。虽然随机现象的结果不能预见，但这并不意味我们在随机现象面前

无能为力。我们仍然可以通过随机试验,以及其他相关知识来认识随机现象的规律。随机试验,简称试验,它必须满足以下的性质:①每次试验的可能结果不是惟一的;②每次试验之前不能确定何种结果会出现;③试验可在相同条件下重复进行。

在随机试验中,可能出现也可能不出现的结果,被称为随机事件,简称事件。试验的结果可能是一个简单事件,也可能是一个复杂事件。简单事件就是不可以再分解的事件,又称为基本事件。复杂事件是由简单事件组合而成的事件。基本事件还可被称为样本点,设试验有 n 个基本事件,分别记为 $\omega_i(i=1,2,\cdots,n)$。集合 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$,称为样本空间,$\Omega$ 中的元素就是样本点。

【例 4-1】投掷一粒均匀的六面体骰子,出现的点数有可能是 1,2,3,4,5,6 共六种。这六种结果是基本结果,不可以再分解成更简单的结果了,所以 $\Omega=\{1,2,3,4,5,6\}$ 为该试验的样本空间。"出现点数是奇数"的结果就不是简单事件,它是由基本事件 $\{1\}$,$\{3\}$ 和 $\{5\}$ 组合而成的。我们通常用大写字母 A,B,C,……来表示随机事件,设 A 表示"出现点数是奇数",则 $A=\{1,3,5\}$;设 B 表示"出现点数是偶数",则 $B=\{2,4,6\}$。

与随机现象相对应的是确定性现象。所谓确定性现象,是指在一定的条件下,其结果能够明确预见的现象。例如,在通常气压条件下,水在摄氏零度时一定会结冰,这是一个必然的结果。为了方便与统一,我们把确定性现象的结果也看做一种特殊的随机事件:必然事件用样本空间 Ω 表示;不可能出现的试验结果称为不可能事件,用空集 Φ 表示。

我们把"A 发生或 B 发生"事件记为 $A\cup B$;把"A 与 B 同时发生"事件记为 $A\cap B$,或 AB。[例 4-1]中:$A\cup B=\Omega$,是必然事件;$AB=\Phi$,是不可能事件。如果 $AB=\Phi$,称 A 与 B 不相容。

(二)概率

1.概率的定义。概率就是指随机事件发生的可能性,或称为机率,是对随机事件发生可能性的度量。如何计算概率?最直观的方式就是进行重复试验,通过试验的频率来计算概率。

例如,不断重复地投掷一枚均匀硬币,出现正面的频率会稳定在 1/2 附近。历史上,有人为此作过试验,结果如表 4-1 所示。

表 4-1 频率试验结果

实验者	投掷硬币的次数(n)	正面朝上的次数(m)	频率 $=m/n$
Demorgan	2 048	1 061	0.518 0
Buffon	4 040	2 048	0.506 9
Pearson(1)	12 000	6 019	0.501 6
Pearson(2)	24 000	12 012	0.500 5

上述试验表明,试验次数越多,出现正面朝上的频率越接近0.5。这是因为均匀的硬币掷出正面、反面的可能性都是一样大的。

下面给出概率的频率定义,也称为概率的统计定义:

进行 n 次重复试验,随机事件A发生的次数是 m 次,发生的频率是 m/n,当试验的次数 n 很大时,如果频率在某一数值 p 附近摆动,而且随着试验次数 n 的不断增加,频率的摆动幅度越来越小,则称 p 为事件A发生的概率,记为:

$$P(A)=p \tag{4.1}$$

由于频率取值在0~1之间,因此:$0\leqslant p\leqslant 1$。

如果随机试验的样本空间是有限集合,所有样本点出现的可能性相同,则事件A的概率可根据以下公式计算:

$$P(A)=\frac{m}{n}=\frac{\text{A包含的样本点个数}}{\text{样本点总数}} \tag{4.2}$$

这样的概率计算模型,称为古典模型。

【例4-2】设一个袋子中装有白球2个,黑球3个,从中随机摸出1只球,问刚好是白球的概率有多大?

解:

由于摸出的任何1只球都形成一个基本事件,所以样本点总数为 $n=5$。用A表示摸出的是白球事件,则A由两个基本点组成,即A={白球,白球},$m=2$。因此,刚好摸出白球的概率为:

$$P(A)=m/n=2/5=0.4$$

2.概率的基本性质。概率最基本的两个性质是:性质1:$P(A)\geqslant 0$;性质2:$P(\Omega)=1$。性质1称为概率的非负性,性质2称为概率的规范性。除此外,概率的性质还有可加性:性质3:若事件A与事件B互不相容,则:

$$P(A\cup B)=P(A)+P(B) \tag{4.3}$$

推论1　不可能事件的概率为0,即:

$$P(\Phi)=0$$

$\bar{A}$ 表示A的对立事件,即它们二者必有一事件发生,但又不能同时发生,则有如下推论。

推论2

$$P(\bar{A})=1-P(A) \tag{4.4}$$

这个推论非常有意义,其提供的思想是:当直接对事件A求概率有困难时,可以考虑去求其对立事件的概率。

【例4-3】袋中装有4只黑球和1只白球,每次从袋中随机地摸出1只球,并换入1只黑球。连续进行,问第三次摸到黑球的概率是多少?

解:

记 A 为"第三次摸到黑球",则 $\bar{A}$ 为"第三次摸到白球"。先计算 $P(\bar{A})$。

由于袋中只有 1 只白球,如果某一次摸到了白球,换入了黑球,则袋中只有黑球了。所以 $\bar{A}$ 相当于第一、第二次都是摸到黑球,第三次摸到白球。注意这是一种有放回的摸球,样本点总数为 5^3,$\bar{A}$ 包含的样本点个数是 $4^2\times1$。故:$P(\bar{A})=(4^2\times1)/5^3=16/125$,所以 $P(A)=1-P(\bar{A})=1-(4^2\times1)/5^3=109/125$。

如果直接计算 $P(A)$则要困难多了。

3.事件的独立性。如果事件 B 发生与否对事件 A 没有影响,可以说,事件 A 与事件 B 两者的出现存在某种"独立性",数学上严格的定义是:

定义　对事件 A 与 B,若 $p(AB)=p(B)p(A)$,则称它们是统计独立的,简称相互独立。

独立性在统计学中是非常重要的概念,我们在许多地方要用到它。

二、随机变量

为了便于应用数学方法来分析随机现象,有必要对随机试验的结果进行量化处理,因此要引进随机变量的概念。在随机现象中,有些试验结果直接与数量有关。例如,投一枚骰子,可能出现的点数是:{1,2,3,4,5,6}。试验结果可用点数这个变量 X 进行描述,X 的可能取值就是:1,2,3,4,5,6。每一取值代表一种试验的结果。有些随机试验的结果表现为定性形式,但是,定性结果也可以通过赋值的方式,进行量化处理。例如,从班上 45 人(其中男生 20 人)中任意抽出一人,恰好是男生。在这里,试验结果是随机的,抽出一人是男是女,本身不具有数量性。如果对一个变量 X 进行赋值,使试验结果与 X 的取值一一对应起来:抽出男生,$X=1$;抽出女生,$X=0$,则变量 X 的取值就完全反映出了试验的结果,达到了对试验结果量化处理的目的。

随机变量 X 是定义在样本空间 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$ 上的一个函数,这个函数的取值随试验的结果不同而变化。这个函数还要求满足条件:对任意的实数 x,$X<x$ 是随机事件,且概率存在。如果随机变量所有可能的取值是有限的,或可排成一列的,这种随机变量称为离散型随机变量,如投骰子试验中出现的点数,就是一个离散型随机变量;另一种情况是随机变量的取值范围是一个区间或整个数轴,这种随机变量称为连续型随机变量。连续型随机变量取值充满整个区间,不能像离散型随机变量那样,可以一一列举,这种性质的不同使得我们须用不同的方式来描述他们的概率分布。

(一)离散型随机变量的概率分布

设离散型随机变量 X 的所有可能取值为 $x_1,x_2,\cdots,x_n,\cdots$,相应的概率为 $p(x_1)$,$p(x_2),\cdots,p(x_n),\cdots$。用表格统一表示,如表 4-2 所示。

表 4-2

X	x_1	x_2	…	x_n	…
P	$p(x_1)$	$p(x_2)$	…	$p(x_n)$	…

这称为离散型随机变量 X 的概率分布。也可简单记为：

$$p(X=x_i)=p(x_i) \quad (i=1,2,\cdots)$$

概率分布有以下性质：

1. $0\leqslant p(x_i)\leqslant 1 \quad (i=1,2,\cdots)$；

2. $\sum_i p(x_i)=1$。

离散型随机变量 X 的期望值为：

$$\mu=E(X)=\sum_i x_i p(x_i) \tag{4.5}$$

随机变量的期望值也即其平均值，是随机变量分布的集中趋势，即分布的中心位置。容易验证期望值满足性质：$E(\alpha X_1+\beta X_2)=\alpha E(X_1)+\beta E(X_2)$，其中 X_1,X_2 都是随机变量，α,β 是任意常数。并且这个性质可推广到多个随机变量情形。

方差为：

$$\sigma^2=Var(X)=E(X-\mu)^2=\sum_i(x_i-\mu)^2p(x_i) \tag{4.6}$$

方差的平方根，称为标准差。方差 σ^2 或标准差 σ 反映随机变量 X 对其期望值的离散程度，σ^2 或 σ 越小，说明期望值的代表性越好；σ^2 或 σ 越大，说明期望值的代表性越差。也容易验证：对于任意的 α，$\sigma^2(\alpha X)=\alpha^2\sigma^2(X)$ 成立。

在一些问题中，我们只对试验中某事件 A 是否出现感兴趣，如果 A 发生，我们称“成功”，否则称“失败”。像这样只有两种结果的试验称为贝努里试验。设 A 出现的概率为 p，我们重复进行 n 次贝努里试验，称为 n 重贝努里试验，有时也简称贝努里试验。

以 B_k 表示 n 重贝努里试验中事件 A 正好出现 k 次这一事件，则

$$p(B_k)=C_n^k p^k q^{n-k} \quad (k=0,1,2,\cdots,n) \tag{4.7}$$

显然，n 重贝努里试验中事件 A 出现的次数 k 是个随机变量，其取值范围是从 0 到 n。它的分布如表 4-3 所示。

表 4-3　二项分布表

k	0	1	2	…	n
$P(B_k)$	$C_n^0p^0q^n$	$C_n^1p^1q^{n-1}$	$C_n^2p^2q^{n-2}$	…	$C_n^np^nq^0$

由于式(4.7)是二项展开式的通项，故该分布称为二项分布。

（二）连续型随机变量的概率分布

对于连续型随机变量 X，它的取值不能一一列出，因此其概率分布形式不能同离

散型随机变量一样，通过表格方式全部表现出来。但是，对任意的实数 x，由随机变量的定义知，$X<x$ 是一随机事件，可以对它求概率，记为 $F(x)=p(X<x)$，该函数就是随机变量的分布函数。分布函数的导数称为密度函数，记作 $p(x)$。通过对密度函数积分，可得到随机变量 X 在点 x 附近或在一个区间上取值的概率。注意，连续型随机变量取一个固定的值的概率为 0。

连续型随机变量的密度函数有以下的性质：

$$①p(x)\geqslant 0;②\int_{-\infty}^{\infty}p(x)\mathrm{d}x=1;③p(a<X<b)=\int_{a}^{b}p(x)\mathrm{d}x. \tag{4.8}$$

其中，$p(a<X<b)$ 表示事件 $a<X<b$，即随机变量 X 的取值落在区间 (a,b) 内的概率。连续型随机变量在 (a,b) 区间上定积分的几何意义就是由 x 轴、被积函数 $p(x)$、直线 $x=a$ 和 $x=b$ 所围成的面积，如图 4－1 所示。

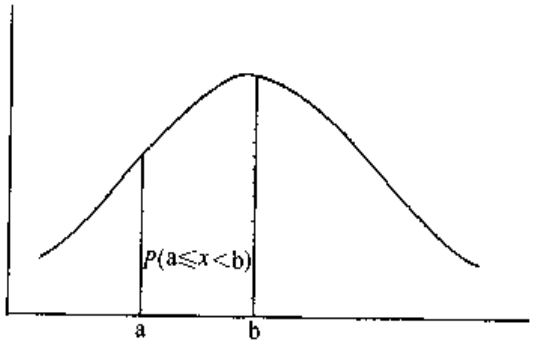


图 4－1 概率的几何意义

连续型随机变量 X 的期望值为：

$$\mu=E(X)=\int_{-\infty}^{\infty}xp(x)\mathrm{d}x \tag{4.9}$$

方差为：

$$\sigma^2=E(X-\mu)^2=\int_{-\infty}^{\infty}(x-\mu)^2p(x)\mathrm{d}x \tag{4.10}$$

可以验证，对连续型随机变量，$E(\alpha X_1+\beta X_2)=\alpha E(X_1)+\beta E(X_2)$ 和 $\sigma^2(\alpha x)=\alpha^2\sigma^2(X)$ 也都成立。

正态分布是最重要的一种连续型随机变量分布，原因有三：第一，它是最常见的一种分布，许多随机变量服从或近似服从正态分布。如同龄人的身高、体重，农作物的产量，学生考试的成绩等等；第二，许多有用的分布可以由正态分布推导出来，如卡方分

布、t 分布和 F 分布都可由正态分布导出；第三，正态分布在一定条件下，还是一些其他分布的近似分布，如大样本下的 t 分布与正态分布近似。

如果连续型随机变量 X 的密度函数为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty<x<\infty) \tag{4.11}$$

则称随机变量 X 服从均值为 μ，方差为 σ^2 的正态分布，记为 $X\sim N(\mu,\sigma^2)$。正态分布的密度函数图形一般称为正态曲线，它是一条以均值为中心的对称钟型曲线，如图 4－2 所示。

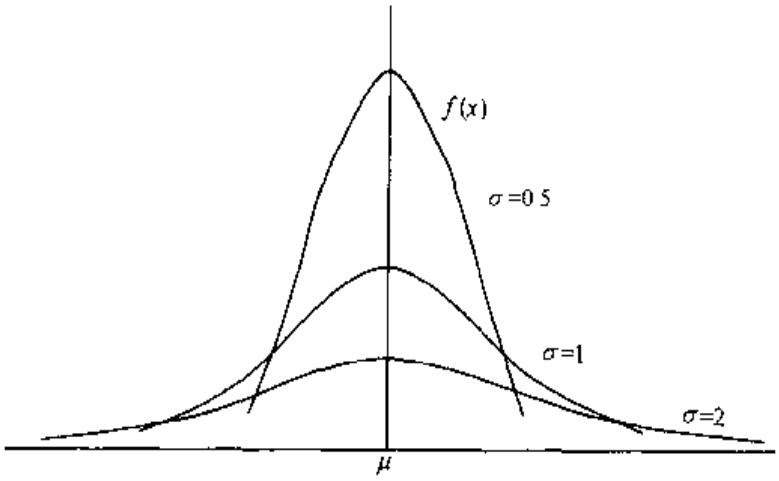


图 4－2　正态曲线

从图 4－2 可看到：μ 是该分布的中心，σ 是标准差，反映分布的离散程度，σ 越大，分布曲线越平缓，离散程度越大；σ 越小，分布曲线越陡峭，说明分布越集中。

如果一个正态分布的 $\mu=0,\sigma=1$，则称该正态分布为标准正态分布，相应的随机变量称为标准正态随机变量，用 Z 表示，即 $Z\sim N(0,1)$，相应的分布密度函数为：

$$p(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \qquad (-\infty<z<\infty) \tag{4.12}$$

标准正态随机变量在区间 $[-Z,Z]$ 取值的概率 $F(Z)$，可通过查标准正态分布概率表获得。

【例 4－4】设随机变量 Z 服从标准正态分布。求以下概率的大小。

①$p(-1<Z<1)$；②$p(0<Z<1.25)$；③$p(1<Z<1.25)$；④$p(Z>1)$。

解：

(1)概率 $p(-1<Z<1)$ 是由标准正态分布的密度函数，在区间 $(-1,1)$ 上的曲边梯形的面积，如图 4－1 所示，该概率(或该面积)可直接按临界值 $Z=1$ 查表得到：

$p(-1<Z<1)=0.682\ 7$。

(2)由标准正态分布关于 $Z=0$ 的对称性可知，

$p(0<Z<1.25)=\frac{1}{2}\times P(-1.25<z<1.25)$，按临界值 $Z=1.25$ 查表得，

$p(-1.25<Z<1.25)=0.788\ 7$。所以：

$$p(0<Z<1.25)=\frac{1}{2}\times 0.788\ 7=0.394\ 4$$

(3)由对称性与面积的分割关系可知：

$$p(1<Z<1.25)=\frac{1}{2}\times[P(\ \ 1.25<z<1.25)-p(-1<Z<1)]$$

$$=\frac{1}{2}\times(0.788\ 7-0.682\ 7)=0.053$$

(4)由于标准正态分布密度函数曲线下的面积为 1，所以：

$$p(Z>1)=\frac{1}{2}\times[1-p(-1<Z<1)]=\frac{1}{2}\times[1-0.682\ 7]=0.158\ 65$$

可以验证，若随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，则随机变量 $Z=\frac{X-\mu}{\sigma}$ 服从标准正态分布，即 $Z\sim N(0,1)$。

$Z=(X-\mu)/\sigma$ 是一个非常重要的变换，称为标准化变换。我们可以利用这个变换来计算非标准正态分布的概率，计算出随机变量取值落于任何区间的概率。

【例 4－5】假定学生某门学科的考试成绩服从均值为 60 分、标准差为 12 分的正态分布。那么某一学生的成绩在 60 分到 75 分之间的概率应为多少？

解：

设 X 表示学生的成绩，要计算的概率是 $p(60<X<75)$。首先对 X 进行标准化：

$$Z=(X-60)/12$$

将计算 X 的概率转化为计算 Z 的概率：

$$p(60<X<75)=p\left(\frac{60-60}{12}<\frac{X-60}{12}<\frac{75-60}{12}\right)=P(0<Z<1.25)=0.394\ 4$$

【例 4－6】设随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，试分别求 X 落在以 μ 为中心，以 σ，1.96σ，2σ 和 3σ 为半径的区间内的概率。

解：

作变换 $Z=\frac{X-\mu}{\sigma}$，则 Z 服从标准正态分布。所求概率分别为：

$$p(\mu-\sigma<X<\mu+\sigma)=p\left(\frac{\mu-\sigma-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{\mu+\sigma-\mu}{\sigma}\right)$$

$$=P(-1<Z<1)=68.27\%$$

$$p(\mu-1.96\sigma<X<\mu+1.96\sigma)=p\left(\frac{\mu-1.96\sigma-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{\mu+1.969\sigma-\mu}{\sigma}\right)$$

$$= P(-1.96 < Z < 1.96) = 95\%$$

$$p(\mu - 2\sigma < X < \mu + 2\sigma) = p\left(\frac{\mu - 2\sigma - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{\mu + 2\sigma - \mu}{\sigma}\right)$$

$$= P(-2 < Z < 2) = 95.45\%$$

$$p(\mu - 3\sigma < X < \mu + 3\sigma) = p\left(\frac{\mu - 3\sigma - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{\mu + 3\sigma - \mu}{\sigma}\right)$$

$$= P(-3 < Z < 3) = 99.73\%$$

从计算结果可知：随机变量 X 落在以 μ 为中心，以 3σ 为半径的区间外的概率只有 $1-99.73\% = 0.27\%$，是一个非常小的概率。一般可以认为 X 不会落到以 μ 为中心，以 3σ 为半径的区间之外，这就是所谓的"3σ 原则"。

除正态分布外，常用的连续型随机变量的分布还有卡方分布（或称 χ^2-分布）、t-分布、F-分布。

1. χ^2-分布。设 $X_1, X_2, \cdots, X_n$ 是相互独立，且服从标准正态分布的随机变量，则称随机变量 $\chi^2 = \sum_{i=1}^{n} X_i^2$ 所服从的分布为自由度为 n 的 χ^2-分布。

2. t-分布。t-分布也称为学生（Student）分布。设 X 服从标准正态分布，Y 服从自由度为 n 的卡方分布，且它们相互独立，则随机变量 $T = X/(\sqrt{Y/n})$ 所服从的分布是自由度为 n 的 t-分布。t-分布的分布曲线与正态分布曲线相似，数学上可以证明，当自由度 n 充分大时，t-分布与标准正态分布很近似。一般当 n 大于或等于 30 时，t-分布与标准正态分布的差别已非常小，可用标准正态分布代替它。

3. F-分布。设 X 和 Y 是相互独立的 χ^2-分布，自由度分别是 m 和 n，则称随机变量 $F = (X/Y)\cdot(n/m)$ 所服从的分布为 F-分布，(m, n) 称为它的自由度。该分布在后面的方差分析中将用到。

第二节　抽样分布

一、抽样的基本概念

抽样推断涉及的基本概念有：总体与样本、样本容量与样本个数、总体参数与样本统计量、重复抽样与不重复抽样。关于总体与样本的概念，我们已在第一章作了介绍，这里不再重复，只介绍样本容量与样本个数、总体参数与样本统计量、重复抽样与不重复抽样三对概念。

(一)样本容量与样本个数

1.样本容量。样本是从总体中抽出的部分单位的集合,这个集合的大小称为样本容量,一般用 n 表示,它表明一个样本中所包含的单位数。样本容量大,样本误差会小,但调查费用必增加,反之,样本容量过小,又将导致抽样误差增大,甚至失去抽样推断的价值。因此,在抽样设计中应根据调查目的认真考虑合适的样本容量。一般地,样本单位数大于30个的样本称为大样本,不超过30个的样本称为小样本。

2.样本个数。样本个数又称样本可能数目,它是指从一个总体中可能抽取多少个样本。样本个数的多少与抽样方法有关。关于样本个数的计算我们将在"重复抽样与不重复抽样"中介绍。

(二)总体参数与样本统计量

1.总体参数。总体分布的数量特征就是总体的参数,也是抽样统计推断的对象。常见的总体参数有:总体的平均数指标,总体成数(比重)指标,总体分布的方差、标准差,等等。它们都是反映总体分布特征的重要指标。平均数反映总体分布的集中趋势,方差或标准差反映总体分布的离中趋势。总体成数(也称总体比率)指标是指总体中具有某性质的单位数目在总体中所占的比重,它反映了总体的结构特征。这些指标的计算前面第三章已有介绍,这里不再重复。

2.样本统计量。与总体参数相对应的是样本的统计量。由于样本是从总体中随机地抽出来的,而样本统计量是样本的一个函数,因此,它们是随机变量。我们利用统计量来估计和推断总体的有关参数。与总体参数相对应,常见的统计量有:

$$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n}\left(=\frac{\sum Xf}{\sum f}\right) \tag{4.13}$$

$$P=\frac{n_1}{n} \tag{4.14}$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\left(=\frac{\sum(X-\bar{X})^2 f}{\sum f-1}\right) \tag{4.15}$$

$$S=\sqrt{S^2} \tag{4.16}$$

以上式中,$\bar{X}$ 是样本平均数,P 是样本成数,S^2 与 S 分别表示样本的方差与标准差。n 是样本容量,n_1 是样本中具有某种特征的单位数目,f 是在分组样本资料下的权数。

应当指出,尽管样本统计量的计算公式与总体参数的计算公式形式上十分类似,但两者有重要区别:总体参数是常数,计算总体参数公式中所用到的总体各单位的标志值是确定的具体数值,用小写字母 x 表示。样本统计量是随机变量;计算样本统计量的公式中所用的样本单位标志值在未具体观察前是不确定的随机变量,用大写字母 X 表

示;计算总体参数公式中所使用的单位数包括总体的所有单位,通常用大写字母 N 表示;计算样本统计量所使用的单位数只是样本所包含的单位数,即单位容量,用小写字母 n 表示;计算总体参数公式中所使用的权数 F 是整个总体分组资料下的权数,计算样本统计量所使用的权数 f 是样本分组资料下的权数。

(三)重复抽样与不重复抽样

简单抽样的抽样方法有重复抽样与不重复抽样。

1.重复抽样。重复抽样(或重置抽样)是指从总体中抽出一个样本单位,记录其标志值后,又将其放回总体中继续参加下一轮样本单位的抽取。重复抽样的特点是:第一,n 个单位的样本是由 n 次试验的结果构成的。第二,每次试验是独立的,即其试验的结果与前次、后次的结果无关。第三,每次试验是在相同条件下进行的,每个单位在多次试验中选中的机会(概率)是相同的。在重复试验中,样本可能的个数是 N^n,N 为总体单位数,n 为样本容量。

2.不重复抽样。不重复抽样亦称为不重置抽样,即每次从总体中抽取一个单位,登记后不放回原总体,不参加下一轮抽样。下一次继续从总体余下的单位中抽取样本。其特点是:第一,n 个单位的样本由 n 次试验结果构成,但由于每次抽出不重复,所以实质上相当于从总体中同时抽取 n 个样本单位。第二,每次试验结果不是独立的,上次中选情况影响下次抽选结果。第三,每个单位在多次(轮)试验中中选的机会是不等的。不重复抽样,如果是考虑顺序,其样本可能个数为 $N!/(N-n)!$;如果不考虑顺序,其样本可能个数为 $N!/[(N-n)!\ n!]$。

二、抽样分布

(一)重复抽样分布

1.样本平均数的抽样分布。设从总体中抽出的样本为 $X_1,X_2,\cdots,X_n$,由于是重复抽样,每个 $X_i,(i=1,2,\cdots,n)$都是从总体中随机抽出的,都是与总体同分布的随机变量,并且是相互独立的。我们设总体的平均数为 μ,方差为 σ^2,则样本平均数的期望值与方差分别是:

$$E(\bar{X})=E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{1}{n}[E(X_1)+E(X_2)+\cdots+E(X_n)]=\mu$$

$$\sigma^2(\bar{X})=\sigma^2\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{1}{n^2}[\sigma^2(X_1)+\sigma^2(X_2)+\cdots+\sigma^2(X_n)]=\frac{\sigma^2}{n} \qquad (4.17)$$

从以上的式子我们知道,样本平均数分布的中心与总体的分布中心完全相同,方差是总体分布方差的 $1/n$。因此,样本平均数分布的集中趋势优于总体分布自身的集中趋势。由于样本平均数能"集中"分布于总体平均数附近,我们可以考虑用样本平均数

来估计总体的平均数。用样本统计量去估计总体参数难免有误差，样本变量的离散程度越大，产生误差的可能性也越大。我们用抽样平均数的标准差来反映抽样平均数与总体平均数的平均误差程度，称其为抽样平均误差，记为：

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{4.18}$$

抽样平均误差是总体标准差的 $1/\sqrt{n}$，通常比总体标准差小得多。

【例 4－7】某班组有 5 个工人，他们的单位工时工资分别是 4，6，8，10，12 元，现用重置抽样方式从 5 个工人中抽出 2 人，计算样本的平均工时工资的抽样平均误差。

解：

总体分布的平均数与方差分别是：

$$\mu=\frac{\sum x}{N}=\frac{4+6+8+10+12}{5}=8(\text{元})$$

$$\sigma^2=\frac{\sum(x-\mu)^2}{N}=\frac{(4-8)^2+(6-8)^2+(8-8)^2+(10-8)^2+(12-8)^2}{5}$$

$$=8$$

抽样平均误差 $$\sigma_{\bar{x}}=\frac{\sigma(X)}{\sqrt{n}}=\frac{\sqrt{8}}{\sqrt{2}}=2(\text{元})$$

2. 样本成数分布。总体成数 p 是指具有某种特征的单位在总体中的比重。在前面我们已经知道，成数是一个特殊平均数，设总体单位总数目是 N，总体中有该特征的单位数是 N_1。设 X 是 0，1 变量，即：总体单位有该特征，则 X 取 1，否则取 0，于是有：

$$p=\frac{N_1}{N}=\bar{X} \tag{4.19}$$

现从总体中抽出 n 个单位，如果其中有相应特征的单位数是 n_1，则样本成数是：

$$P=\frac{n_1}{n} \tag{4.20}$$

P 也是一个随机变量，利用样本平均数分布性质的结论，即有：

$$E(P)=p$$

$$\sigma(P)=\frac{\sigma(X)}{\sqrt{n}}=\frac{\sqrt{p(1-p)}}{\sqrt{n}}=\sqrt{\frac{p(1-p)}{n}}$$

【例 4－8】已知一批产品的合格率为 90%，现采用重置抽样方式从中取出 400 件，求样本合格率的抽样平均误差。

解：

$$\sigma(P)=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.9\times0.1}{400}}=1.5\%$$

由于样本容量大，样本成数的平均误差就大大减小。

(二)不重置抽样分布

1.样本平均数分布。仍以例4-7资料为例,在不重置抽样条件下,所有样本平均数如表4-4所示。从表4-4中可整理出样本平均数的分布如表4-5所示。

表4-4 样本工时平均工资 单位:元

样本变量	4	6	8	10	12
4	—	5	6	7	8
6	5	—	7	8	9
8	6	7	—	9	10
10	7	8	9	—	11
12	8	9	10	11	—

表4-5 样本工时平均工资分布

样本工时平均工资 $\bar{x}$(元)	频数	频率
5	2	2/20
6	2	2/20
7	4	4/20
8	4	4/20
9	4	4/20
10	2	2/20
11	2	2/20
合计	20	1

根据表4-5的分布数据,可计算样本平均工资与其标准差:

$$E(\bar{X})=\frac{\sum\bar{x}f}{\sum f}=\frac{1}{20}(5\times2+6\times2+7\times4+8\times4+9\times4+10\times2+11\times2)$$
$$=8(\text{元})$$

$$\sigma^2(\bar{X})=\frac{\sum[\bar{x}-E(\bar{x})]^2f}{\sum f}=\frac{1}{20}[(5-8)^2\times2+(6-8)^2\times2+(7-8)^2\times4$$
$$+(8-8)^2\times4+(9-8)^2\times4+(10-8)^2\times2+(11-8)^2\times2]$$
$$=3$$

$$\sigma(\bar{X})=\sqrt{3}(\text{元})$$

计算结果表明:在不重置抽样条件下,样本平均数分布的中心还是总体的中心;而抽样平均误差比重置抽样要小。

不重置抽样的平均误差比重置抽样小的原因是很直观的:不重置抽样排除了"每次抽出的都是极端值"的可能,这显然对降低抽样误差有利。数学上可以证明,在不重置

抽样条件下，

$$\sigma_x=\sqrt{\frac{\sigma^2(X)}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.21}$$

不重置抽样与重置抽样比，多了一个系数$\sqrt{\frac{N-n}{N-1}}\approx\sqrt{1-\frac{n}{N}}$，这个系数称为不重置抽样的修正系数。由于该系数在0,1之间，因此，不重置抽样平均误差比重置抽样小。当N远大于n时，修正系数近似1，修正与否对平均误差几乎没有影响，这时可以不考虑抽样方式差异，都按重置抽样处理。

2.样本成数抽样分布。由于样本成数是一个特殊的平均数，不重置抽样也有如下性质：

$$E(P)=p \tag{4.22}$$

$$\sigma(P)=\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \tag{4.23}$$

三、大数定理与中心极限定理

（一）大数定理

大数定理又称作大数法则。人们在观察个别事物时，是连同一切个别的特性来观察的。个别现象受偶然因素影响，有各自不同的表现。但是，对总体大量观察后再进行平均，就能使偶然因素的影响相互抵消，消除由个别偶然因素引起的极端性影响，从而使总体平均数稳定下来，反映出事物变化的一般规律，这就是大数定理的意义。

大数定理：独立同分布的随机变量$X_1,X_2,\cdots X_n,\cdots$，设它们的平均数为$\mu$，方差为$\sigma^2$。则对任意的正数$\varepsilon$，有：

$$\lim_{n\to\infty}p\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1 \tag{4.24}$$

该定理说明，当n充分大时，独立同分布的一系列随机变量，其平均数与它们共同的期望值之间的偏差，可以有很大的把握被控制在任意给定的范围之内。由于从总体中抽出的样本是独立且与总体同分布的，因此，当样本容量n充分大时，样本平均与总体平均之间的误差可以有很大的把握被控制在任意给定的范围之内，这就是人们用样本平均估计总体平均的理论根据。

由于成数指标是一个特殊的平均数，大数定理对成数指标自然也成立：设m是n次试验中事件A发生的次数，p是事件A发生的概率，则对于任意小的正数ε，有

$$\lim_{n\to\infty}p\left\{\left|\frac{m}{n}-p\right|<\varepsilon\right\}=1 \tag{4.25}$$

即当n充分大时，事件A发生的频率接近（依概率收敛于）事件A发生的概率，反

映了频率在大量重复试验过程中的稳定性。该定理称为贝努里大数定理,它提供了用频率代替概率的理论根据。

(二)中心极限定理

1.正态分布的再生定理。相互独立的两个正态随机变量相加之和仍服从正态分布,这就是正态分布的再生性。因此,从服从正态分布的总体中抽出一个容量是 n 的样本,则样本平均数 $\bar{X}$ 也服从正态分布。如果总体的平均数是 μ,标准差是 $\sigma(X)$,则样本平均数所服从的正态分布的中心仍是 μ,标准差是抽样平均误差 $\sigma_{\bar{x}}$。

2.中心极限定理。从正态分布的再生定理,可推出的样本平均数 $\bar{X}$ 服从正态分布 $N(\mu,\sigma_{\bar{x}}^2)$,前提条件是总体服从正态分布。在客观实际中,总体服从正态分布的条件不会总是成立的。在非正态总体的场合,样本平均数服从什么分布,当只有我们了解了中心极限定理后,这个问题才能明确。

中心极限定理:随机变量 $X_1,X_2\cdots X_n,\cdots$ 相互独立,且服从同一分布,该分布存在有限的期望和方差:$E(X_i)=\mu,\sigma^2(X_i)=\sigma^2,(i=1,2,\cdots)$。则当 n 趋于无穷大时,算术平均数 $\bar{X}=(\sum_{i=1}^{n}X_i)/n$,近似服从正态分布,即:

$$\bar{X}\sim N(\mu,\sigma^2/n)$$

从上述定理可以得出结论:无论总体服从何种分布,只要它的期望值与方差存在,我们就可以通过增大样本容量 n 的方式,保证样本平均数 $\bar{X}$ 近似服从正态分布。也就是说,大样本的平均数近似服从正态分布。

对于成数指标,我们设总体成数是 p,样本成数是 P,则当样本容量充分大时,P 近似服从 $N(p,\sigma_p^2)$。

第三节 总体参数估计

一、总体参数估计概述

总体参数估计就是以样本统计量来估计总体参数。参数估计应满足以下两个要求:一是估计的精度要求,二是可靠性要求。所谓精度就是估计误差的最大范围,即误差的最大值,可通过极限误差来反映;所谓可靠性是指估计结果正确的概率大小。

设待估计的总体参数是 θ,用以估计该参数的统计量是 $\hat{\theta}$ 抽样估计的极限误差是 Δ,即 $|\hat{\theta}-|\leqslant\Delta$。极限误差是根据研究对象的变异程度和分析任务的性质来确定的允许误差范围。显然,Δ 越小,估计的精度要求越高,Δ 越大,估计的精度要求越低。极

限误差的确定要以实际需要为基本标准。比如,对航天元器件的估计误差,就要求控制在极小的范围内;而对一些小商品,如纽扣的合格率估计,其估计误差就可以控制在较大的范围里,因为这种误差,对消费者、对厂商的负面影响都有限。

可靠性是抽样估计本身正确性的一个概率保证,通常称为估计的置信度。我们知道,对于连续型随机变量,它在一个点上取值的概率为零,因此,对服从连续型分布的抽样统计量,直接用它去估计总体参数值很难说是可靠的。用统计量估计总体参数值,称为点估计。点估计完全正确的概率通常为0。因此,我们更多的是考虑用样本统计量去估计总体参数的范围,这就是区间估计。例如,通过抽样估计某班学生某课程平均成绩的范围,这时我们要考虑这个估计正确的概率大小问题。一种极端估计是:平均成绩在0与100分之间。显然这个范围的估计,正确的概率很大,100%的正确,但这个估计无精度可言。如提高精度,估计平均成绩在70分到80分之间,这时估计正确的把握性肯定低于1。可见估计中精度要求与可靠性要求是一对矛盾。

二、点估计

(一)点估计的定义

点估计就是根据总体参数与样本统计量之间的内在联系,直接以样本统计量作为相应总体参数的估计量,点估计又称为定值估计。在统计中经常使用的点估计量有:

$$\begin{cases}\hat{\mu}=\bar{X}\\ \hat{p}=P\\ \hat{\sigma}^2=s^2=\dfrac{\sum(X-\bar{X})^2}{n-1}\end{cases}\tag{4.26}$$

上式中 $\hat{\mu},\hat{p},\hat{\sigma}^2$ 分别表示总体平均数、总体成数与总体方差的估计量。

【例4-9】对某企业的产品进行抽样检验,设抽出100件产品,其中不合格产品5件,试估计该企业产品的合格率是多少?

我们可以通过样本的合格率来估计企业产品的合格率。样本合格率 $p=95/100=95\%$,我们估计该企业产品的合格率是95%。

(二)估计的评价标准

点估计的优点是,直接给出了总体参数的估计值。不足之处是不能提供估计误差的信息。样本统计量是一个随机变量,从一次抽样的结果来判断一个统计量的优劣是不确切的,必须通过多次试验或从抽样分布的特点出发,才能判断这个估计量是否为理想的估计量。

判断点估计优良性包括三条标准:无偏性、有效性和一致性。

1.无偏性。用 θ 表示总体的待估计参数,$\hat{\theta}$ 是估计 θ 的样本统计量,我们说 $\hat{\theta}$ 是 θ

的无偏估计，指的是 θ 满足：

$$E(\hat{\theta})=\theta \tag{4.27}$$

无偏性要求用来估计总体参数的样本统计量，其分布是以总体参数真值为中心的，在一次具体的抽样估计中，估计量或者大于总体参数，或者小于总体参数；但是，在进行重复抽样估计的过程中，所有估计量的平均数应该等于待估的总体参数。这说明，无偏估计要求估计量没有系统偏差。

由于 $E(X)=\mu$，所以样本平均是总体平均的一个无偏估计。有了无偏性标准，我们可以解释为什么对总体方差进行估计时，样本方差公式所除的不是样本容量 n，而是 $n-1$。证明如下：

$$\begin{aligned}
E(S^2)&=E[\frac{1}{n-1}\sum(X_i-\bar{X})^2]\\
&=\frac{1}{n-1}E\sum[(X_i-\mu)-(\bar{X}-\mu)]^2\\
&=\frac{1}{n-1}E[\sum(X_i-\mu)^2-n(X-\mu)^2]\\
&=\frac{1}{n-1}[\sum E(X_i-\mu)^2-nE(\bar{X}-\mu)^2]\\
&=\frac{1}{n-1}\left[n\sigma^2-n\frac{\sigma^2}{n}\right]\\
&=\sigma^2
\end{aligned}$$

这说明 $S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ 是总体方差的无偏估计。当样本容量很大时，$1/n$ 与 $1/(n-1)$ 相差不大，样本方差的公式，可以直接除以 n，与总体方差计算公式保持一致。

2. 有效性。$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是总体参数 θ 的无偏估计量，如果 $\sigma^2(\hat{\theta}_1)<\sigma^2(\hat{\theta}_2)$，则说明估计量 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。设总体的方差是 σ^2，我们有：$\sigma^2(X_i)=\sigma^2$，$\sigma^2(\bar{X})=\sigma^2/n$。显然，样本平均数的方差比样本中某个单位的标志值的方差要小，只是其方差的 $1/n$，所以作为估计量，样本平均数更加有效。

3. 一致性。一致性是指随着样本容量不断增大，样本统计量接近总体参数的可能性就越来越大，或者，对于任意给定的偏差控制水平，两者间偏差高于此控制水平的可能性越来越小，接近于0。用公式表示就是：

$$\lim_{n\to\infty}p\{|\hat{\theta}-\theta|<\varepsilon\}=1 \tag{4.28}$$

公式中，ε 为一任意小的数。上式说明，当 n 充分大时，$\hat{\theta}$ 与 θ 之间的偏差，可以有很大的把握被控制在任意给定的范围之内。当 n 趋于无穷大时，估计量 $\hat{\theta}$ 依概率收敛于 θ。

由大数定理我们知道,抽样平均数是总体平均数的一致估计。因此,样本平均是总体平均的一个无偏、有效且满足一致性要求的估计量;因为成数是一个特殊的平均数,该结论对成数估计也成立。

三、区间估计

(一)区间估计的含义

点估计给出总体参数的具体估计值,但这个估计值误差有多大?可靠性如何?这些问题点估计都不能回答。区间估计则弥补了点估计这些方面的不足。

所谓区间估计,就是估计总体参数的区间范围,并要求给出区间估计成立的概率值。设 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是两个统计量($\hat{\theta}_1<\hat{\theta}_2$),分别作为总体参数 θ 区间估计的下限与上限,则要求:

$$P(\hat{\theta}_1\leqslant\theta\leqslant\hat{\theta}_2)=1-\alpha \tag{4.29}$$

式中,$\alpha(0<\alpha<1)$是区间估计的显著性水平,其取值大小由实际问题确定,经常取1%、5%和10%;$1-\alpha$ 称为置信度。区间估计的特点是:给出总体参数的一个估计区间,总体参数恰好在这个区间内的概率不要求达到1,可放低要求,减去一个小概率的显著性水平,即达到 $1-\alpha$ 就行了。

由于 θ 作为总体参数,是固定不变的常数,它或者在给出的区间内或者在给出的区间外,(4.29)式的概率或者是0或者是1,不可能是 $1-\alpha$。怎么来理解这个概率?由于 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是统计量,$\hat{\theta}_1\leqslant\theta\leqslant\hat{\theta}_2$ 表示用区间$(\hat{\theta}_1,\hat{\theta}_2)$去框 θ,虽然每次结果只能是:此区间或者框住了 θ,或者没有框住 θ,估计结论或者正确或者错误,但是如果多次重复估计的话,则平均100次估计中,只有 100α 次估计是错误的,有 $100(1-\alpha)$次估计是正确的。

(二)平均数的区间估计

对平均数的区间估计,分两种情况:

1.总体方差 σ^2 已知。由抽样分布定理知道,如果总体服从正态分布,则样本平均数 $\bar{X}\sim N(\mu,\sigma_{\bar{x}}^2)$;如果总体正态性不成立,但是样本容量 n 充分大时,近似地也有 $\bar{X}\sim N(\mu,\sigma_{\bar{x}}^2)$。因此,

$$z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\sim N(0,1)$$

这样,对于给出的显著性水平 α,通过(反)查标准正态分布表可得到临界值 $z_{\alpha/2}$,满足:

$$P(-z_{\alpha/2}\leqslant z\leqslant z_{\alpha/2})=1-\alpha$$

注意到:

$$-z_{\alpha/2}\leqslant z\leqslant z_{\alpha/2}$$

$$\Leftrightarrow -z_{\alpha/2}\leqslant\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\leqslant z_{\alpha/2}$$

$$\Leftrightarrow -z_{\alpha/2}\sigma_{\bar{x}}\leqslant\bar{X}-\mu\leqslant z_{\alpha/2}\sigma_{\bar{x}}$$

$$\Leftrightarrow\bar{X}-z_{\alpha/2}\sigma_{\bar{x}}\leqslant\mu\leqslant\bar{X}+z_{\alpha/2}\sigma_{\bar{x}}$$

因此,总体平均数在显著性水平是 α 时的区间估计为:

$$\bar{X}\pm z_{\alpha/2}\sigma_{\bar{x}} \tag{4.30}$$

反过来,如果平均数抽样的极限误差是 Δ,即:$|\bar{X}-\mu|\leqslant\Delta$,等价于一个区间估计:$\bar{x}-\Delta\leqslant\bar{X}\leqslant\bar{x}+\Delta$,这个区间估计的置信度是多少呢?

$$P(\bar{X}-\Delta\leqslant\mu\leqslant\bar{X}+\Delta)=P(|\bar{X}-\mu|\leqslant\Delta)$$

$$=P\left(\left|\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}\right|\leqslant\frac{\Delta}{\sigma_{\bar{x}}}\right)=P(|z|\leqslant\frac{\Delta}{\sigma_{\bar{x}}})$$

通过临界值 $z_{\alpha/2}=\frac{\Delta}{\sigma_{\bar{x}}}$,去查标准正态分布,可得出上式的置信度是 $1-\alpha$。

临界值 $z_{\alpha/2}$与置信度 $1-\alpha$,两者间是密切相关的,通过查标准正态表,可互相确定。如果置信度提高,区间估计的概率增大,α 就要相应地减小,临界值 $z_{\alpha/2}$增大;反之,区间估计的概率减小,临界值也减小。因此,也称临界值 $z_{\alpha/2}$为概率度,简记为 z,用来间接衡量区间估计的概率大小。

关于极限误差、抽样平均误差、概率度三者间的关系,有如下结果:

$$z=\frac{\Delta}{\sigma_{\bar{x}}}$$

或者

$$\Delta=\sigma_{\bar{x}}\cdot z \tag{4.31}$$

以上关系式不仅在对平均数进行区间估计时有效,在对成数指标估计时也适用。

对于给出的置信水平 $1-\alpha$,得到置信区间的公式:

重复抽样时,

$$\bar{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{4.32}$$

不重复抽样时,

$$\bar{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.33}$$

【例 4-10】某地区的电视台委托调查公司估计地区内居民平均每日的看电视时间。调查公司随机抽取了 100 名居民进行调查,样本数据显示平均每人每天看电视时间是 4 个小时。如果已知总体的标准差 $\sigma=1.5$ 小时。试求:

1.该地区内居民每天看电视的平均时间的置信区间(置信度是95%);

2.如果要求估计的误差不超过27分钟,这时置信度是多少?

解:

已知 $\bar{X}=4$ 小时,$n=100$,$\sigma=1.5$ 小时,$1-\alpha=95\%$,$\alpha=5\%$。这时查标准正态分布表,可得临界值:

$$z_{\alpha/2}=z_{0.025}=1.96$$

由于样本容量在地区居民人数中所占的比重太小,重复与不重复抽样效果相差不大,按重复抽样计算,区间估计是:

$$\bar{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=4\pm1.96\times\frac{1.5}{\sqrt{100}}=4\pm0.29$$

因此,可以95%的置信度,估计该地区内居民每天看电视的平均时间在3.71到4.29个小时之间。

要求极限误差等于27分钟,即 $\Delta=0.45$ 小时。这时概率度:

$$z=\frac{\Delta}{\sigma/\sqrt{n}}=\frac{0.45}{1.5/\sqrt{100}}=3$$

查概率表知这时的概率是99.73%,区间估计是:

$(\bar{x}-\Delta,\bar{x}+\Delta)=(4-0.45,4+0.45)=(3.55,4.45)$。

因此结论是:可用99.73%的概率保证,估计该地区内居民每天看电视的平均时间在3.55到4.45个小时之间。

2.总体方差 σ^2 未知。当总体服从正态分布但方差未知时,可用样本的标准差S代替总体标准差。这时统计量是:$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$

t 服从的分布不是标准正态分布,而是自由度为 $n-1$ 的 t-分布。因此,总体均值的区间估计是:

重复抽样时,

$$\bar{X}\pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \tag{4.34}$$

不重复抽样时,

$$\bar{X}\pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.35}$$

以上区间估计公式中,不仅总体标准差 σ 用样本标准差 S 代替,而且临界值 $t_{\alpha/2,n-1}$,要查 t-分布表(自由度为 $n-1$)得到。但是,在大样本场合,t-分布与标准正态分布非常接近,可直接从标准正态分布表中查临界值。

【例4-11】麦当劳餐馆在7星期内抽查49位顾客的消费额(元)如下,求在概率

90%的保证下,顾客平均消费额的估计区间。

15	24	38	26	30	42	18
30	25	26	34	44	20	35
24	26	34	48	18	28	46
19	30	36	42	24	32	45
36	21	47	26	28	31	42
45	36	24	28	27	32	36
47	53	22	24	32	46	26

解:

第一步:通过 Excel 进行统计计算(具体计算过程参见本章第五节)可得到:

$$\bar{X}=\frac{\sum x}{n}=32, S=\sqrt{\frac{\sum(x-x)^2}{n-1}}=9.45, \sigma_{\bar{x}}=\frac{S}{\sqrt{n}}=\frac{9.45}{\sqrt{49}}=1.35$$

点估计:麦当劳餐馆总体顾客平均消费额为 32 元。

第二步:根据给定的置信度 $F(z)=90\%$,查概率表得 $z=1.64$。

第三步:计算 $\Delta=z\cdot\sigma_{\bar{x}}=1.64\times1.35=2.2$ 元。据此估计,

总体平均消费额下限 $=\bar{x}-\Delta=32-2.2=29.8$ 元,

总体平均消费额上限 $=\bar{x}+\Delta=32+2.2=34.2$ 元。

区间估计:以 90% 的概率保证,麦当劳餐馆顾客消费额在 29.8~34.2 元之间。

(三)成数指标的区间估计

成数指标是一个特殊的平均数。所以,类似于总体平均数的区间估计,总体成数的区间估计是:

$$P\pm z_{\alpha/2}\sigma_p \tag{4.36}$$

式中的成数抽样平均误差在重复抽样条件下是:

$$\sigma_p=\sqrt{\frac{P(1-P)}{n}}$$

在不重复抽样的条件下是:

$$\sigma_p=\sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

注意:在实践中,由于总体成数常常未知,这时,上述抽样平均误差公式中的总成数用样本成数代替。

【例 4-12】某工厂要估计一批总数 5 000 件的产品的废品率,于是随机抽出 400 件产品进行检测,发现有 32 件废品。试给出该批产品的废品率的区间估计(置信度是 90%)。

解：

$n=400, N=5\ 000$，样本废品率 $P=32/400=8\%$。置信度 $1-\alpha=90\%$，$\alpha=10\%$，$\alpha/2=5\%$。查标准正态分布表得：

$$z_{\alpha/2}=z_{0.05}=1.645$$

因此，这批产品废品率的区间估计是：

$$\begin{aligned}&P\pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}\\&=0.08\pm1.645\times\sqrt{\frac{0.08\times(1-0.08)}{400}}\times\sqrt{\frac{5\ 000-400}{5\ 000-1}}\\&=0.08\pm0.021\end{aligned}$$

即这批产品的废品率在 5.9% 与 10.1% 之间。

四、样本容量的确定

在前面我们已经知道，极限误差、概率度与抽样平均误差三者间的数量关系是：$\Delta=z\cdot\sigma_{\bar{x}}$。当抽样平均误差保持不变时，极限误差与概率度两者间关系是：Δ 增大，z 也增大了，Δ 减小，z 也减小了。因此，抽样估计的精度与可靠性之间存在矛盾，要提高精度（Δ 减小），需以牺牲概率度（z 减小）为代价；要提高概率度（z 增大），又要以牺牲估计精度（Δ 增大）为代价。在 $\sigma_{\bar{x}}$ 不变的情况下，这对矛盾是不可调和的；但是，降低抽样平均误差后，就可以同时提高估计的精度与概率度。比如，通过增加样本容量 n 来达到降低抽样平均误差的目标。这时应该考虑，样本容量 n 究竟取多大合适？这就是样本容量的确定问题。

（一）估计总体均值时样本容量的确定

1. 总体方差已知，重复抽样，这时有：

$$\Delta=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

上式两边平方整理后可得：

$$n=\frac{z_{\alpha/2}^2\sigma^2}{\Delta^2}\tag{4.37}$$

这就是在给定极限误差、概率度要求下，至少应抽取的样本容量。

2. 总体方差已知，不重复抽样，这时有：

$$\Delta=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

上式两边平方并进行整理得：

$$n=\frac{Nz_{\alpha/2}^2\sigma^2}{N\Delta^2+z_{\alpha/2}^2\sigma^2}\tag{4.38}$$

(二)估计成数时样本容量的确定

1.重复抽样

$$n=\frac{z_{\alpha/2}^{2}P(1-P)}{\Delta_P^2} \tag{4.39}$$

2.不重复抽样

$$n=\frac{Nz_{\alpha/2}^{2}p(1-p)}{N\Delta_P^2+z_{\alpha/2}^{2}p(1-p)} \tag{4.40}$$

(三)使用上述公式时应注意的问题

1.计算样本容量时,一般总体的方差与成数都是未知的,可用有关资料替代:一是用历史资料已有的方差与成数代替;二是在进行正式抽样调查前进行几次试验性调查,用试验中方差的最大值代替总体方差;三是成数方差在完全缺乏资料的情况下,就用成数方差的最大值0.25代替。

2.如果进行一次抽样调查,同时估计总体均值与成数,用上面的公式同时计算出两个样本容量,可取一个最大的结果,同时满足两方面的需要。

3.上面的公式计算结果如果带小数,这时样本容量不按四舍五入法则取整数,取比这个数大的最小整数代替。例如计算得到:$n=56.03$,那么,样本容量取57,而不是56。

【例4-13】对企业产品合格率进行抽样调查,根据历史上进行的二次调查资料,合格率分别是15%和13%,这次调查要求误差不超过5%,概率保证程度为95%,问至少要抽出多少产品作为样本?

解:

已知$\alpha=5\%$,$\Delta_P=0.05$,$z_{\alpha/2}=1.96$。按历史上的两次调查资料,分别计算成数方差为:$0.15\times(1-0.15)=0.127\ 5$,和$0.13\times(1-0.13)=0.113\ 1$。取方差最大者,因此选$P=15\%$。由于企业产品数量一般都较大,抽出样本在总体中所占的比重很小,无论是重复抽样还是不重复抽样,结果相差不大,可按重复抽样方式计算,所以至少应抽取的样本容量是:

$$n=\frac{z_{\alpha/2}^{2}P(1-P)}{\Delta_P^2}=\frac{1.96^2\times0.15\times(1-0.15)}{0.05^2}=195.922(\text{件})$$

应抽取196件产品进行检验。

【例4-14】对某型号电池进行电流强度检验,根据以往正常生产的经验数据,已知电流强度的标准差$\sigma=0.4$安培,合格率$P=90\%$。采用随机重复抽样方式,需要在99.73%的概率保证下,抽样平均电流的误差范围不超过0.08安培,抽样合格率误差范围不超过5%,试求必要的抽样单位数。

解：

已知，$1-\alpha=99.73\%$，$z_{\alpha/2}=3$，按抽样平均数与成数计算的样本容量分别是：

$$n_1=\frac{z_{\alpha/2}{}^2\sigma^2}{\Delta_{\bar{x}}^2}=\frac{3^2\times0.4^2}{0.08^2}=225(个)$$

$$n_2=\frac{z_{\alpha/2}{}^2P(1-P)}{\Delta_P^2}=\frac{3^2\times0.9\times0.1}{0.05^2}=324(个)$$

取以上计算结果中较大者，即 $n=324$，应抽取 324 个电池作样本以保证抽样调查的准确性。

第四节 抽样设计

抽样估计效果好坏，关键是抽样平均误差的控制。抽样平均误差小，抽样效果从整体上看就是理想的；否则，抽样效果就不理想。从前面的分析知道，抽样平均误差受以下几方面因素的影响：一是总体的变异性，即与总体的标准差大小有关；二是样本容量；三是抽样方法。还有一个重要的因素，就是抽样的组织形式。抽样的组织形式有：简单随机抽样、类型抽样、等距抽样、整群抽样、阶段抽样等。不同抽样组织设计意味着对总体信息不同程度的利用，将需要不同的调查成本，它们之间的抽样效果也存在较大的差异。前面我们讨论的抽样组织形式，都是简单随机抽样。接下来我们要讨论其他几种抽样组织方式及其抽样平均误差。

一、类型抽样

类型抽样也称分层抽样，它是按一定标志对总体各单位进行分类，然后分别从每一类中按随机原则抽取一定的单位构成样本。类型抽样的前提是对总体的结构有一定的了解，为了充分利用这些信息，提高估计的精确性，对总体按确定标志进行分类，保证抽出的样本与总体尽可能保持相似的结构。例如，抽样调查一个城市的居民收入分配状况，如果历史资料反映了该城市居民的贫富结构：高收入者、中等收入者与低收入者的比例结构，我们可以按此结构分类，分别从高收入者、中等收入者与低收入者中按一定的比例抽取样本。这样就可以避免样本全来自某一收入阶层所产生的系统偏差。

设总体由 N 个单位组成，按对总体的认识，把总体分为 k 组，使得：

$$N=N_1+N_2+\cdots+N_k$$

然后相应从各组中分别按随机方式抽出 $n_1,n_2,\cdots,n_k$ 个单位组成样本。设样本

容量为 n，它满足：

$$n = n_1 + n_2 + \cdots + n_k$$

采用比例抽样方式，我们从每一类 N_i 抽取 n_i 时要求两者间保持合适的比例，也就是保持各组样本单位数与总体同组单位数之比，等于样本容量与总体单位数之比，即：

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n}{N} \tag{4.41}$$

所以各组的样本单位数应为：

$$n_i = \frac{nN_i}{N} \qquad (i = 1, 2, \cdots, k) \tag{4.42}$$

设样本是：$X_{11}, X_{12}, \cdots, X_{1n_1}, X_{21}, X_{22}, \cdots, X_{2n_2}, \cdots, X_{k1}, X_{k2}, \cdots, X_{kn_k}$，

样本抽出来后：

第一步：计算各组平均数

$$\bar{X}_i = \frac{\sum_j X_{ij}}{n_i} \quad (i = 1, 2, \cdots, k)$$

第二步：将各组平均数以各组单位数或各样本组单位数为权数，进行加权平均，求出样本平均数

$$\bar{X} = \frac{\sum_i N_i \bar{X}_i}{N} = \frac{\sum_i n_i X_i}{n} \tag{4.43}$$

第三步：计算类型抽样的抽样平均误差

$$\sigma_{\bar{x}}^2 = \frac{\sum_i n_i^2 \sigma_{\bar{x}_i}^2}{n^2} \tag{4.44}$$

其中，$\sigma_{\bar{x}_i}^2$ 是从第 i 组中抽样计算平均数的方差，如果每组中都按重复方式抽样，这个样本平均数的标准差就是抽样平均误差：

$$\sigma_{\bar{x}_i} = \frac{\sigma_i}{\sqrt{n_i}} \tag{4.45}$$

其中，$\sigma_i \approx \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n_i}}$ 是第 i 组的组内标准差。记 $\overline{\sigma_i^2} = \frac{\sum_i n_i \sigma_i^2}{n}$，代入(4.44)式，得到重复抽样条件下样本平均数的抽样平均误差为：

$$\sigma_x = \sqrt{\frac{\sum_i n_i^2 \cdot \frac{\sigma_i^2}{n_i}}{n^2}} = \sqrt{\frac{1}{n} \cdot \frac{\sum_i n_i \sigma_i^2}{n}} = \sqrt{\frac{\overline{\sigma_i^2}}{n}} \tag{4.46}$$

同样可计算在不重复抽样条件下的抽样平均误差为：

$$\sigma_x = \sqrt{\frac{\overline{\sigma_i^2}}{n} \cdot \left(1 - \frac{n}{N}\right)} \tag{4.47}$$

【例 4－15】假设某农场种植小麦 1 200 亩，根据其地理条件划分为甲、乙、丙三类，按 5% 的比例总共抽取 60 亩进行调查，结果如表 4－6 所示。试以 95% 的概率估计农场平均亩产量的区间范围。

表 4－6

麦田类别	总体面积(亩)N_i	抽查面积(亩)n_i	抽样平均亩产(公斤)	亩产标准差(公斤)
甲	600	30	300	10
乙	360	18	230	12.5
丙	240	12	200	18
合计	1 200	60	259	12.7

解：

样本平均数是：

$$\bar{x}=\frac{\sum_i n_i\bar{x}_i}{n}=\frac{30\times300+18\times230+12\times200}{60}=259(\text{公斤})$$

$$\overline{\sigma^2}=\frac{\sum_i n_i\sigma_i^2}{n}=\frac{30\times10^2+18\times12.5^2+12\times18^2}{60}=161.7(\text{公斤})$$

所以，类型抽样的平均误差是：

$$\sigma_{\bar{x}}=\sqrt{\frac{\overline{\sigma^2}}{n}}=\sqrt{\frac{161.7}{60}}=1.65(\text{公斤})$$

按题意，$\alpha=5\%$，$z_{\alpha/2}=1.96$，所以 $\Delta=z_{\alpha/2}\sigma_{\bar{x}}=1.96\times1.65=3.234$(公斤)。

估计区间为：

$$[\bar{x}-\Delta,\bar{x}+\Delta]=[255.76,262.25]$$

所以，估计总体平均亩产量在 255.76 公斤到 262.25 公斤之间。

二、整群抽样

整群抽样就是将总体各单位分成若干群，然后从其中随机抽取部分群，对中选的群进行全面调查的抽样组织方式。

在总体单位数很大时，如果直接从总体中抽取总体单位，有时是很困难的，比如从一个大城市中的所有大学生中抽样了解大学生的基本情况，这个城市的大学生人数有几十万之众，直接抽取样本单位有许多困难。如果按整群抽样，以班级为抽样单位，从全部学校的所有班级中抽出部分班级，调查抽中的班级，就方便多了。

设总体的全部 N 个单位被划分为 R 群，每群含有 M 个单位。现在从总体 R 群中随机抽出 r 群组成样本，对中选的群中的所有单位进行全面调查。群的平均数是：

$$\bar{X}_i=\frac{\sum_{j=1}^{M}X_{ij}}{M}\qquad(i=1,2,\cdots,r)\tag{4.48}$$

样本平均数是：

$$\bar{X}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{M}X_{ij}}{rM}=\frac{\sum_{i=1}^{r}\bar{X}_i}{r}\tag{4.49}$$

从上式可以看出，整群抽样实质上是以群代替总体单位，以群平均数代替总体单位标志值之后的简单随机抽样。群间方差是：

$$\delta^2=\frac{\sum_i(\mu_i-\mu)^2}{R}\tag{4.50}$$

或者由样本数据估计：

$$\delta^2=\frac{\sum_i(\bar{X}_i-\bar{X})^2}{r}\tag{4.51}$$

因此，样本平均数的抽样平均误差是：

$$\sigma_{\bar{x}}=\sqrt{\frac{\delta^2}{r}\left(\frac{R-r}{R-1}\right)}\tag{4.52}$$

上式中出现修正系数$\frac{R-r}{R-1}$，这是因为整群抽样都采用不重复抽样。

【例 4－16】假设某水泥厂大量连续生产 100 公斤装水泥，一昼夜产量为 14 400 袋，平均每分钟产量 10 袋。现每隔 144 分钟抽取一分钟的产量（10 袋为一群），一昼夜共抽取 100 袋水泥，观察结果如表 4－7 所示。

表 4－7　水泥生产情况表

样本群编号	各群每袋平均重量	与样本平均数的离差	离差平方
1	98	－3	9
2	102	1	1
3	104	3	9
4	106	5	25
5	100	－1	1
6	98	－3	9
7	100	－1	1
8	96	－5	25
9	100	－1	1
10	106	5	25
合计	1 010	——	106

试根据表中的数据，计算样本平均数的抽样平均误差，并以 95% 的概率估计每包水泥重量的区间范围。

解：

按题意，每分钟产量为一群，总体被分为 $R=24\times60=1\ 440$ 群，

$r=1\ 440/144=10$。样本平均 $\bar{X}=\frac{\sum_{i=1}^{r}\bar{X}_i}{r}=\frac{1\ 010}{10}=101$(公斤)

先求出组间方差：

$$\delta^2=\frac{\sum_i(\bar{X}_i-\bar{X})^2}{r}=\frac{106}{10}=10.6$$

所求的抽样平均误差：

$$\sigma_{\bar{x}}=\sqrt{\frac{\delta^2}{r}\left(\frac{R-r}{R-1}\right)}=\sqrt{\frac{10.6}{10}\left(1-\frac{10}{1\ 440}\right)}$$
$$=1.026(\text{公斤})$$

概率保证是 95%，即 $\alpha=5\%$，概率度 $z_{\alpha/2}=1.96$。

所以每包水泥的重量的置信区间是：

$$[\bar{X}\pm z_{\alpha/2}\sigma_{\bar{x}}]$$
$$=(101-1.96\times1.026,101+1.96\times1.026)$$
$$=(98.99,103.01)$$

即每包水泥的估计区间约在 99 公斤到 103 公斤之间。

三、等距抽样

等距抽样又称机械抽样或系统抽样，它是将总体各单位按某一标志进行排列，然后按固定的间隔来抽取样本单位的抽样组织形式。

根据需要抽取的样本单位数 n 和总体的单位数 N，可以计算出等距抽样的间隔大小：

$$k=N/n \tag{4.53}$$

先从排序后顺序是 $1,2,\cdots,k$ 的第一部分中随机抽出第 i 个单位，然后在顺序是 $k+1,k+2,\cdots,2k$ 的第二部分中取出第 $k+i$ 个单位，再从顺序是 $2k+1,2k+2,\cdots,3k$ 的第三部分抽取第 $2k+i$ 个单位，依次类推，最后从顺序是 $(n-1)k+1,(n-1)k+2,\cdots,nk$ 的第 n 部分抽取第 $(n-1)k+i$ 个单位，一共 n 个单位构成样本。

总体排序标志是由总体的有关辅助信息确定的，与调查标志两者间可以有关也可以无关。如家计调查，按门牌号码排序就是无关标志。但是，选择排序标志与实际调查标志间如果存在密切联系，要比无关标志排列的机械抽样更为优越。如农产量调查按

平均亩产量高低排序，职工家计调查按平均工资多少进行排序，都可缩小各单位间的差异程度，有利于提高样本的代表性。

等距抽样的间隔，应避免与现象本身的节奏性或循环周期相重合。例如，进行农作物产量调查时，抽样间隔就应避免与农作物垅长或间距相重合；进行工业产品质量调查时，产品抽样时间间隔不宜和上下班时间相一致，否则，就会引起系统偏差而影响样本的代表性。

用等距抽样方式抽取一个样本后，就可以计算样本平均数。关键是这个平均数的平均误差如何确定。一般说来，排序后总体被分成 n 个部分，每一部分包含 k 个单位，从中随机抽取一个单位，其余单位情况未知，每一部分中的方差不可计算，一般也没有历史资料可以替代它们。因此，直接计算等距抽样的平均误差是有困难的，只能以间接方式计算其近似值。如果用来排队的标志与所要研究的目的没有关系，第一个单位是随机抽取的，那么它的抽样误差就与简单随机抽样误差相接近。为了方便起见，可以采用简单随机抽样的平均误差代替等距抽样平均误差：

$$\sigma_x=\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} \tag{4.54}$$

等距抽样一般都是不重复抽样。总体方差 σ^2 不知，常用样本方差代替。

【例 4－17】某块麦地长 300 米，宽 120 米，包括 120 条垅，每垅长 300 米。现从这块麦地按等距抽样的方式，抽取 50 个 2 米长垅为样本单位进行实割实测。

样本距离为麦垅总长除以样本单位数，即 $300\times120/50=720$（米）。现从地角一边样本距离一半处抽取第一个样本单位，即从 360 米点前后各 1 米为第一个样本单位，以后每隔 720 米取一个样本单位，一直抽取出 50 个样本单位为止。实测各样本单位产量如表 4－8 所示。

表 4－8　各样本单位小麦产量

样本产量 X（公斤）	单位数 n	nX	$X-\bar{X}$	$(X-\bar{X})^2n$
0.8	6	4.8	－0.4	0.96
1.0	12	12.0	0.2	0.48
1.2	14	16.8	0	0
1.4	12	16.8	0.2	0.48
1.6	6	9.6	0.4	0.96
合计	50	60	—	2.88

试计算平均亩产量的抽样平均误差，并以 95% 的概率保证估计这块麦地的亩产量和总产量。

解：

样本平均产量 $\bar{X}=\frac{\sum_i n_i X_i}{n}=\frac{60}{50}=1.2$(公斤)

样本单位标准差 $S=\sqrt{\frac{\sum_i (X-\bar{X})^2 n_i}{n}}=\sqrt{\frac{2.88}{50}}=0.24$(公斤)

样本单位的抽样平均误差 $\sigma_x=\sqrt{\frac{S^2}{n}}=\sqrt{\frac{0.24^2}{50}}=0.034$(公斤)

这块麦地的面积是:300×120=36 000(平方米),折合为 360×0.15=54(亩)。

由于样本单位垅长是 2 米,所以每亩含样本单位数是:

1/2×总垅长÷面积=1/2×36 000/54,约等于 333 个。

平均亩产量=样本平均产量×每亩含样本单位数

平均亩产量是:1.2×18 000/54=400(公斤)

平均亩产量的抽样平均误差=每亩含样本单位数×样本单位的抽样平均误差=18 000/54×0.034=11.33(公斤)

由于概率保证是 95%,即 $\alpha=5\%$,所以概率度 $z_{\alpha/2}=1.96$。

亩产量的区间估计是:

$$[\bar{X}\pm z_{\alpha/2}\sigma_{\bar{x}}]=[400\pm1.96\times11.33]=[377.79,422.21]$$

即亩产量估计在 377.79 公斤到 422.21 公斤之间。

总产量的区间估计是:

$$[N\bar{X}\pm N\cdot z_{\alpha/2}\sigma_{\bar{x}}]=[54\times400\pm54\times1.96\times11.33]=[20\ 400.83,22\ 799.34]$$

即以 95% 的概率保证,这块麦地的总产量估计在 20 401 公斤到 22 799 公斤之间。

四、阶段抽样

如果总体的范围很大,有必要采用阶段抽样的组织形式。所谓阶段抽样,就是先从总体中抽出较大的范围的单位,再从选的大单位中抽较小范围的单位,依次类推,最后从更小的范围抽出样本基本单位。这种抽样方式在我国的农产量调查、职工家计调查中都很适用:先从全国抽出各个省,再从抽中的省中抽出县、市,最后抽出样本的基本单位等等。

为了简化问题,以下分析两阶段抽样平均误差的控制问题。两阶段抽样在组织技术上可以看成是整群抽样和类型抽样的结合。设总体分成 R 组,每组 M 个单位。两阶段抽样就是:第一阶段用整群抽样方式从总体的全部 R 组(群)中,随机抽取 r 组(群);第二阶段用类型抽样方式从每个中选组中抽出 m 个样本单位。

样本平均数可以这样计算:

先计算第 i 组的样本平均数 $\bar{X}_i$

$$\bar{X}_i = \frac{\sum_{j=1}^{m} X_{ij}}{m} \qquad (i = 1, 2, \cdots, r) \tag{4.55}$$

再计算样本的平均数 $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{m} X_{ij}}{rm} = \frac{\sum_{i=1}^{r} \bar{X}_i}{r} \tag{4.56}$$

两阶段抽样的平均误差是由两部分构成的：第一部分是第一阶段从总体全部组中抽部分组所引起的组间误差；第二部分是由第二阶段在中选组中抽部分单位所引起的组内平均误差。在总体 R 组中抽取 r 组，又在 r 组中每组 M 个单位抽取 m 个单位的情况下，样本平均方差 $\sigma_{\bar{x}}^2$ 应等于组平均数组间方差的 $1/r$ 以及各组内方差平均数的 $1/rm$ 两项之和。再考虑阶段抽样是不重置抽样，各项必须乘以各自的修正系数，所以样本平均数的抽样平均误差为：

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{r}\left(\frac{R-r}{R-1}\right) + \frac{\sigma_2^2}{rm}\left(\frac{M-m}{M-1}\right)} \tag{4.57}$$

式中，$\sigma_1^2 = \sum(\mu_i - \mu)^2/R$ 为组间方差；$\sigma_2^2 = \sum\sigma_i^2/R$ 为组内方差的平均数。

应用以上公式，在得不到总体资料的情况下，可以用样本资料来代替。不过用样本组间方差代替总体组间方差，用样本组内方差代替总体组内方差，上述抽样平均误差公式就不是无偏估计，而必须加以适当的修正才能成为无偏估计量。修正后的抽样误差公式为：

$$\sigma_{\bar{x}} = \sqrt{\frac{S_1^2}{r}\left(\frac{R-r}{R-1}\right) + \frac{S_2^2}{rm}\left(\frac{r}{R}\right)\left(\frac{M-m}{M-1}\right)} \tag{4.58}$$

式中，$S_1^2 = \sum(\bar{X}_i - \bar{X})^2/r$ 为样本的组间方差；$S_2^2 = \sum S_{2j}^2/r$ 为各样本组内方差的平均数。

【例 4－18】对某地的木材蓄积量进行抽样调查。将总体 50 块面积相等的地块划分为 10 个区域，各区域包含 5 个地块。现在用两阶段抽样，先从 10 个区域中抽选 30%，再从中选的区域中抽选 60% 地块组成样本，实测各地块的木材量如表 4－9。试据此计算抽样的平均误差。

样本平均数：

$$\bar{X} = \frac{\sum \bar{X}_i}{r} = \frac{585 + 650 + 850}{3} = 695(M^3)$$

各区平均数区间方差：

$$S_1^2 = \frac{\sum(\bar{X}_i - \bar{X})^2}{r} = \frac{(585-695)^2 + (650-695)^2 + (850-695)^2}{3}$$
$$= 12\ 717$$

表 4-9 各地块木材积蓄量

区域	地块积蓄量(M^3) x_{ij}	样本平均数 $\bar{x}_i$	离差 $(x_{ij}-\bar{x}_i)$	离差平方 $(x_{ij}-\bar{x}_i)^2$
Ⅰ	576	585	-9	81
	585		0	0
	594		+9	81
Ⅱ	635	650	-15	225
	650		0	0
	665		+15	225
Ⅲ	820	850	-30	900
	850		0	0
	880		+30	900
	6 255	695	—	—

各区内方差：

$$S_{21}^2=\frac{\sum_j(X_{1j}-\bar{X}_1)^2}{3}=\frac{81+81}{3}=54$$

$$S_{22}^2=\frac{\sum_j(X_{2j}-\bar{X}_2)^2}{3}=\frac{225+225}{3}=150$$

$$S_{23}^2=\frac{\sum_j(X_{3j}-\bar{X}_3)^2}{3}=\frac{900+900}{3}=600$$

区内方差平均数：

$$S_2^2=\frac{\sum_j S_{2j}^2}{r}=\frac{54+150+600}{3}=268$$

$$R=10,\ r=R\times30\%=10\times30\%=3(\text{区})$$

$$M=5,\ m=M\times60\%=5\times60\%=3(\text{地块})$$

抽样平均误差：

$$\sigma_{\bar{x}}=\sqrt{\frac{S_1^2}{r}\left(\frac{R-r}{R-1}\right)+\frac{S_2^2}{rm}\left(\frac{r}{R}\right)\left(\frac{M-m}{M-1}\right)}$$

$$=\sqrt{\frac{12\ 717}{3}\left(\frac{10-3}{10-1}\right)+\frac{268}{3\times3}\left(\frac{3}{10}\right)\left(\frac{5-3}{5-1}\right)}=57.46(M^3)$$

五、不同抽样组织设计的比较

抽样实施前，进行抽样组织设计的目的是提高样本的代表性，增加抽样的效果。其中简单随机抽样是最基本的组织形式，在没有总体的有关信息时，人们一般都采用简单随机抽样。但是，在掌握了总体的有关信息情况下，在抽样设计阶段，如果能充分利用

上这些有关信息，有利于提高抽样的精确度和可靠度。当然，在抽样设计时，如何降低抽样的成本费用等因素，也是应该考虑的。限于篇幅，在此就不比较各抽样组织方式的费用大小了。

（一）简单随机抽样是基本抽样组织方式

抽样推断效果如何，依赖于所抽出样本的质量；样本的质量好坏，就看样本对总体的代表性如何。代表性好的样本，能从总体中"带出"更多有关总体的信息，样本分布与总体分布两者间存在密切的联系性与相似性。要做到这一点，首先就要避免抽样人员的"主观"影响，克服由此产生的"偏见"，因此，抽取样本都要求满足"随机性"。例如，类型抽样先将总体分成不同的部分，再从每一个部分中抽出一些单位，共同组成样本。虽然划分过程不存在随机性，但是，因为每一部分都需进行抽样，从每部分中取样时，就必须遵从随机原则了。对整群抽样也是这样，虽然对抽中的群进行的是全面调查，不存在随机性要求，但是，群是如何抽取的？是把"群"看成抽样单元，通过简单随机抽样抽出来的，在抽取群时，就满足随机原则了。任何概率推断，抽样设计时都应考虑在某个阶段或某个环节上遵从随机性要求，否则的话，样本的代表性就值得怀疑，抽样推断也就无从进行。再者，我们抽样的最终目的是为了进行统计推断，统计推断是概率推断，以随机变量的概率分布作为推断的基础。典型调查、重点调查都是进行判断抽样、决定是非概率的调查方式，不一定能体现总体的结构，根据这样的样本来推断总体的参数，其可靠度如何无从量化。由此可见，简单随机抽样是抽样设计的基础。

（二）类型抽样与整群抽样比较

类型抽样的平均误差与组间方差无关，它决定于组内方差的平均水平；整群抽样的平均误差与组内方差无关，它决定于组间方差大小。总体方差等于组间方差加上组内方差平均数。我们由此可推导出减小类型抽样与整群抽样平均误差的方法。

1.减小类型抽样平均误差的办法。提高组间方差，降低组内方差。具体来说，就是使类型抽样的各部分内部单位差异尽可能地小，不同类型间的差异尽可能地大。如果组间方差接近于总体方差，说明组内方差接近于0，这时组内单位基本上没有差异，这是一种极端情况，此时类型抽样的平均误差接近于0。

2.减小整群抽样平均误差的办法。为了降低抽样平均误差，应该设法降低群间方差。可通过提高群内方差方法达到降低群间方差的目的。

因此，类型抽样与整群抽样对总体进行分组的要求刚好是相反的：类型抽样要尽量提高组间方差降低组内方差；整群抽样应尽量提高组内方差降低组间方差。也就是说，类型抽样时，尽量使各组内的单位差异减小，各组间的单位差异增大；整群抽样则刚好相反。

于是，我们得到结论：类型抽样的前提是对总体的结构事先有一定的认识，有辅助

信息可资利用,这种信息与所研究的问题有密切关系,可以作为分类的依据。通过分类把总体中调查标志差异比较接近的单位归为一组,减少组内差异,再从各组中抽出样本,这样的样本就对总体有更大的代表性。所以类型抽样的特点是充分利用总体的已有信息。整群抽样可以对抽中的群进行集中调查,是一种较为方便有效的抽样组织方式;在总体单位没有原始资料可资利用时,整群抽样有利于提高抽样的效率。但是,正是由于整群抽样的抽样单位较集中,限制了样本在总体中分配的均匀性,所以,整群抽样有时代表性不是很理想,抽样误差较大。在实际抽样中,通常要适当增加一些样本单位,以缩小抽样误差,提高抽样推断的准确度。

(三)阶段抽样平均误差的控制

阶段抽样的特点是将抽样过程分成了几个阶段,每个阶段都有可能引起抽样的误差,因此阶段抽样误差的控制必须落实到抽样的各个阶段,抽样过程只要有一个阶段误差失控都有可能使整个抽样达不到理想效果。两阶段抽样误差控制,要具体到第一阶段的整群抽样的误差控制与第二阶段的类型抽样的误差控制两方面。实际中,人们还可以控制抽样误差,在各个阶段进行整体性平衡,以求达到最理想的效果。如两阶段抽样平均误差,既取决于组间方差也取决于组内平均方差,但组间方差是主要的因素。这是因为组间方差通常大于组内方差,而且抽样平均误差是由组间方差的 $1/r$ 和组内平均方差的 $1/rm$ 决定的,前者大于后者。所以在组织两阶段抽样时在相同样本容量要求下,适当增加第一阶段的组数,比增加第二阶段的单位数,能更显著地提高抽样效果。

第五节 Excel 在概率计算与参数估计中的运用

一、利用 Excel 进行概率计算

【例 4-19】例 4-5 的 Excel 分析方法如下:

(一)用 NORMDIST()函数计算

因为 $p(60<X<75)=p(X<75)-p(X<60)$

所以,在任一空单元格内输入如下公式,"=NORMDIST(75,60,12,TRUE)-NORMDIST(60,60,12,TRUE)"。其中,NORMDIST 函数的中间两个参数分别表示均值 60 和标准差 12,参数 TRUE 表示函数返回的是累积概率。

（二）用 NORMSDIST()函数计算

经过标准化后，$p(60<X<75)=p(0<Z<1.25)=p(Z<1.25)-p(Z<0)$

因此，在任一空单元格内输入如下公式，"=NORMSDIST(1.25)-NORMSDIST(0)"。

此外，其他一些分布（如 t-分布、二项分布等）的概率计算，与正态分布计算类似，只是所应用的函数不同而已。

二、区间估计

【例 4-19】我们以例 4-11 的求解为例，说明利用 Excel 进行区间估计的具体步骤：

第一，构造工作表。首先在各个单元格输入以下的内容。其中，A1:A50 单元格为样本数据——消费额，B2:B11 单元格为计算过程中要用到的样本统计量、置信水平、中间变量及最终结果的名称，C2:C11 单元格为相对应的计算公式。其中，置信水平由用户直接输入，如图 4-3 所示。

	A	B	C
1	消费额	总体平均数的估计	
2	15	样本容量	=COUNT(消费额)
3	24	样本均值	=AVERAGE(消费额)
4	38	样本标准差	=STDEV(消费额)
5	26	置信水平	0.9
6	30	抽样标准差	=样本标准差/SQRT(样本个数)
7	42	自由度	=样本容量-1
8	18	z值	=ABS(NORMSINV((1-置信水平)/2))
9	30	置信区间半径	=z值*抽样标准差
10	25	置信区间上界	=样本均值-置信区间半径
11	26	置信区间下界	=样本均值+置信区间半径

图 4-3

第二，定义变量名。先定义样本数据的变量名。选定 A1:A50 单元格，执行菜单命令"插入"→"名称"→"指定"，用鼠标点击"首行"选项，然后点击"确定"按钮。

再将 B2:B11 单元格的名称定义为 C2:C11 单元格各个公式计算结果的变量名。选定 B2:C11 单元格，执行菜单命令"插入"→"名称"→"指定"，用鼠标点击"最左列"选项，然后点击"确定"按钮。即可得到图 4-4 所示的结果。

通过样本数据，可以估计出麦当劳餐馆里顾客平均消费额是 32 元，并且能以 90% 的概率保证顾客平均消费额在 29.78 元至 34.22 元之间。

这里需要注意如下两点：①构造以上的工作表仅是为了让读者清楚具体的计算步骤，在实际应用中，可将一些中间步骤整合在一起，仅用一个公式就可以计算出置信区

间半径。定义消费额的变量名后，在任一空单元格内输入如下公式即可："=ABS(NORMSINV(0.1/2))*STDEV(消费额)/SQRT(COUNT(消费额))"。该公式

	A	B	C
1	消费额	总体平均数的估计	
2	15	样本个数	49
3	24	样本均值	32
4	38	样本标准差	9.451631253
5	28	置信水平	0.9
6	30	抽样标准误	1.350233066
7	42	自由度	48
8	18	z值	1.644853
9	30	置信区间半径	2.220934861
10	25	置信区间下界	29.77906514
11	26	置信区间上界	34.22093486

图 4-4

即相当于计算：$z_{\alpha/2}\cdot(S/\sqrt{n})$。②在总体方差已知的情况下，可借助函数CONFIDENCE进行区间估计。

本章小结

1.概率与概率分布。基本概念有：随机现象、随机事件、概率、随机变量、正态分布。引入随机变量是为了更好地利用数学工具分析随机现象。正态分布在统计推断中起着核心作用。设 $X\sim N(\mu,\sigma^2)$，作变换：$Z=(x-\mu)/\sigma$ 则 Z 服从标准正态分布。该变换称为"标准化变换"。

在大样本场合，抽样分布都可用正态分布去近似。

2.抽样与抽样分布。抽样的基本概念有：总体与样本、参数与统计量、样本容量与样本个数、重复与不重复抽样。

抽样分布的性质：

(1) $E(\bar{X})=\mu$，即样本平均数的期望(平均)等于总体的平均，这个性质与抽样方式无关。

$$(2)\sigma_{\bar{x}}=\begin{cases}\dfrac{\sigma}{\sqrt{n}} & \text{(重置抽样)}\\[2ex] \dfrac{\sigma}{\sqrt{n}}\cdot\sqrt{1-\dfrac{n}{N}} & \text{(不重置抽样)}\end{cases}$$

不重置抽样比重置抽样多了一个修正系数 $\sqrt{1-n/N}$；当总体标准差未知时，可考

虑用样本标准差代替它。

3.参数估计。参数估计分为点估计与区间估计两种。在点估计时,要注意估计方法的优良性要求,即无偏性、有效性与一致性要求。在区间估计时,要注意给出估计的概率保证的特殊要求。区间估计主要内容可归纳为图4-5。

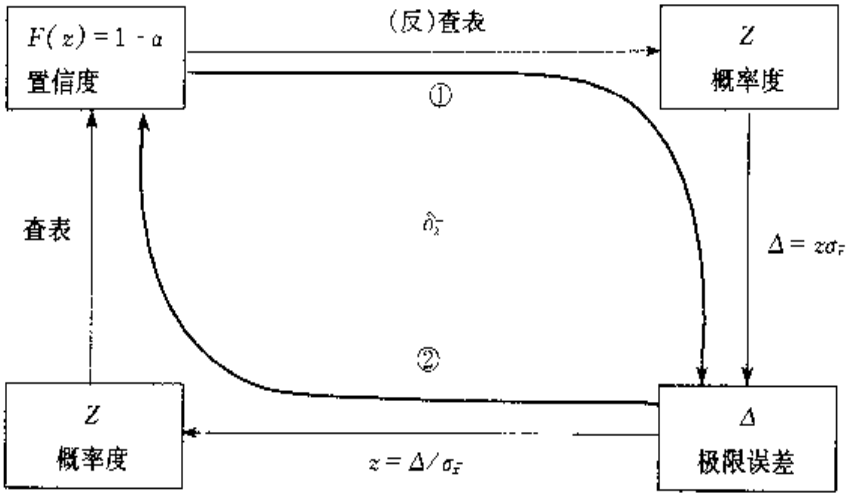


图4-5

图4 5展示了参数估计的两套模式:

(1)是根据置信度 $F(z)$ 的要求,估计出抽样的极限误差 Δ,并可由此求出置信区间:$\bar{x}\pm\Delta$。

(2)是根据给定的极限误差 Δ,求出区间估计的概率保证。

两套模式都离不开抽样平均误差 $\sigma_{\bar{x}}$,$\sigma_{\bar{x}}$ 处在参数估计的核心位置。

4.抽样设计。主要介绍了类型抽样、整群抽样、等距抽样和阶段抽样四种形式。请大家注意类型抽样与整群抽样在分组时的要求刚好相反:类型抽样分组要求尽量扩大组间差异,缩小组内差异,达到降低抽样平均误差的目的;整群抽样分组时则要求扩大组内差异,缩小组间差异,达到提高抽样效果的目的。

5.在抽样分布与参数估计中,Excel主要有以下几个方面的应用:①利用有关函数计算正态分布、t-分布、二项分布的概率;②利用有关函数直接计算样本均值、样本方差、样本标准差的点估计值;③构建工作表进行区间估计。

思考与练习题

一、单项选择题：

1.抽样平均误差与极限误差间的关系是(　　)。

a. 抽样平均误差大于极限误差

b. 抽样平均误差等于极限误差

c. 抽样平均误差小于极限误差

d. 抽样平均误差可能大于、等于或小于极限误差

2.在其他条件不变的情况下，如果允许误差缩小为原来的1/2，则样本容量(　　)。

a. 扩大为原来的4倍

b. 扩大为原来的2倍

c. 缩小为原来的1/2

d. 缩小为原来的1/4

3.类型抽样影响抽样平均误差的方差是(　　)。

a. 组间方差　　b. 组内方差

c. 总方差　　d. 允许误差

4.当样本单位数充分大时，样本估计量充分地靠近总体指标的可能性趋于1，称为抽样估计的(　　)。

a.无偏性　　b.一致性

c.有效性　　d.充分性

二、多项选择题：

1.影响抽样平均误差的因素有(　　)。

a.总体标志变异程度　　b.样本容量　　c.抽样方式

d.抽样的组织形式　　e.样本标志值的大小

2.抽样估计的抽样平均误差(　　)。

a.是不可避免要产生的　　b.是可以通过改进调查方法消除的

c.是可以事先计算的　　d.只有调查结束之后才能计算

e.其大小是可以控制的

3.确定样本容量时，一般用以下方法取得近似的总体方差估计值(　　)。

a.参考以往调查的经验资料

b.以试点调查的样本方差来估计

c.在做成数估计时,用成数方差最大值0.25来代替

d.假定总体不存在标志变异,方差为零

三、问答题:

1.某小组5个工人的周工资分别为140,160,180,200,220元,现在用重置抽样的方法从中抽出2个工人的工资构成样本。

要求:

(1)计算总体工人平均工资和标准差。

(2)列出样本平均工资的抽样分布。

(3)计算样本平均工资的平均数,并检验是否等于总体平均工资。

(4)计算样本平均工资的标准差。

(5)按公式计算抽样平均误差,并验证是否等于(4)的结果。

(6)用不重置抽样方法抽取样本,重新计算(1)~(5)。

2.某市居民家庭人均年收入是服从$\mu=4\ 000$元,$\sigma=1\ 200$元的正态分布,求该市居民家庭人均年收入:

(1)在5 000~7 000元之间的概率。

(2)超过8 000元的概率。

3.某工厂有1 500个工人,用简单随机重置抽样方法抽出50个工人作为样本,调查其工资水平,结果如表4-10。

表4-10

月工资水平(元)	524	534	540	550	560	580	600	660
工人数(人)	4	6	9	10	8	6	4	3

要求:

(1)计算样本平均数和抽样平均误差。

(2)以95.45%概率保证,估计该厂工人的月平均工资和工资的区间。

4.为了调查某地区人口总数,在该地区15 000户中以不重复抽样方法随机抽选30户作为样本,每户人口数据如下:

5,6,3,3,2,3,3,3,4,4

3,2,6,4,3,5,4,5,3,3

4,3,3,1,2,5,3,4,2,4

要求:

以95.45%的概率,推断该地区人口总数。若要求人口总数误差不超过33 000人,

则至少需抽取多少户作为样本。

5.采用简单随机重置抽样的方法,从2 000件产品中抽查200件,其中合格品190件。

要求:

(1)计算合格品率及其抽样平均误差。

(2)以95.45%概率保证程度,对合格品率和合格品数量进行区间估计。

(3)如果极限误差为2.31%,则其概率保证程度是多少?

6.某大学欲对学生身高进行调查,从该校160个班中按不重复抽样方法随机抽取16个班作为样本,对16个班的全体同学均测量身高,得知平均身高172.05cm,各班之间身高的方差为60.5。

要求:

以90%的概率推断该校学生身高的可能范围。若改为每个中选班再进行一次抽样,并假设每班均为30人,按不重复抽样方法在每班抽选15人作为第二阶段单位。且算得平均身高为172.33cm,平均班内方差为125,仍以90%的概率推断该校学生平均身高的范围。

第五章

假设检验

★ 本章学习要点与要求 ★

本章介绍假设检验的基本概念以及参数检验与非参数检验的主要方法。通过学习，要求：①掌握统计检验的基本概念，理解该检验犯两类错误的可能；②熟练掌握总体均值与总体成数指标的各种检验方法，包括 z-检验、t-检验和 p-值检验；③掌握基本的非参数检验方法，包括符号检验、秩和检验与游程检验；④能利用 Excel 进行假设检验。

第一节　假设检验概述

一、假设检验的基本概念

假设检验是统计推断的另一种方式，它与区间估计的区别主要在于：区间估计是用给定的大概率推断出总体参数的范围，而假设检验是以小概率为标准，对由总体的状况所做出的假设进行判断。假设检验与区间估计结合起来，构成完整的统计推断内容。假设检验分为两类：一类是参数假设检验，另一类是非参数假设检验。本章将分别讨论这两类检验方法。

进行假设检验，首先要对总体的分布函数形式或分布的某些参数做出假设，然后再根据样本数据和“小概率原理”，对假设的正确性做出判断。这种思维方法与数学里的

"反证法"很相似,"反证法"先将要证明的结论假设为不正确的,作为进一步推论的条件之一使用,最后推出矛盾的结果,以此否定事先所作的假设。反证法所认为矛盾的结论,也就是不可能发生的事件,这种事件发生的概率为零,该事件是不能接受的现实。其实,我们在日常生活中,不仅不肯接受概率为0的事件,而且对小概率事件,也持否定态度。比如,虽然偶尔也有媒体报导陨石降落的消息,但人们不必担心天空降落的陨石会砸伤自己。

所谓小概率原理,即指概率很小的事件在一次试验中实际上不可能出现。这种事件称为"实际不可能事件"。

小概率的标准是多大?这并没有绝对的标准,一般我们以一个所谓显著性水平$\alpha(0<\alpha<1)$作为小概率的界限,α的取值与实际问题的性质有关。所以,统计检验又称显著性检验。

下面通过一个具体例子来说明假设检验是怎样进行的。

【例5-1】消费者协会接到消费者投诉,指控某品牌纸包装饮料存在容量不足,有欺骗消费者之嫌。包装上标明的容量为250毫升。消费者协会从市场上随机抽取50盒该品牌纸包装饮品,测试发现平均含量为248毫升,小于250毫升。这是生产中正常的波动,还是厂商的有意行为?消费者协会能否根据该样本数据,判定该饮料厂商欺骗了消费者呢?

上述例子中,消费者协会实际要进行的是一项统计检验工作,检验总体平均容量是否等于包装上注明的250毫升。即,检验总体平均$\mu=250$是否成立。这就是一个原假设(Null Hypothesis),通常用H_0表示,即:

$$H_0:\mu=250$$

与原假设对立的是备选假设(Alternative Hypothesis)H_1,备选假设是在原假设被否定时另一种可能成立的结论。备选假设比原假设还重要,这要由实际问题来确定,一般把期望出现的结论作为备选假设。上例中可能的备选假设有三种。

第一种:如果消费者协会希望知道的是,该品牌饮料的平均容量是否为标明的250毫升,则$H_1:\mu\neq250$

第二种:如果消费协会希望知道该品牌饮料的平均容量是否少于标明的250毫升,则$H_1:\mu<250$

第三种:如果消费者协会希望知道该品牌饮料的平均容量是否大于标明的250毫升,则$H_1:\mu>250$

由于备选假设不同,可将假设检验分为双侧(边、尾)检验(Two Tailed Test)和单侧(边、尾)检验(One Tailed Test)。对此,我们在后面将进一步说明。

原假设与备选假设确定之后,我们要构造一个统计量来决定是"接受原假设,拒绝

备选假设”，还是“拒绝原假设，接受备选假设”。对不同的问题，要选择不同的检验统计量。检验统计量确定后，就要利用该统计的分布以及由实际问题中所确定的显著性水平，来进一步确定检验统计量拒绝原假设的取值范围，即拒绝域。在给定的显著性水平 α 下，检验统计量的可能取值范围被分成两部分：小概率区域与大概率区域。小概率区域就是概率不超过显著性水平 α 的区域，是原假设的拒绝区域；大概率区域是概率为 $1-\alpha$ 的区域，是原假设的接受区域。如果样本统计量落入拒绝域，我们就拒绝原假设，接受备选假设，认为样本数据支持备选假设的结论；如果样本统计量落入接受区域，我们就接受原假设，认为没有充分证据证明备选假设结论为真。请注意，我们这里使用的判断语气比较委婉，原因是：拒绝域是小概率区域，按小概率原理应该拒绝原假设，但是，小概率事件不是完全不可能事件，还是有可能发生的；接受区域是大概率区域，大概率事件也不是必然事件。无论是接受原假设还是拒绝原假设，都有产生判断失误的可能。因此，不宜将统计检验的结论过于绝对化。

二、两种类型的错误

统计假设检验是通过比较检验统计量的样本数值，来作决策。统计量是随机变量，据之所作的判断不可能保证完全正确。一般来说，决策结果存在以下四种情形：

原假设是真实的，判断结论是接受原假设，这是一种正确的判断；

原假设是不真实的，判断结论是拒绝原假设，这也是一种正确的判断；

原假设是真实的，判断结论是拒绝原假设，这是一种产生“弃真错误”的判断；

原假设是不真实的，判断结论是接受原假设，这又是一种产生“取伪错误”的判断。

以上四种判断可归纳为如表 5-1 的形式。

表 5-1　统计决策表

	接受 H_0	拒绝 H_0
H_0 真实	判断正确	弃真错误(α)
H_0 不真实	取伪错误(β)	判断正确

以上的弃真错误也称作假设检验的“第一类错误”；取伪错误也称作假设检验的“第二类错误”。无论是第一类错误还是第二类错误，都是检验结论失真的表现，都是应尽可能地加以避免的情形，如果不能完全避免，也应该对其发生的概率加以控制。

第一类错误产生的原因是：在原假设为真的情况下，检验统计量不巧刚好落入小概率的拒绝区域。因此，犯第一类错误的概率大小就等于显著性水平的大小，即等于 α。我们可以通过控制显著性水平大小的方式，来控制犯第一类错误的可能性大小。α 定得越小，犯第一类错误的可能性就越小，例如 $\alpha=0.05$，表示犯第一类错误的可能性为 5%，100 次判断中，产生弃真性错误的次数是 5 次；进一步降低显著性水平，取

$\alpha=0.01$,这时犯第一类错误的概率下降为1%。所以统计学上,又称第一类错误为α错误。

第二类错误是"以假为真"的错误,即把不正确的原假设,当做正确的而将它接受了的错误。犯第二类错误大小的概率记为β,因此,统计学上称第二类错误为β错误。犯第二类错误的概率与犯第一类错误的概率是密切相关的,在样本一定的条件下,α小,β就增大;α大,β就减小。为了同时减小α和β,只有增大样本容量,减小抽样分布的离散性,这样才能达到目的。它们的这种关系可以通过正态分布的统计予以检验,如图5-1所示。

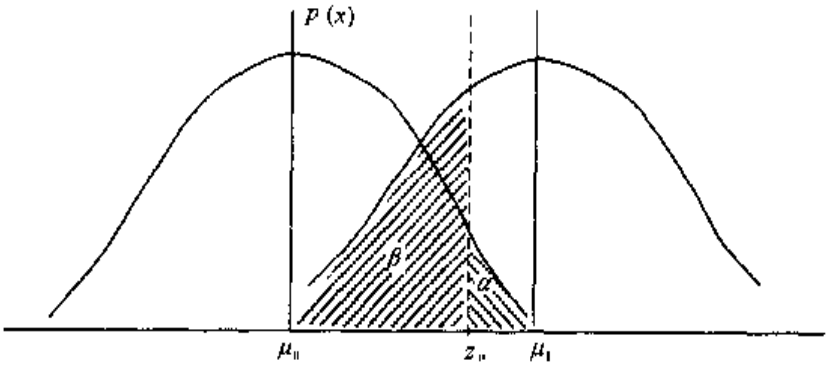


图5-1 α与β关系示意图

【例5-2】按照法律,在证明被告有罪之前应先假定他是无罪的。也就是原假设是H_0:被告无罪;备选假设H_1:被告有罪。法庭可能犯的第一类错误是:被告无罪但判他有罪;第二类错误是:被告有罪但判他无罪。犯第一类错误的性质是"冤枉了好人",第二类错误的性质是"放过了坏人"。为了减小"冤枉好人"的概率,应尽可能接受原假设,判被告无罪,这就有可能增大了"放过坏人"的概率;反过来,为了不"放过坏人",增大拒绝原假设的概率,相应地就又增加了"冤枉好人"的可能性,这就是α与β的关系。当然,这只是在"一定的证据下"的两难选择。如果进一步收集有关的证据,在充分的证据下,就有可能做到既不冤枉好人,又不放过坏人。在现有证据不充分的条件下,法庭控制两类错误概率的实践是:按案件的性质决定首先要控制哪一类错误的概率,如果案件将来对社会危害大,就要控制少犯第二类错误的概率,免得放过的坏人继续危害社会;如果案件对社会没有什么大的危害,不妨"放他一马",免得冤枉了好人,影响当事人"一生的前程"。

三、检验功效

检验效果的好与坏,与犯两类错误的概率都有关。一个有效的检验首先是犯第一

类错误的概率 α 不能太大，否则的话，就经常产生弃真现象；另外，β 错误就是取伪的错误，在犯第一类错误概率得到控制的条件下，犯取伪错误的概率也要尽可能地小，或者说，不取伪的概率 $1-\beta$ 应尽可能增大。$1-\beta$ 越大，意味着当原假设不真实时，检验判断出原假设不真实的概率越大，检验的判别能力就越好；$1-\beta$ 越小，意味着当原假设不真实时，检验结论判断出原假设不真实的概率越小，检验的判别能力就越差。可见 $1-\beta$ 是反映统计检验判别能力大小的重要标志，我们称之为检验功效或检验力。

前面分析说明，第一类错误和第二类错误是一对矛盾体，在其他条件不变时，减小犯第一类错误的可能性，势必增加犯第二类错误的可能性；增大第一类错误的可能性，又能减小犯第二类错误的可能性。可见 α 的大小，影响到 β 的大小，进而影响到 $1-\beta$ 的大小。犯第一类错误的概率或检验的显著性水平 α 是影响检验力的一个重要因素。在其他条件不变的情况下，显著性水平 α 增大，β 随之减小，检验功效就增强。可见取 $\alpha=0.1$ 时比取 $\alpha=0.01$ 时，检验的功效强，检验力大。

我们在统计检验中，一般都是首先控制犯第一类错误的概率，也就是显著性水平 α 都尽量取较小的值，尽量避免犯弃真的错误，在其他条件不变时，β 就增大，检验的功效就减弱。该如何来调和这一对相互对抗的矛盾呢？惟一的办法就是增大样本容量，因为增加样本容量能够既保证满足较小的 α 需要，同时又能减小犯第二类错误的概率 β，抵消检验功效的衰减。可见样本容量大小是影响检验功效大小的一个重要因素，可通过增大样本容量的方法提高检验功效。然而，实际上样本容量 n 的增加也是有限制的，兼顾 α 与 β 很困难，这时，鉴于 α 风险一般比 β 风险重要，首先考虑的还是控制 α 风险。

影响检验功效大小的另一个因素是原假设与备选假设间的差异程度。如果这两个假设间的差异是非常明显的，这时原假设不真而取伪的可能性就减小，即 β 就减小，检验功效就强。否则的话，就较难通过检验把原假设与备选假设区分开来，也就会影响检验功效的提高。

第二节　总体参数检验

一、单侧检验与双侧检验

在前面介绍的统计检验程序中，我们看到，通过确定的检验统计量、事先给出的显著性水平，可以找出一个临界值，将统计量的取值范围划分成接受区域与拒绝区域两部分。拒绝区域是检验统计量取值的小概率区域，我们可以将这个小概率区域安排在检

验统计量分布的两端,也可以安排在分布的一侧,分别称作双侧检验与单侧检验。单侧检验又按拒绝域在左侧还是在右侧而分为左侧检验与右侧检验两种。我们通过服从正态分布的检验统计量如图 5-2 所示。

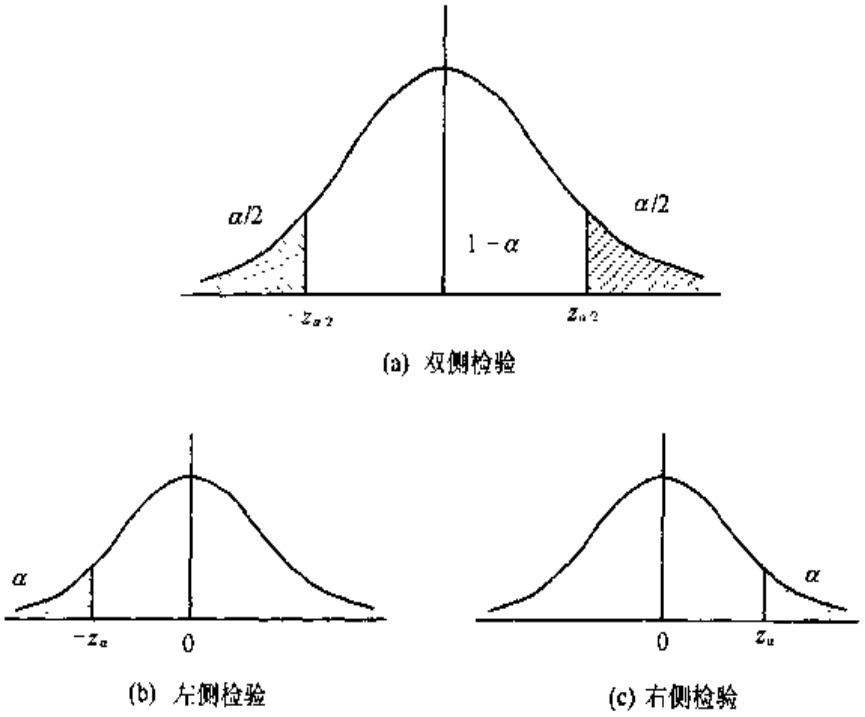


图 5-2 双侧、单侧检验的拒绝域分配

一个统计检验究竟是使用双侧检验还是使用单侧检验,单侧检验时,是使用左侧检验还是使用右侧检验,这都取决于备选假设的性质。比如,当总体标准差已知时,我们对正态总体的均值进行检验,使用的检验统计量是:

$$z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \tag{5.1}$$

按备选假设的不同,存在三种情形:

(1)$H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$

(2)$H_0:\mu=\mu_0, H_1:\mu<\mu_0$

(3)$H_0:\mu=\mu_0, H_1:\mu>\mu_0$

对于第一种情形,备选假设是总体均值不等于一个给定的 μ_0,检验统计量取极端值,不论是在极端大的右侧取值,还是在极端小的左侧取值,都有利于拒绝原假设,接受备选假设。因此,拒绝原假设的拒绝域被安排在左右两侧,使用的是双侧检验。对于第二种情形,备选假设是总体均值小于一个确定的 M_0,检验统计量越在极端大的右侧取值,越有利于说明总体的均值较大,这对备选假设是不利的。因此,拒绝域不宜安排在

右侧，安排到左侧比较合适，应使用左侧检验。对于第三种情形，备选假设是总体均值大于一个确定的 μ_0，检验统计量在极端小的左侧取值时，对备选假设是不利的。因此拒绝原假设的拒绝域被安排在左侧，即使用单侧检验中的右侧检验。最后我们可以归纳结论：用单侧检验还是双侧检验，使用左侧检验还是右侧检验，决定于备选假设中的不等式形式与方向。与"不相等"对应的是双侧检验，与"小于"相对应的是左侧检验，与"大于"相对应的是右侧检验。

二、总体平均数的检验

（一）先考虑总体标准差 σ 已知的情形

在例 5－1 中，设按历史资料，总体的标准差是 4 毫升。我们通过检验总体均值是否等于 250 毫升，来判断饮料厂商是否欺骗了消费者。程序如下：

第一步：确定原假设与备选假设。

$$H_0: \mu = 250, H_1: \mu < 250$$

以上的备选假设是总体均值小于 250 毫升，因为消费者协会希望通过样本数据推断出厂商的欺骗行为（大于 250 毫升一般不会发生）。因此使用左侧检验。

第二步：构造出检验统计量。

我们知道，如果总体的标准差已知，则正态总体（正常情况下，生产饮料的容量服从正态分布）的抽样平均数，也服从正态分布，对它进行标准化变换，可得到：

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{5.2}$$

可用 z 作为检验统计量。

第三步：确定显著性水平，确定拒绝域。

显著水平由实际问题确定，我们这里取 $\alpha = 0.05$，使用左侧检验，拒绝域在左边，查标准正态分布表得临界值：$-z_\alpha = -1.645$，拒绝域是 $z < -1.645$。

第四步：计算检验统计量的数值。

将样本平均数 $\bar{X} = 248$，$n = 50$，代入检验统计量得：

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{248 - 250}{4/\sqrt{50}} = -3.54 < -1.645$$

第五步：判断。

检验统计量的样本取值落入拒绝域。拒绝原假设，接受备选假设，认为有足够的证据说明该种纸包饮料的平均容量小于包装盒上注明的 250 毫升，该厂商有欺骗行为。

（二）总体标准差未知的情形

我们在区间估计时已知，当总体的标准差或方差未知时可相应地用样本的标准差

与方差代替它们。但这时检验统计量就不服从标准正态分布了。事实上这时的检验统计量是：

$$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1) \tag{5.3}$$

式中，S 是样本标准差，检验统计量服从的是自由度为 $n-1$ 的 t-分布。这里用 t 作为检验总体均值的统计量，称之为 t-检验统计量（t-test statistic）。但是，在大样本场合（样本容量 n 大于 30 时），t-统计量与标准正态分布统计量近似，通常用 z 检验代替 t 检验。

【例 5-3】某厂采用自动包装机分装产品，假定每包产品的重量服从正态分布，每包标准重量为 1 000 克，某日随机抽查 9 包，测得样本平均重量为 986 克，样本标准差是 24 克。试问在 $\alpha=0.05$ 的显著性水平上，能否认为这天自动包装机工作正常？

解：

第一步：确定原假设与备选假设。

$$H_0:\mu=1\,000, H_1:\mu\neq 1\,000$$

以上的备选假设是总体均值不等于 1 000 克，因为只要均值偏离 1 000 克，都说明包装机工作不正常。因此应使用双侧检验。

第二步：构造出检验统计量。

由于总体标准差未知，用样本标准差代替，相应检验统计量是 t-统计量：

第三步：确定显著性水平，确定拒绝域。

$\alpha=0.05$，查 t-分布表（自由度 $n-1=8$），得临界值是 $t_{\alpha/2}(n-1)=t_{0.025}(8)=2.306$，拒绝域是 $|t|\geqslant 2.306$。

第四步：计算检验统计量的数值。

样本平均数 $\bar{x}=986$，$n=9$，$S=24$，代入 t-检验统计量得：

$$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}=\frac{986-1\,000}{24/\sqrt{9}}=-1.75$$

第五步：判断。

由于 $|t|<2.306$，检验统计量的样本取值落入接受区域，所以接受 H_0。样本数据说明这天的自动包装机工作正常。

我们不难发现，t-检验与正态检验十分相似，不同之处只是在确定临界值时，查的分布表不同，而且，在大样本场合两者检验过程可完全相同。

三、总体成数的检验

由前一章的抽样分布我们知道，样本成数是一个特殊的平均数，当样本容量较大

时，下列统计量服从标准正态分布：

$$z=\frac{P-p}{\sqrt{\frac{p(1-p)}{n}}} \tag{5.4}$$

式中，小写字母 p 代表总体的成数，大写字母 P 代表样本的成数。

以上的 z 统计量可以用作总体成数检验的检验统计量。

【例 5－4】某企业声明有 30% 以上的消费者对其产品质量满意。如果随机调查 600 名消费者，表示对该企业产品满意的有 220 人。试在显著性水平 $\alpha=0.05$ 下，检验调查结果是否支持企业的自我声明。

解：

第一步：作出假设。

$$H_0: p=30\%, H_1: p>30\%$$

以上的备选假设是企业自我声明的结论，我们希望该企业说的是实话。因此使用右侧检验。

第二步：构造 z 检验统计量。

第三步：确定拒绝域。

显著水平 $\alpha=0.05$，查标准正态分布表得临界值：$z_\alpha=1.645$，拒绝域是 $z>1.645$。

第四步：计算检验统计量的数值。

样本成数 $p=220/600=0.37$，总体假设的成数 $P=0.3$，代入 z 检验统计量得：

$$z=\frac{P-p}{\sqrt{\frac{p(1-p)}{n}}}=\frac{0.37-0.3}{\sqrt{0.3(1-0.3)/600}}=3.7$$

第五步：判断。

检验统计量的样本取值 $z=3.7>1.645$，落入拒绝域。拒绝原假设，接受备选假设，认为样本数据证明该企业声明属实。

四、p-值检验

p-值检验是国际上流行的检验格式。该检验格式是通过计算 p-值，再将它与显著性水平 α 作比较，决定拒绝还是接受原假设。所谓 p-值就是拒绝原假设所需的最低显著性水平。p-值判断的原则是：如果 p-值小于给定的显著性水平 α，则拒绝原假设；否则，接受原假设。或者，更加直观的原则是：如果 p-值很小，拒绝 H_0，p-值很大，接受 H_0。p-值检验无需针对不同的显著性水平，先查分布表确定临界值，然后才能进行检验判断；而是可直接把计算机计算出来的 p-值与显著性水平进行比较，立刻作出统计决策。Excel 等统计分析软件都直接给出了 p-值。请大家注意的是这里的 p-值

是指概率,不要与前面的成数指标相混淆。

p-值实际上是检验统计量超过(大于或小于)由样本数据所得数值的概率。因此,p-值与检验统计量的分布、是双侧检验还是单侧检验、是左侧还是右侧检验都有关系。

(一)z 检验的 p-值

z 检验统计量服从正态分布,可利用标准正态分布表计算 p-值。先看总体的均值检验的 p-值计算公式,以 z_0 表示检验统计量的抽样数据,则 p-值的计算方法如下:

如果 $H_1:\mu\neq\mu_0$, p-值 $=2p(z>|z_0|)$ (5.5)

如果 $H_1:\mu\geqslant\mu_0$, p-值 $=p(z>z_0)$ (5.6)

如果 $H_1:\mu\leqslant\mu_0$, p-值 $=p(z<z_0)$ (5.7)

总体成数检验的 p-值计算公式,与上述三式完全相同,只需将总体均值换成总体成数就行了。

【例 5-5】利用 p-值检验重新检验例 5-1。

解:

第一步、第二步与例 5-1 完全相同,故省略之。

第三步:计算样本统计的数值。

样本平均数 $\bar{X}=248$, $n=50$,代入检验统计量得:

$$z_0=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}=\frac{248-250}{4/\sqrt{50}}=-3.54$$

第四步:计算 p-值。

使用左侧检验,p-值 $=p(z<z_0)$。查标准正态分布表得:

$$p\text{-值}=\frac{1}{2}[1-F(3.54)]=\frac{1}{2}(1-0.999\,68)=0.000\,16$$

第五步:判断。

p-值小于给出的显著性水平(0.05),拒绝原假设,接受备选假设,与例 5-1 的结论相同。

【例 5-6】某国总统选举中,有位候选人几个月前的支持率是 60%。近期的一项调查,访问了 500 人,发现他的支持率变成了 55%。显著性水平取 0.05,试用 p-值方法,检验他的支持率是否下降了?

解:

第一步:确定原假设与备选假设。

$$H_0:P=60\%,\ H_1:P<60\%。$$

第二步:构造检验统计量。

第三步:计算检验统计量的数值。

$P=55\%$，代入检验统计量中，得：

$$z_0=\frac{P-p}{\sqrt{\frac{p(1-p)}{n}}}=\frac{0.55-0.6}{\sqrt{\frac{0.6\times(1-0.6)}{500}}}=-2.28$$

第四步：计算 p-值。

由于是左侧检验，p-值计算为：

p-值 $=p(z<z_0)=p(z<-2.28)=0.011\ 3$

第五步：判断。

p-值小于显著性水平 0.05，拒绝原假设，接受备选假设，有足够的证据证明该候选人的支持率已经下降了。

（二）t-检验的 p-值

进行 t-检验时，p-值的计算与 z 检验的 p-值计算方法相同，两者的区别在于：进行 t-检验其 p-值不是查标准正态分布表，而是查 t-分布表，自由度是 n-1。在大样本场合，这些差别都可以不存在，因为这时 t-分布与标准正态分布近似。由中心极限定理我们还可知道，在大样本场合，检验统计的分布还不受总体分布的正态性限制。

我们可以归纳 p-值检验的步骤如下：①建立原假设与备选假设；②确定检验统计量及其分布；③将样本观测值代入检验统计量计算出其样本数值；④计算 p-值；⑤将 p 值与显著性水平 α 相比较，作出判断。

第三节　非参数检验

一、自由分布检验概述

第二节介绍的统计检验属于参数假设检验，都是先对样本所属总体的性质作出若干的假定，或对总体的分布形状加以限定，然后对总体的有关参数情况进行统计假设检验。因此，参数检验又称为限定分布检验。自由分布检验与限定分布检验不同，它是指在假设检验时不对总体分布的形状加以限制的检验。与参数检验相对应，自由分布检验又称为非参数检验，但这里的非参数并不是指“不含参数或没有参数”。

自由分布检验与参数检验相比，具有以下优点：

首先，检验条件比较宽松，适应性强。自由分布检验对资料的要求不像参数检验那样严格，它适合于处理诸如非正态的、方差不等的或分布形状未知的资料。

其次，自由分布检验的方法比较灵活，用途广泛。它不但可以应用于处理测量层次较高的定距、定比数据，也适用于层次较低的定类、定序数据。对于那些不能进行加、减、乘、除运算的定类数据与定序数据，可使用符号检验、秩和检验等方法进行检验。

再次，自由分布检验的计算相对简单。由于自由分布的检验方法不用复杂计算，一般使用计数方法就可以了，因此，它的计数过程与结果都比较简单、直观与明显。

自由分布检验由于其要求的条件简单，由此产生的缺点也是明显的。由于它对原始数据中包含的信息利用得不够充分，检验的功效也就相对较弱。当总体的分布形式已知时，基于这种分布类型的参数方法，一般说来要优于非参数方法。例如，对于一批资料，可同时适用于参数的 t-检验、非参数的符秩检验和符号检验。其检验功效是，t-检验的最好，符秩检验次之，符号检验最差。这主要是由于符号检验对信息的利用最不充分。所以参数检验与自由分布检验是针对不同情况提出的两种统计方法，它们各有优缺点，可以互为补充。

下面简单介绍几种最常用的自由分布检验方法。

二、符号检验

符号检验虽然是自由分布检验中最简单的一种检验方法，但是这种方法却十分常用。该方法是建立在以正、负号表示样本数据与假设参数值差异关系基础上的，因此被称为符号检验。该方法既适用于单样本场合，也适用于配对样本场合。

（一）单样本场合的符号检验

在单样本的场合，我们可以用符号检验方法，检验总体的中位数是否在某一指定的位置。

假设总体中位数的真值是 A，即 $H_0: M_e = A$，再从样本观测结果：

$$x_1, x_2, \cdots, x_n$$

每个数据都减去 A，只记录其差数的符号，即，当 $x_i > A$ 时，记正号；当 $x_i < A$ 时，记负号；当 $x_i = A$ 时，将此数据剔除，不记录。设正号个数是 n^+，负号个数是 n^-，从理论上看，当原假设为真时，n^+ 与 n^- 应该很接近；若两者相差太远，就有理由拒绝原假设。这就是中位数检验的基本原理。

【例 5－7】设有 20 个工人，他们一天生产的产品件数，抽样结果如下：

168 163 160 172 162 168 152 153 167 165 164 142 173 166 160 165 171 186 167 170

试以 $\alpha = 0.10$ 的检验水平，判定总体中位数是否是 160。

解：

第一步：作出假设。

$H_0: M_e = 160, H_1: M_e \neq 160$

由备选假设知，这个检验是双侧的。

第二步：计数。

对样本数据，大于 160 的记下“+”，小于 160 的记下“-”，等于 160 的予以剔除（以 0 记之），结果如下：

+ + 0 + + + - - + + + - + + 0 + + + + +

计数以上“+”的个数是 $n^+ = 15$，“-”的个数 $n^- = 3$，剔除数据 2 个。最后有效的样本个数为 $n = n^+ + n^- = 18$。

第三步：确定拒绝域。

显著水平 $\alpha = 0.10$，由于进行双侧检验，拒绝域分布在两边，每侧概率 $\alpha/2 = 0.05$，查二项分布临界值表，得到拒绝域的临界值是 13。

第四步：选择 n^+，n^- 较大者，再与临界值比较。

结果是 15>13。

第五步：判断。

由上一步的比较结果可知，样本落入拒绝域，所以拒绝原假设，认为样本数据不能证明总体中位数等于 160 件。

（二）配对样本场合的符号检验

在实验过程中，我们经常将实验组数据与对照组数据进行比较，检验实验结果是否存在显著的差异，这就是一种配对样本场合。可通过符号检验方法来确定两组配对样本间是否有显著差异。

样本配对场合与单样本场合的符号检验，基本原理是一致的。设从两个总体中分别抽出一个容量相等的样本，然后将两样本的数据进行一一配对，得到一组配对值。再将各对配对值相减，记录下差数的符号，计算出“+”的个数 n^+ 与“-”的个数 n^-。如果两个样本的总体差异不显著，配对值之差的正负号出现的概率各是 1/2，则 n^+ 与 n^- 应当非常接近；如果 n^+、n^- 相差太大的话，说明两总体存在显著差异。

【例 5-8】兹给出成对实验中用两种不同饲料喂养生猪的增重结果，如表 5-2 所示。

表 5-2 生猪增重对照表

配对序号	A 饲料	B 饲料	符号	配对序号	A 饲料	B 饲料	符号
1	25	19	+	10	28	26	+
2	30	32	-	11	32	30	+
3	28	21	+	12	29	25	+
4	34	34	0	13	30	29	+

续表

配对序号	A饲料	B饲料	符号	配对序号	A饲料	B饲料	符号
5	23	19	+	14	30	31	−
6	25	25	0	15	31	25	+
7	27	25	+	16	29	25	+
8	35	31	+	17	23	20	+
9	30	31	−	18	26	25	+

取 $\alpha=0.05$,用符号检验,问饲料A是否对生猪的增重更有利?

解:

第一步:作出假设。

$$H_0: p=1/2, H_1: p>1/2$$

原假设是两种饲料对生猪增重相同,正、负号出现的概率相同,都是1/2。备选假设是A饲料对生猪增重更有利,正号出现的概率大于负号出现的概率。因此,这个检验是一个右侧检验。

第二步:计数。

"+"的个数是 $n^+=13$,"−"的个数 $n^-=3$,剔除数据2个。最后有效的样本个数为 $n=n^++n^-=16$。

第三步:确定拒绝域。

显著水平 $\alpha=0.05$,由于进行右侧检验,拒绝域分布在右边,查二项分布临界值表,得到拒绝域的临界值是12。

第四步:比较正号个数与临界值大小。

结果是13>12。

第五步:判断。

由上一步的比较结果可知,样本落入拒绝域,所以拒绝原假设,样本数据证明A饲料对生猪增重更加有利。

符号检验的优点是,它对总体的分布、方差的同一性等都不作任何假定,只需通过差数确定正负号的个数就可以了。它适用于测量要求不太精密的场合,也适合于那些不能和不宜于定量测量而只能定序测量的场合。然而,这也是符号检验的缺点所在,它只考虑差异的方向,不考虑数量差异的程度,从而丧失了一部分可资利用的信息。

三、秩和检验

为了弥补符号检验对数据信息利用不充分的不足,可以考虑用秩和检验方法。秩和检验方法也可用于检验两个独立的样本是否来自同一个总体,或判断总体间是否存在显著性的差异。它与符号检验最主要的区别是:符号检验只考虑样本间差数的符号,

而秩和检验还要考虑差数的顺序，比符号检验利用数据信息更加充分，因此，检验功效就更强。应注意的是，秩和检验也只是利用样本差数的顺序，并没有利用样本差数数值本身，因而比参数检验利用样本数值本身的信息要逊色一些；但是，参数检验受总体分布已知这一条件的限制，而秩和检验就不受这个限制，这又是秩和检验的长处。

设分别从两个未知的总体独立、随机地抽取容量为 n_1 和 n_2 的样本，把样本容量较小的总体称为总体Ⅰ，容量较大的总体称为总体Ⅱ。如果两样本容量相等，就把任意一个总体称作总体Ⅰ，另一个称作总体Ⅱ，这里不妨设 $n_1 < n_2$。

现将两个样本混合起来，并按数据的大小，从小到大排列编号，每个数值的编号就是它的秩次。如果混合样本中有若干个相同的数值，则把它们的秩次进行简单算术平均，用此平均值作为这些数值的秩次，计算来自总体Ⅰ的 n_1 个数据在混合样本中的秩次之和，记为 T。

显然 T 最小的可能值是 $T_1 = 1+2+3+\cdots+n_1 = [n_1(n_1+1)]/2$；最大的可能值是 $T_2 = (n_2+1)+(n_2+2)+\cdots+(n_2+n_1) = n_1[(n_2+1)+(n_2+n_1)]/2$。如果两个总体分布无显著差异，则 T 值不应太大或太小，约等于中间值 $(T_1+T_2)/2$；如果总体Ⅰ分布于总体Ⅱ的右边，T 将接近其最大值 T_2；如果总体Ⅰ位于总体Ⅱ的左边，T 将接近于它的最小值 T_1。因此，我们可以用秩和 T 作为检验的统计量。

由于 T 的分布与 n_1, n_2 的大小都有关，因此秩和检验中的临界值的确定有两种方法。

第一种方法，当 n_1 和 n_2 都不超过 10 时，查“秩和检验表”确定临界值；

第二种方法，当 n_1 和 n_2 都超过 10 时，秩和 T 服从正态分布：

$$T \sim N[n_1(n_1+n_2+1)/2,\ n_1n_2(n_1+n_2+1)/12] \tag{5.8}$$

先对 T 进行标准化变换，再利用标准正态分布表，确定检验的临界值。

下面通过具体的例子，说明秩和检验的应用。

【例 5-9】有 A,B 两家厂商供应同一种商品，两家的商品价格与性能一致，但使用寿命是否一致有待检验。今分别从两家生产的产品中抽出样本，测定产品使用寿命(单位:小时)如下所示。

A厂商产品：	5	11	6	9	7	10
B厂商产品：	8	6	10	7	8	

试以 0.05 的显著性水平，检验两厂商产品寿命是否有差异?

解：

第一步：作出假设。

$$H_0: M_A = M_B, H_1: M_A \neq M_B$$

原假设是两厂商生产的产品没有差异，平均寿命相同，备选假设是平均寿命不相

同，是双侧检验。

第二步：求秩和。

将样本混合、排列：

数据：	5	6	$\underline{6}$	7	$\underline{7}$	$\underline{8}$	$\underline{8}$	9	10	$\underline{10}$	11
秩号：	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9.5	9.5	11

以上数据下面划横线的为B厂商产品寿命。B厂商产品样本容量小，看做总体Ⅰ，$n_1=5$。A厂商产品是总体Ⅱ，$n_2=6$。总体1的秩和 $T=2.5+4.5+6.5+6.5+9.5=29.5$。

第三步：确定拒绝域。

显著水平 $\alpha=0.05$，进行双侧检验，查“秩和检验表”，$n_1=5$，$n_2=6$，得临界值 $T_1(\alpha)=20$，$T_2(\alpha)=40$。

第四步：比较秩和与临界值大小。

结果是：$20<29.5<40$，即 $T_1(\alpha)<T<T_2(\alpha)$。

第五步：判断。

样本落入接受域，所以接受原假设，样本数据证明A，B两厂商产品的寿命也是一致的。

四、游程检验

游程检验可用来检验一个样本是否是“随机”地来自总体。样本中具有某特征的单位分布得越“零乱”，越不具有规律性、倾向性，越说明样本的随机性强。我们通过“游程”概念，来反映样本分布的特征。

设 $x_1,x_2,\cdots,x_n$ 为一个样本的观测结果，并设样本中存在两种不同类型的单位，一类记为a，设有 n_1 个这种单位；另一类记为$\underline{a}$，设有 n_2 个单位。把样本按某一顺序（先后或大小）进行排列，可以得到一个由a或$\underline{a}$组成的字符串。主要的问题是，如何说明这个字符串中a与$\underline{a}$的分布是有“规律”的，还是很“零乱”的。我们不妨从分析一个具体字符串的一些典型分布情形开始。设一个由8个a和7个$\underline{a}$组成的字符串，排列后的顺序选择出下列四个典型结果：

第一种情形：a a a a a a a a $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$

第二种情形：a a a a $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$ a a a a $\underline{a}$ $\underline{a}$ $\underline{a}$

第三种情形：a $\underline{a}$ $\underline{a}$ a a a $\underline{a}$ a $\underline{a}$ $\underline{a}$ $\underline{a}$ a $\underline{a}$ a a

第四种情形：a $\underline{a}$ a $\underline{a}$ a $\underline{a}$ a $\underline{a}$ a $\underline{a}$ a $\underline{a}$ a $\underline{a}$ a

所谓游程，就是指一个重复出现的某一种类字符串片段。同一类的游程出现的次

数,称游程数。不同类游程数的和,称总游程数。例如,第一种情形中,a 的游程只有一个,它的游程数(记为 R^1)是 1;$\underline{a}$ 的游程也只有一个,其游程数(记为 R^2)也是 1,总游程数记为 R,$R=R^1+R^2=2$。第二种情形中,a 的游程有两个,它的游程数 $R^1=2$;$\underline{a}$ 的游程也有两个,其游程数 R^2 也是 2,总游程数记为 $R=R^1+R^2=4$。同理,第三种情形中,a 的游程数 $R^1=5$,$\underline{a}$ 的游程数 $R^2=4$,总游程数 $R=R^1+R^2=9$。第四种情形中,$R^1=8$;$R^2=7$;$R=15$。

显然,总游程数 R 的最小值是 2,在 n_1 与 n_2 不相等时,R 的最大值是 $2\min(n_1,n_2)+1$;在 $n_1=n_2$ 时,R 的最大值是 n_1+n_2。当 R 取得最大、最小值时,字符串的排列是最有规律的。R 取值适中时,字符串的排列就较零乱。上面的第一、第四种情形都是 a 与 $\underline{a}$ 排列得有规律情形;第二、第三种情形是较为"零乱"的排列情形。因此,R 可以作为检验样本随机性的统计量。

【例 5-10】从某校中抽出 16 个学生进行英语口试,按他们参加口试的顺序排列,成绩如下:

61 74 70 63 64 58 82 78 60 76 85 72 68 54 62 56

试以 0.05 的显著性水平,判断样本是否具有随机性。

解:

可以计算出这个成绩分布的中位数是:$(64+68)\div 2=66$

学生按其口试成绩是小于中位数还是大于中位数,被分为两类,分别记为 a 和 $\underline{a}$。按口试进行的顺序,我们得到一个字符串:

a $\underline{a}$ $\underline{a}$ a a a $\underline{a}$ $\underline{a}$ a $\underline{a}$ $\underline{a}$ $\underline{a}$ $\underline{a}$ a a a

对上述排列的字符串进行随机性检验。

第一步:作出假设。

H_0:样本是随机样本,H_1:样本不具有随机性。这是一个双侧检验。

第二步:求游程总和 R。

$R^1=4$,$R^2=3$,$R=7$。

第三步:确定拒绝域。

显著水平 $\alpha=0.05$,进行双侧检验,查"游程检验 R 临界值表",$(n_1=8,n_2=8)$,得临界值 $R_1(\alpha)=4$,$R_2(\alpha)=14$。

第四步:比较 R 与临界值大小。

结果是:$4<7<14$,即 $R_1(\alpha)<R<R_2(\alpha)$。

第五步:判断。

接受原假设,认为样本具有随机性。

上面通过查表确定检验的临界值，这个方法要求样本容量 n_1 与 n_2 都相当小。当 n_1 与 n_2 较大时(其中一个大于 20)，直接查表方法就不可行了。不过，对于大样本场合，游程总数 R 近似服从正态分布，因此，可以用正态分布统计量来确定检验的临界值。R 的均值与方差公式是：

$$E(R)=\frac{2n_1n_2}{n_1+n_2}+1 \tag{5.9}$$

$$\sigma^2(R)=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \tag{5.10}$$

【例 5－11】以前面的例 5－10 的资料，改用 z 检验来判断样本的随机性。

解：

计算游程总数的均值与标准差：$E(R)=9$；$\sigma(R)=1.93$。

构造 z 检验统计量：$z=\frac{R-E(R)}{\sigma(R)}$，$z$ 服从标准正态分布。

临界值 $z_{\alpha/2}=z_{0.025}=1.96$。

计算统计量的样本数值：$z=(7-9)/1.93=-1.036$。

$|z|<z_{\alpha/2}$，接受原假设，与例 5－10 的结论一致。

第四节 Excel 在假设检验中的运用

一、总体参数检验

Excel 中总体参数检验可利用有关的函数进行。下面用一个例子说明。

【例 5－12】某城镇去年居民家庭平均每人每月生活费收入 275 元。根据抽样调查，今年该城镇 50 户居民家庭平均每人每月生活费收入如下：

367	322	294	273	237	398	327	298	276	246
311	355	240	275	296	324	382	229	264	288
235	271	291	319	360	226	262	286	309	352
337	222	260	284	304	343	217	259	283	303
200	253	281	301	329	212	257	281	303	332

试问该城镇居民家庭平均每人每月生活费收入今年与去年比较是否明显提高($\alpha=0.05$)？

解：

设今年该城镇各居民家庭平均每人每月生活费收入用 Y 表示，其总体平均数为 μ，可设定以下假设：$H_0:\mu=275$；$H_1:\mu>275$。

分析步骤如下：

1. 输入数据。在 A2:A51 单元格中存变量 Y 的数据，变量名称“Y”放在 A1 单元格。

2. 计算 t-检验统计量。在任一个空单元格内，输入如下公式，“=(AVERAGE(Y)-275)/(STDEV(Y)/SQRT(COUNT(Y)))”。3. 计算 t 单侧临界值。在另一个空单元格内，输入如下公式，“=TINV(2*0.05,COUNT(Y)-1)”。

4. 根据上述计算结果作出判断。

二、非参数检验

在非参数检验中，Excel 比较容易实现符号检验，不易实现秩和检验与游程检验，但也可使用 Excel 进行一些辅助的运算，以提高效率。

下面以一个实例来说明 Excel 在单样本场合符号检验中的应用。

【例 5-13】本章例 5-7 的 Excel 分析步骤如下。

1. 输入数据。如图 5-3 所示，A2:A21 输入产品件数，A1 输入列标志“产品件数”。

2. 计算 n^+。在 C2 单元格输入公式“=COUNTIF(A2:A21,">160")”即可，该函数表示统计 A2:A21 单元格中大于 160 的数据点的个数。

3. 计算 n^-。在 C3 单元格输入公式“=COUNTIF(A2:A21,"<160")”。

4. 计算 n。在 C4 单元格输入公式“=C2+C3”。

	A	B	C
1	产品件数		
2	168	n^+	15
3	163	n	3
4	160	n	18
5	172	临界值	13
6	162		

图 5-3

5. 计算拒绝域的临界值。二项分布临界值可用 Excel 的分布函数求得。在 C5 单元格中输入公式“=CRITBINOM(C4,0.5,0.95)+1”即可。其中，第一个参数存放的是 n；第二个参数是一次试验中成功的概率，根据二项分布临界值表的要求，固定为 0.5；第三个参数是概率保证度的临界值，对于单侧检验，它等于 $1-\alpha$，对于双侧检验，它等于 $1-\alpha/2$。因为 CRITBINOM 返回的是使累积二项式分布概率大于等于 $1-\alpha$(或 $1-\alpha/2$)的最小值，所以根据符号检验的要求，应在上述公式中加 1。计算结果如图 5-3 所示。

6. 判断。根据以上计算出的数据可进行判断。

本章小结

1.假设检验分为两类:参数检验与非参数检验。

2.假设检验的基本概念有:小概率原理、原假设与备选假设、单侧检验与双侧检验、取伪错误与弃真错误、检验功效或检验力等。

3.统计检验的步骤如下:①确定原假设与备选假设;②构造检验统计量;③给出显著性水平,确定检验统计量的临界值与拒绝域;④根据样本数据,计算检验统计量的数值;⑤作出判断。

4.本章中出现的检验方法可归纳如表5-3所示。

表5-3

类型	检验目标	检验统计量或检验方法	服从的分布
参数检验	①σ已知条件下对正态总体均值的检验; ②大样本场合总体均值检验	$z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
	σ未知条件下对正态总体均值的检验	$t=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	$t(n-1)$
	对总体成数指标的检验	$z=\frac{P-p}{\sqrt{\frac{p(1-p)}{n}}}$	$N(0,1)$
非参数检验	①单样本场合检验总体中位数是否在某一指定位置; ②双样本场合检验两组配对样本是否有显著差异	符号检验	
	检验两独立样本是否来自同一总体,或判断不同总体是否存在显著差异	秩和检验	
	检验样本的随机性	游程检验	

5.p-值检验是通过直接计算检验统计量在样本数据下的概率,来检验原假设是否成立。该检验形式在统计分析软件中经常使用。

6.本章所介绍的假设检验的内容,可根据有关公式和Excel中的函数来进行分析。

思考与练习题

一、单项选择题：

1.将由显著性水平所规定的拒绝域平分为两部分，置于概率分布的两边，每边占显著性水平的1/2，这是(　　)。

a. 单侧检验　b. 双侧检验　c. 右单侧检验　d. 左单侧检验

2.检验功效定义为(　　)。

a. 原假设为真时将其接受的概率　b. 原假设不真时将其舍弃的概率

c. 原假设为真时将其舍弃的概率　d. 原假设不真时将其接受的概率

3.符号检验中，(+)号的个数与(-)号的个数相差较远时，意味着(　　)。

a. 存在试验误差(随机误差)　b. 存在着条件误差

c. 不存在什么误差　d. 既有抽样误差，也有条件误差

4.得出两总体的样本数据如下：

甲：8　6　10　7　8　　乙：5　11　6　9　7　10

秩和检验中，秩和最大可能值是(　　)。

a. 15　b. 48　c. 45　d. 66

二、多项选择题：

1.显著性水平与检验拒绝域的关系(　　)。

a. 显著性水平提高(α变小)，意味着拒绝域缩小

b. 显著性水平降低，意味着拒绝域扩大

c. 显著性水平提高，意味着拒绝域扩大

d. 显著性水平降低，意味着拒绝域扩大化

e. 显著性水平提高或降低，不影响拒绝域的变化

2.β错误(　　)。

a. 是在原假设不真实的条件下发生

b. 是在原假设真实的条件下发生

c. 决定于原假设与实际值之间的差距

d. 原假设与实际值之间的差距越大，犯β错误的可能性就越小

e. 原假设与实际值之间的差距越小，犯β错误的可能性就越大

三、问答题：

1.假设某产品的重量服从正态分布，现在从一批产品中随机抽取16件，测得平均重量为820克，标准差为60克，试以显著性水平$\alpha=0.01$与$\alpha=0.05$，分别检验这批产品的平均重量是否是800克。

2.某品牌某种型号的彩电规定无故障时间为10 000小时，厂家采取了改进措施，现在从新批量彩电中抽取100台，测得平均无故障时间为10 150小时，标准差为500小时，能否据此判断该种彩电无故障时间有显著增加($\alpha=0.01$)？

3.某加油站经理希望了解驾车人士在该加油站的加油习惯。在一周内，他随机地抽取100名驾车人士调查，得到如下结果：平均加油量等于13.5加仑，样本标准差是3.2加仑，有19人购买无铅汽油。

试问：

(1)设0.05的显著性水平，是否有证据说明平均加油量并非12加仑？

(2)计算(1)的p-值。

(3)设0.05的显著性水平，是否有证据说明少于20%的驾车者购买无铅汽油？

(4)计算(3)的p-值。

(5)在加油量服从正态分布的假设下，若样本容量为25，计算(1)和(2)。

4.某市全部职工中，平常订阅某种报纸的占40%，最近从订阅率来看似乎出现减少的现象，随机抽200户职工家庭进行调查，有76户职工订阅该报纸，问报纸的订阅率是否显著降低($\alpha=0.05$)？

5.某型号的汽车轮胎耐用里程按正态分布，其平均耐用里程为25 000公里。现在从某厂生产的该型号轮胎中随机取10个进行里程测试，结果数据如下：

25 400	25 600	25 300	24 900	25 500
24 800	25 000	24 800	25 200	25 700

要求：

根据以上数据，检验该厂轮胎的耐用里程是否存在显著性的差异($\alpha=0.05$)。再用p-值重新检验，比较两结论是否一致。

6.从某铁矿南北两段各抽取容量为10的样本，随机配成10对如下：

南段含铁量	28	20	4	32	8	12	16	48	8	20
北段含铁量	20	11	13	10	45	15	11	13	25	8

要求：

用符号检验法，在$\alpha=0.05$的条件下，检验“南北两段含铁量无显著差异”的假设。

7.在14对条件相同的地块上分别播下种子A和种子B，其收获量记录如表5－4，试以显著性水平$\alpha=0.05$，用秩和检验法检验两种种子的收获量是否存在显著性的差

异。

表 5-4 种子收获量记录 单位:公斤

A种子	B种子	A种子	B种子
33	48	44	34
18	17	25	37
40	24	46	47
50	22	36	13
54	38	53	27
30	41	35	30
39	20	42	25

8.某加油站有两种商标的汽油 A 和 B,某天售出的 50 桶汽油可按商标 A 和 B 排成这样的顺序:

AABAABABBAAABBABBABBABBAB

AABBBBAABABABAAABAAAAABBA

试问:在显著性水平 $\alpha=0.05$ 条件下,这一序列是否有随机性?

第六章

方差分析

★ 本章学习要点与要求 ★

本章介绍方差分析的理论与方法。通过学习，要求：①掌握方差分析的基本概念，包括因子、因子水平、单(双)因素、交互作用等；②掌握方差分解的思想，能针对单因素、双因素(有、无交互作用)的方差分析，构造出对原假设进行检验的 F 统计量；③能够利用 Excel 进行方差分析。

第一节　方差分析的基本问题

一、方差分析问题的提出

方差分析简称 ANOV(Analysis of Variance)，该统计分析方法能一次性地检验多个总体均值是否存在显著差异。前一章的假设检验主要是检验两总体的均值是否差异显著。对于多个总体均值是否差异显著的问题，如果按照每一对总体进行一次检验，显然要花费较多的时间。因此，方差分析所提供的处理方法比两两比较的处理方法要方便得多。

【例 6－1】某连锁商业企业在同城三个不同地点开了三家分店，从这三家分店随机抽 5 天的营业额资料如表 6－1 所示。

表 6-1 三家分店日营业额资料表 单位:万元

	第一家分店	第二家分店	第三家分店
第一天	10	7	14
第二天	12	11	8
第三天	9	8	12
第四天	8	13	10
第五天	11	10	11

试分析这三家分店的平均日营业额是否相同,从而确定地点因素是否对日均营业额有影响($\alpha=0.05$)。

如果把每一个分店的日营业额看成一个总体,以上问题的实质是检验这三个总体的均值是否相等:

$$H_0:\mu_1=\mu_2=\mu_3 \tag{6.1}$$

$$H_1:\mu_1,\mu_2,\mu_3 \text{ 三者不全相等} \tag{6.2}$$

其中,μ_1,μ_2,μ_3 分别为三分店的平均日营业额。

如果检验结果接受原假设,则有充分证据表明地点因素对分店的日营业额没有实质性影响;如果拒绝原假设,则有充分证据说明地点因素对日营业额有显著影响。具体情况如图 6-1 所示。

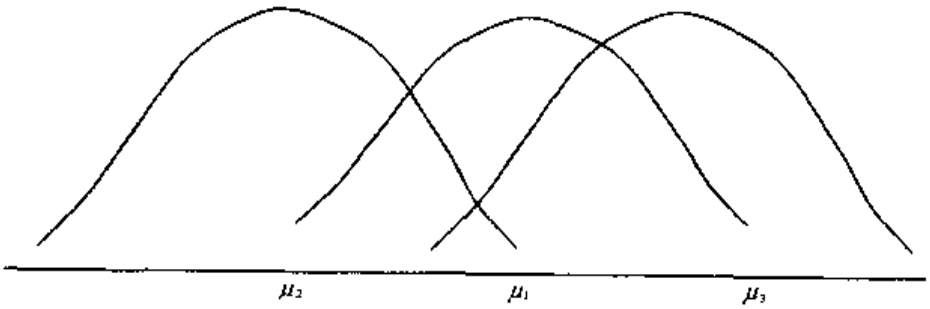


图 6-1(a)地点因素有显著影响(三个均值的排列是任意的)

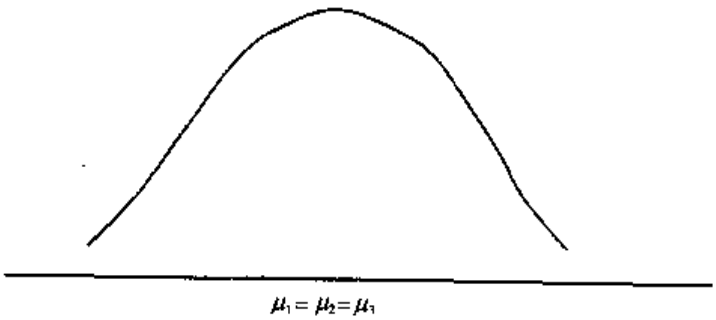


图 6-1(b) 地点因素没有显著影响

二、方差分析的基本概念

在具体学习方差分析之前，我们先介绍方差分析中的基本概念：

1.因素。因素又称因子，是在实验中或在抽样时发生变化的“量”，通常用A,B,C,……表示。例6-1中的“地点”就是一个变化的因素，这个因素是可能对分店的日均营业额产生影响的因子。方差分析的目的就是分析因子对实验或抽样的结果有无显著影响。如果在实验中变化的因素只有一个，这时的方差分析称为单因素方差分析；在实验中变化的因素不只一个时，就称多因素方差分析。双因素方差分析是多因素方差分析的最简单情形。

2.水平。因子在实验中的不同状态称做水平。如果因子 A 有 r 个不同状态，就称它有 r 个水平，可用 $A_1,A_2,\cdots,A_r$ 表示。我们要针对因素的不同水平或水平的组合，进行实验或抽取样本，以便了解因子的影响。

3.交互影响。当方差分析过程中的影响因子不惟一时，这些因子间是否独立，是否存在相互影响，是有必要注意的问题。如果因子间存在相互作用，我们称之为“交互影响”；如果因子间是相互独立的，则称为无交互影响。交互影响有时也被称为交互作用，是对实验结果产生作用的一个新因素，在分析过程中，有必要将它的影响作用也单独分离开来。

三、方差分析的原理

(一)方差的分解

从图6-1可以看出，如果地点因素对三个分店的日均营业额没有影响，则从三个分店中抽出的样本数据的波动，是由随机因素引起的；如果地点因素对分店日营业额有显著影响，则样本数据波动就有二个来源：一个是随机波动，一个是地点影响或因子影响。样本数据的波动，可通过离差平方和来反映，这个离差平方和可分解为组间方差与组内方差两部分。组间方差反映出不同的因子对样本波动的影响；组内方差则是不考虑组间方差的纯随机影响。离差平方和的分解是我们进入方差分析的“切入点”，这种方差的构成形式为我们分析现象变化提供了重要的信息。如果组间方差明显高于组内方差，说明样本数据波动的主要来源是组间方差，(地点)因子是引起波动的主要原因，可以认为因子对实验的结果存在显著的影响；反之，如果波动的主要部分来自组内方差，则因子的影响就不明显，没有充足理由认为因子对实验或抽样结果有显著作用。

(二)检验统计量

因素或因素间“交互作用”对观测结果的影响是否显著，关键要看组间方差与组内方差的比较结果。当然，产生方差的独立变量的个数对方差大小也有影响，独立变量个

数越多,方差就有可能越大;独立变量个数越少,方差就有可能越小。为了消除独立变量个数对方差大小的影响,我们用方差除以独立变量个数,得到"均方差(Mean Square)",作为不同来源方差比较的基础。引起方差的独立变量的个数,称做"自由度"。因此,我们得到检验因子影响是否显著的统计量是一个 F 统计量:

$$F=\frac{\text{组间均方差}}{\text{组内均方差}} \tag{6.3}$$

F 统计量越大,越说明组间方差是主要的方差来源,因子影响是显著的;F 越小,越说明随机方差是主要的方差来源,因子的影响不显著。

第二节 单因素方差分析

一、单因素条件下离差平方和的分解

单因素指的是因子惟一。为了检验该因子的不同水平下的总体均值是否有显著差异,我们可针对因子的不同水平进行试验或抽样;把因子处在不同水平上抽得的样本看做是来自不同总体的样本,然后检验这些不同总体的均值是否相等。在因子 A 的不同水平下抽样的结果如表 6-2 所示。

表 6-2

因素水平 \ 样本	1	2	…	n	合计	均值
A_1	X_{11}	X_{12}	…	X_{1n}	$X_{1.}$	$\bar{X}_1$
A_2	X_{21}	X_{22}	…	X_{2n}	X_2	$\bar{X}_{2.}$
⋮	⋮	⋮	…	⋮	⋮	⋮
A_r	X_{r1}	X_{r2}	…	X_{rn}	$X_{r.}$	$\bar{X}_r$
合计					$X_{..}$	X

表中的 X_{ij} 是在 A_i 水平上,第 j 个样本单位的数据。另外:

$$X_{i.}=\sum_{j=1}^{n}X_{ij},\quad X_{..}=\sum_{i=1}^{r}\sum_{j=1}^{n}X_{ij}$$
$$\bar{X}_{i.}=\frac{X_{i.}}{n},\quad \bar{X}=\frac{X_{..}}{nr}\qquad (i=1,2,\cdots,r)$$

即 $\bar{X}_{i.}$ 是在因素水平 A_i 上的平均数,$\bar{X}$ 是样本总平均。

总离差平方和记为 SST(Sum of Squares for Total),则:

$$SST=\sum_i\sum_j(X_{ij}-\bar{X})^2 \tag{6.4}$$

它反映了样本数据总的波动程度。按前面的平方和分解思路，我们有：

$$\begin{aligned}\sum_i\sum_j(X_{ij}-\bar{X})^2&=\sum_i\sum_j[(X_{ij}-\bar{X}_{i.})+(\bar{X}_{i.}-\bar{X})]^2\\&=\sum_i\sum_j(X_{ij}-X_{i.})^2+\sum_i\sum_j(\bar{X}_{i.}-\bar{X})^2+2\sum_i\sum_j(X_{ij}-\bar{X}_{i.})(\bar{X}_{i.}-\bar{X}.)\end{aligned}$$

以上式中，交叉项之和为零，即：

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{i.})(\bar{X}_{i.}-\bar{X})&=\sum_{i=1}^{r}(\bar{X}_{i.}-\bar{X})\sum_{j=1}^{n}(X_{ij}-\bar{X}_{i.})\\&=\sum_{i=1}^{r}(\bar{X}_{i.}-\bar{X})\times 0=0\end{aligned}$$

进一步记：

$$SSA=\sum_i\sum_j(\bar{X}_{i.}-\bar{X})^2=\sum_i n(\bar{X}_{i.}-\bar{X})^2 \tag{6.5}$$

$$SSE=\sum_i\sum_j(X_{ij}-\bar{X}_{i.})^2 \tag{6.6}$$

SSA(Sum of Squares Among Groups)表示的是组间方差总和，是由各组均值差异引起的；SSE(Sum of Squares for Error)表示的是组内方差部分，由随机误差产生。因此，我们得到离差平方和的分解式：

$$SST=SSE+SSA \tag{6.7}$$

二、因素作用显著性的检验

前面已指出，检验因子作用的显著性，实质上就是检验以下的假设：

$$H_0:\mu_1=\mu_2=\cdots=\mu_r \tag{6.8}$$

$$H_1:\mu_1,\mu_2,\cdots,\mu_r\text{ 不全相等} \tag{6.9}$$

原假设是否为真，关键是看 SSA 与 SSE 两者间的相对比较；在两者比较时，我们要剔除引起各种方差的“独立变量个数”——自由度的影响。

SST 是由于 X_{ij} 的波动引起的方差，但是，这里所有的 nr 个变量并不独立，它们必须满足的一个约束条件是：$\sum_i\sum_j(X_{ij}-\bar{X})=0$，真正独立的变量只有 $nr-1$ 个，自由度是 $nr-1$。SSA 是因子在不同水平上的均值 $\bar{X}_{i.}$ 变化而产生的方差。但是，r 个均值 $\bar{X}_i$ 并不是独立的，它们必须满足约束条件：$\sum n(\bar{X}_{i.}-\bar{X})=0$，因此也丢失一个自由度，它的自由度是 $r-1$。SSE 是由所有的 X_{ij} 在各因素水平上围绕均值波动产生的，但是它们必须满足的约束条件是：$\sum(X_{ij}-\bar{X}_{i.})=0(i=1,2,\cdots,r)$，一共 r 个，失去了 r 个自由度，所以 SSE 的自由度是 $nr-r$。SST，SSA 和 SSE 的自由度满足如下关系：

$$nr-1=(r-1)+(nr-r) \tag{6.10}$$

将各方差除以各自的自由度，就得到相应的均方差，即：

$$MSA=\frac{SSA}{r-1};MSE=\frac{SSE}{nr-r} \tag{6.11}$$

MSA 与 MSE 分别是 SSA 与 SSE 的均方差。

有了均方差，我们就可以构造 F 统计量来检验原假设式(6.8)是否为真：

$$F=\frac{MSA}{MSE}\sim F(r-1,nr-r) \tag{6.12}$$

F 值越大，越说明总的方差波动中，组间方差是主要部分，有利于拒绝原假设接受备选假设；反之，F 值越小，越说明随机方差是主要的方差来源，有利于接受原假设，没有充分证据说明待检验的因素对总体波动有显著影响。因此，检验的拒绝域安排在右侧。对于给出的显著性水平 α，查 F 分布表得临界值 $F_\alpha(r-1,nr-r)$，当 $F>F_\alpha$ 时，拒绝原假设，接受备选假设，认为所检验因素对总体有显著影响；当 $F<F_\alpha$ 时，接受原假设，认为没有证据说明所检验的因素对总体有显著影响。如图 6－2 所示。

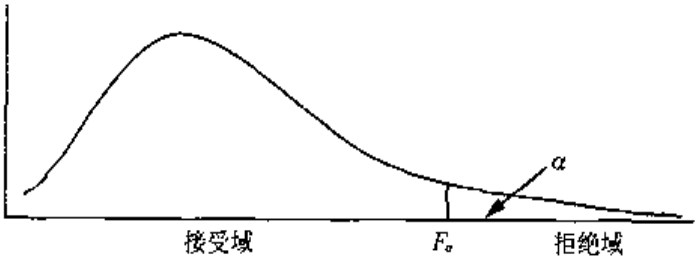


图 6－2　F 检验示意图

现在，我们回头去解决例 6－1 中提出的问题。通过 Excel 得到方差分析的计算结果如表 6－3 所示(参见本章第四节)。

表 6－3　Excel 输出的方差分析表

差异源	SS	df	MS	F	P-value	F crit
组间	4.133 333	2	2.066 667	0.469 697	0.636 215	3.885 29
组内	52.8	12	4.4			
总计	56.933 33	14				

由上表可得：$SST=56.93$，$SSA=4.13$，$SSE=52.8$，$MSA=2.067$，$MSE=4.4$，样本的统计量 $F=2.067/4.4=0.469\ 7$，分析表给出了临界值是 $F_\alpha=3.885$。(显著性水平 $\alpha=0.05$)，$F<F_\alpha$，接受 H_0，即没有充分证据说明三个分店的地点不同对日均营业额产生了影响。如果直接从 P-值进行判断，由于 P-值 $=0.636\ 215>0.05$，结论也是接受原假设。注意：这里的 P-值是通过 F 统计量计算的。

三、单因素方差分析中的几个问题

使用单因素方差分析时，应注意的问题主要是：

第一，方差分析需满足的假设条件。方差分析实质上是对各总体均值相等假设进行检验，为了得到检验统计量的精确分布，要求满足的前提条件是：①样本是独立的随机样本；②各样本皆来自正态总体；③总体方差具有齐性，即各总体方差相等。满足以上的条件，式(6.12)才满足第一个自由度为 $r-1$，第二个自由度为 $nr-r$ 的 F 分布。一般情况下，我们总认为以上的假定条件都是满足的或近似满足的。这一点对多因素分析也有效。

第二，在前面的数据结构中，因素各水平下的样本容量都相等，是为了公式推导的方便。在实际问题中，各总体的样本容量可以相等也可以不等，分析过程和结论都不受影响。在用 Excel 进行方差分析时，计算机也能自动判断各样本的容量大小。

第三，方差分析将所有样本结合在一起，使数据数量增多，提高了分析结果的稳定性(Robustness)。但是，方差分析也存在自身的不足之处，比如，当检验结果是拒绝原假设时，我们认为各总体的均值不等，至于哪个总体均值大，哪个总体均值小，方差分析本身不能立即得出结论。如果要得到各总体均值大小的排序信息，还需要借助"多重比较"法，限于篇幅，这里不作介绍，有兴趣的读者可参考其他资料。[①]

第三节 双因素方差分析

一、无交互作用下的方差分析

设 A 与 B 是可能对试验结果有影响的两个因素，相互独立，无交互作用。设在双因素各种水平的组合下进行试验或抽样，得数据结构如表 6-4 所示。

表中每行的均值 $\bar{X}_{i.}(i=1,2,\cdots,r)$ 是在因素 A 的各个水平上试验结果的平均数；每列的均值 $\bar{X}_{.j}(j=1,2,\cdots,n)$ 是在因素 B 的各种水平上试验的平均数。以上数据的离差平方和分解形式为：

$$SST = SSA + SSB + SSE \tag{6.13}$$

① 参见：王明华. 现代应用统计分析方法. 北京：中国统计出版社，1999，61～70

表 6－4

		因素 B				均值
		B_1	B_2	…	B_n	
因素 A	A_1	X_{11}	X_{12}	…	X_{1n}	$\bar{X}_{1.}$
	A_2	X_{21}	X_{22}	…	X_{2n}	$\bar{X}_2$
	⋮	⋮	⋮	…	⋮	⋮
	A_r	X_{r1}	X_{r2}	…	X_{rn}	$\bar{X}_r$
均值		$\bar{X}_{.1}$	$\bar{X}_{.2}$	…	$\bar{X}_{.n}$	$\bar{X}$

上式中：

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2 \tag{6.14}$$

$$SSA=\sum_{i}\sum_{j}(\bar{X}_{i.}-\bar{X})^2=\sum_{i}n(\bar{X}_{i.}-\bar{X})^2 \tag{6.15}$$

$$SSB=\sum_{i}\sum_{j}(\bar{X}_{.j}-\bar{X})^2=\sum_{j}r(X_{.j}-X)^2 \tag{6.16}$$

$$SSE=\sum_{i}\sum_{j}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X})^2 \tag{6.17}$$

SSA 表示的是因素 A 的组间方差总和，SSB 是因素 B 的组间方差总和，都是由各因素在不同水平下各自的均值差异引起的；SSE 仍是组内方差部分，由随机误差产生。各个方差的自由度是：SST 的自由度为 $nr-1$，SSA 的自由度为 $r-1$，SSB 的自由度为 $n-1$，SSE 的自由度为 $nr-r-n-1=(r-1)(n-1)$。

各个方差对应的均方差是：

对因素 A 而言，

$$MSA=\frac{SSA}{r-1} \tag{6.18}$$

对因素 B 而言，

$$MSB=\frac{SSB}{n-1} \tag{6.19}$$

对随机误差项而言，

$$MSE=\frac{SSE}{nr-r-n-1} \tag{6.20}$$

我们得到检验因素 A 与 B 影响是否显著的统计量分别是：

$$F_A=\frac{MSA}{MSE}\sim F[r-1,(r-1)(n-1)] \tag{6.21}$$

$$F_B=\frac{MSB}{MSE}\sim F[n-1,(r-1)(n-1)] \tag{6.22}$$

【例 6－2】某企业有三台不同型号的设备，生产同一产品，现有五名工人轮流在此

三台设备上操作，记录下他们的日产量如表6－5所示。试根据方差分析说明这三台设备之间和五名工人之间对日产量的影响是否显著($\alpha=0.05$)?

表6－5

	工人1	工人2	工人3	工人4	工人5
设备A	64	72	63	81	78
设备B	75	66	61	73	80
设备C	78	67	80	69	71

解：

检验的假设有两个，第一个假设是针对设备(设为A因素)的：

H_{01}：三台设备对日产量没有显著影响；

H_{11}：三台设备对日产量有显著影响。

第二个假设是针对人员(设为B因素)的：

H_{02}：工人技术对日产量没有显著影响；

H_{12}：工人技术对日产量有显著影响。

将以上数据输入Excel表格中，进行“无重复双因素分析”，Excel输出的方差分析表如表6－6所示。

表6－6 Excel输出的方差分析表

差异源	SS	df	MS	F	P-value	F crit
行(A因素)	10.533 33	2	5.266 667	0.092 371	0.912 73	4.458 968
列(B因素)	161.066 7	4	40.266 67	0.706 226	0.609 69	3.837 854
误差	456.133 3	8	57.016 67			
总计	627.733 3	14				

从表6－6可知：$F_A=0.092<F_{0.05}(2,8)=4.46$，接受$H_{01}$，没有证据证明三台设备对日产量有显著影响；$F_B=0.706<F_{0.05}(4,8)=3.84$，接受$H_{02}$，也没有证据证明五名工人的技术对日产量有显著影响。

二、有交互作用的方差分析

为了研究两个因素是否独立，有无交互作用，我们需要在各个因素水平组合下，进行重复试验；因此，有交互作用时，方差分析的数据结构不同于无交互作用的情形。设因素A与因素B每一对水平搭配下重复试验的次数都是m，得到试验数据结构如表6－7。

表 6－7

		因素 B			
		B_1	B_2	…	B_n
因素 A	A_1	X_{111} X_{112} ⋮ X_{11m}	X_{121} X_{122} ⋮ X_{12m}	…	X_{1n1} X_{1n2} ⋮ X_{1nm}
	A_2	X_{211} X_{212} ⋮ X_{21m}	X_{221} X_{222} ⋮ X_{22m}	…	X_{2n1} X_{2n2} ⋮ X_{2nm}
	⋮	⋮	⋮	…	⋮
	A_r	X_{r11} X_{r12} ⋮ X_{r1m}	X_{r21} X_{r22} ⋮ X_{r2m}	…	X_{rn1} X_{rn2} ⋮ X_{rnm}

表 6－7 中的 X_{ijl}表示的是在因素水平组合(A_i, B_j)下第 l 次试验的结果。在此组合下试验结果的平均值为：

$$\bar{X}_{ij.} = \frac{1}{m}\sum_{l=1}^{m} X_{ijl} \tag{6.23}$$

进一步记：

$$\bar{X}_{i} = \frac{1}{nm}\sum_{j=1}^{n}\sum_{l=1}^{m} X_{ijl} \tag{6.24}$$

$$X_{.j.} = \frac{1}{rm}\sum_{i=1}^{r}\sum_{l=1}^{m} X_{ijl} \tag{6.25}$$

$$\bar{X} = \frac{1}{rnm}\sum_{i}\sum_{j}\sum_{l} X_{ijl} \tag{6.26}$$

则我们类似地有以下的离差平方和分解形式：

$$SST = SSA + SSB + SSAB + SSE \tag{6.27}$$

式中，

$$SST = \sum_{i}\sum_{j}\sum_{l}(X_{ijl} - \bar{X})^2 \tag{6.28}$$

$$SSA = nm\sum_{i}(\bar{X}_i - \bar{X})^2 \tag{6.29}$$

$$SSB = rm\sum_{j}(X_{.j.} - \bar{X})^2 \tag{6.30}$$

$$SSAB = m\sum_i\sum_j(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X})^2 \tag{6.31}$$

$$SSE = \sum_i\sum_j\sum_l(X_{ijl} - \bar{X}_{ij})^2 \tag{6.32}$$

与无交互作用的双因素方差分解相比，这里多出了一项 $SSAB$，它刚好反映了两个因素交互作用的结果。离差平方和 SST，SSA，SSB，$SSAB$ 和 SSE 的自由度分别是 $rnm-1$，$r-1$，$n-1$，$(r-1)(n-1)$和 $rn(m-1)$。我们得到如下的均方差：

$$MSA = \frac{SSA}{r-1} \tag{6.33}$$

$$MSB = \frac{SSB}{n-1} \tag{6.34}$$

$$MSAB = \frac{SSAB}{(r-1)(n-1)} \tag{6.35}$$

$$MSE = \frac{SSE}{rn(m-1)} \tag{6.36}$$

则检验因素 A 与 B 影响是否显著的统计量分别是：

$$F_A = \frac{MSA}{MSE} \sim F(r-1, rnm-rn) \tag{6.37}$$

$$F_B = \frac{MSB}{MSE} \sim F(n-1, rnm-rn) \tag{6.38}$$

检验交互影响是否显著的统计量是：

$$F_{AB} = \frac{MSAB}{MSE} \sim F[(r-1)(n-1), rnm-rn] \tag{6.39}$$

【例 6－3】为了分析光照因素 A 与噪音因素 B 对工人生产有无影响，光照效应与噪音效应有交互作用，在此两因素不同的水平组合下做试验，结果如表 6－8(表中数据为产量)所示。

表 6－8

		因素 B								
		B_1			B_2			B_3		
因素 A	A_1	15	15	17	19	19	16	16	18	21
	A_2	17	17	17	15	15	15	19	22	22
	A_3	15	17	16	18	17	16	18	18	18
	A_4	18	20	20	15	16	17	17	17	17

解：

检验的假设有三个：

H_{01}：光照因素 A 对产量没有显著影响；

H_{11}：光照因素 A 对产量有显著影响；

H_{02}:噪音因素 B 对产量没有显著影响;

H_{12}:噪音因素 B 对产量有显著影响;

H_{03}:光照效应与噪音效应没有交互作用;

H_{13}:光照效应与噪音效应有交互作用。

将以上数据输入 Excel 表格中,进行"有重复双因素分析",Excel 输出的方差分析如表 6-9 所示。

表 6-9 方差分析表

差异源	SS	D_f	MS	F	P-value	F crit
样本(B因素)	28.388 889	2	14.194 4	9.462 96	0.000 93	3.402 83
列(A因素)	2.083 333 3	3	0.694 44	0.462 96	0.710 77	3.008 79
交互	63.833 333	6	10.638 9	7.092 59	0.000 2	2.508 19
内部	36	24	1.5			
总计	130.305 56	35				

从表 6-9 可知:$F_A=0.462\ 96<F_{0.05}(3,24)=3.\ 008\ 79$,接受 H_{01},没有充分证据说明光照对产量有显著影响;$F_B=9.462\ 96>F_{0.05}(2,24)=3.402\ 83$,拒绝 H_{02},有充分证据说明噪音对产量有显著影响;$F_{AB}=7.092\ 59>F_{0.05}(6,24)=2.508\ 19$,拒绝 H_{03},有充分证据说明光照与噪音存在交互作用并由此对产量产生显著影响。

第四节 Excel 在方差分析中的运用

一、单因素方差分析

在 Excel 中,可以使用"方差分析:单因素方差分析"工具来完成单因素方差分析。

【例 6-4】本章例 6-1 的 Excel 统计分析步骤如下。

第一,输入数据。如图 6-3 所示。

第二,调出"方差分析:单因素方差分析"对话框,其主要选项含义如下。

输入区域:在此输入待分析数据区域的单元格引用。该引用必须由两个或两个以上按列或行组织的相邻数据区域组成。本例为"$B$1:$D$6"。

分组方式:如果需要指出输入区域中的数据是按行还是按列排列,请单击"行"或"列"。本例分组方式为"列"。

标志位于第一行/列：如果输入区域的第一行中包含标志项，请选中“标志位于第一行”复选框；如果输入区域的第一列中包含标志项，请选中“标志位于第一列”复选框；如果输入区域没有标志项，则该复选框不被选中，Excel 将在输出表中生成适宜的数据标志。

图 6－3

α：在此输入计算 F 统计临界值的置信度。本例为 0.05。

本例对话框的填写如图 6－3 所示。图中分组方式为“列”方式，因为三家分店的日营业额是按列排列的，即分别排在 B,C,D 列。

第三，单击“确定”按钮，可得方差分析表(结果已在本章第二节的表 6－3 中给出)。

二、无交互作用下的双因素方差分析

该项工作可利用“方差分析：无重复双因素方差分析”工具来完成。

【例 6－5】本章例 6－2 的 Excel 统计分析步骤如下：

第一，输入数据。如图 6－4 所示。

第二，调出“方差分析：无重复双因素分析”对话框，填写如图 6－4 所示。该工具对话框设置与单因素方差分析类似。要注意，本例中“标志”复选框被选中，输入区域必须包括 A 因素与 B 因素的水平标志(如“工人 1”、“工人 4”、“设备 B”等)所在的单元格区域，也即输入区域为“＄M＄1：＄R＄4”，而不是只包括数据的单元格区域“＄N＄2：＄R＄4”。

第三，单击“确定”按钮，得到方差分析表(结果已在本章第三节表6－6中给出)。

图6－4

三、有交互作用的双因素方差分析

该项工作可以使用“方差分析：可重复双因素方差分析”工具来完成。

	A	B	C	D	E
1		A1	A2	A3	A4
2	B1	15	17	15	18
3		15	17	17	20
4		17	17	16	20
5	B2	19	15	18	15
6		19	15	17	16
7		16	15	16	17
8	B3	16	19	18	17
9		18	22	18	17
10		21	22	18	17

图6－5

【例6－6】本章例6－3的Excel统计分析步骤如下：

第一，输入数据，如图6－5所示。其中，B2:B4单元格存放的是在“A1”与“B1”因素水平共同作用下，进行3次试验所得的结果；D5:D7单元格存放的是在“A3”与“B2”因素水平共同作用下，进行3次试验所得的结果，其余类推。

第二，调出“方差分析：可重复双因素分析”对话框，其填写如图6－6所示。该分析工具对话框与单因素方差分析对话框基本相同，只是多了一个“每一样本的行数”编辑框，其中输入包含在每个样本中的行数。本例中，在每种不同因素水平组合下，分别进行了3次试验，因此“每一样本的行数”为“3”。每个样本必须包含同样的行数。另外，在该分析工具对话框中去掉了“标志位于第一行”复选框，但要注意输入区域必须包括因素水平标志(“A1”、“B2”等)所在的单元格区域，也即，输入区域为“$A$1:$E$10”，而不是只包括数据的单元格区域“$B$2:$E$10”。

第三，单击“确定”按钮，得到方差分析表(结果已在本章第三节表6－9中给出)。

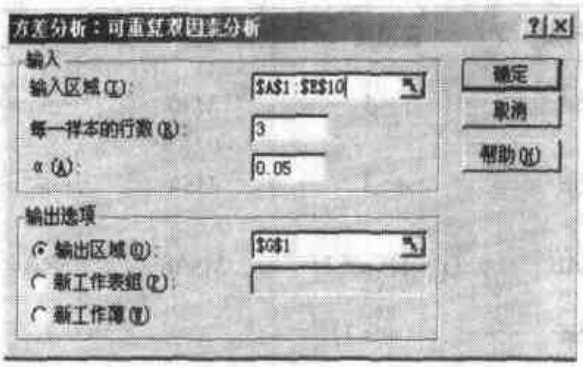


图 6-6

本章小结

1.方差分析的基本概念:在实验或抽样中引起变量变化的因素称为因子,因子的不同取值称为因子的水平;方差分析按因子个数分为:单因子方差分析与多因子方差分析;在多因子场合,还要考虑因子间的交互作用。

2.单因子方差分析表如表 6-10 所示。

表 6-10

差异源	平方和	自由度	均方	F 统计量	$P_r>F$
组间	SSA	$r-1$	MSA	$F=\frac{MSA}{MSE}$	
组内	SSE	$nr-r$	MSE		
总计	SST	$nr-1$			

3.双因子无交互作用方差分析表如表 6-11 所示。

表 6-11

差异源	平方和	自由度	均方	F 统计量	$P_r>F$
A 因素	SSA	$r-1$	MSA	$F_A=\frac{MSA}{MSE}$	
B 因素	SSB	$n-1$	MSB	$F_B=\frac{MSB}{MSE}$	
误差	SSE	$(r-1)(n-1)$	MSE		
总计	SST	$nr-1$			

4.双因子有交互作用方差分析表如表 6-12 所示。

表 6－12

差异源	平方和	自由度	均方	F 统计量	$Pr>F$
A 因素	SSA	$r-1$	MSB	$F_A=\frac{MSA}{MSE}$	
B 因素	SSB	$n-1$	MSA	$F_B=\frac{MSB}{MSE}$	
交互作用	$SSAB$	$(r-1)(n-1)$	$MSAB$	$F_{AB}=\frac{MSAB}{MSE}$	
内部	SSE	$rn(m-1)$			
总计	SST	$rnm-1$			

5. 从以上的方差分析表中的 F 统计量或 $Pr>F$ 的 P-值就可以方便地对原假设进行判断。

6. Excel 中方差分析工具有三种：单因素方差分析、可重复双因素分析和无重复双因素分析。利用这些工具可以方便地进行分析。

思考与练习题

一、单项选择题：

1. 利用"方差分析表"进行方差分析时，该表不包括的项目有(　　)。

a. 方差来源　　b. 离差平方和及其分解

c. 各离差平方和的自由度　　d. 原假设的统计判断

2. 以下对方差分析叙述不正确的是(　　)。

a. 方差分析可以对若干平均值是否相等同时进行检验

b. 进行方差分析要求各水平下的样本容量相同

c. 离差平方和能分解为组内方差与组间方差的和

d. 方差分析方法在社会科学领域也大有用武之地

3. 双因素方差分析有两种类型：一个是有交互作用的，一个是无交互作用的。区别的关键是看这对因子(　　)。

a. 是否独立　　b. 是否都服从正态分布

c. 是否因子的水平相同　　d. 是否有相同的自由度

二、多项选择题：

1. 方差分析针对不同情况可分为(　　)。

a. 单因素方差分析　　b. 多因素方差分析

c. 双因素方差分析　　d. 双因素无交互影响方差分析

e. 双因素有交互影响方差分析

2. 对方差分析的基本原理描述正确的有(　　)。

a. 通过方差的比较,检验各因子水平下的均值是否相等

b. 方差比较之前应消除自由度的影响

c. 方差比较的统计量是 F 统计量

d. 方差分析的实质是对总体均值的统计检验

e. 方差分析的因子只能是定量的,不然就无从进行量化分析

三、问答题:

1. 就双因素有交互影响的方差分析,说明与各离差平方和相对应的自由度各为多少?

2. 某农场为了比较四种不同的肥料对农作物产量的影响,进行了试验并得到如表6-13所示数据。

表6-13

肥料种类	收获量 x				平均收获量
A	98	96	91	66	87.75
B	60	69	50	35	53.50
C	79	64	81	70	73.50
D	90	70	79	88	81.75

试据以上数据推断:肥料对该农作物的收获量有无显著作用?(显著性水平取0.05。)

3. 农田试验中考察三个品种对产量 y 有无影响。试验时将试验田分成三个区组,每个区组又分为三块区田,并将三种品种随机地安排在每个区组的小田上种植。结果如表6-14所示。

表6-14

品种 \ 区田	1	2	3
A	32.6	42.7	35.3
B	36.4	47.1	40.1
C	29.5	32.9	33.6

要求:

以0.05的显著性水平,检验品种对产量有无显著影响。

4. 某厂商在五个地区销售自己生产的产品。包装部门欲了解不同的外包装对各地

消费者的消费有无影响，要求销售部门提供销售资料如表 6－15 所示。

表 6－15

地区＼包装	A1	A2	A3
B1	41	45	34
B2	53	51	44
B3	54	48	46
B4	55	43	45
B5	43	39	51

得到数据资料后，包装设计部门以 0.05 的显著性水平，对该厂产品的包装不同与地区的不同是否影响销售额进行检验，结论应如何？

第七章

相关与回归分析

★ 本章学习要点与要求 ★

本章学习相关与回归分析。具体要求:①掌握有关相关与回归分析的基本概念;②掌握相关系数的计算与检验的方法,理解标准的一元线性回归模型,能够对模型进行估计和检验并利用模型进行预测;③理解标准的多元线性回归模型,掌握估计、检验的基本方法和预测的基本公式,理解复相关系数和偏相关系数及其与单相关系数的区别;④了解常用的非线性函数的特点,掌握常用的非线性函数线性变换以及估计和检验的方法;⑤能够应用 Excel 软件进行相关与回归分析。

第一节 相关与回归分析的基本概念

一、函数关系与相关关系

客观现象总是普遍联系和相互依存的。客观现象之间的数量联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。

当一个或几个变量取一定的值时,另一个变量有确定的值与之相对应,我们称这种关系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 以及该商品价格 P 之间的关系可用 $Y=P\cdot X$ 表示,这就是一种函数关系。一般把作为影响因素的变量称为自变量,把发生对应变化的变量称为因变量。在上例中,Y 是因变量,

P 与 X 则是自变量。

当一个或几个相互联系的变量取一定数值时，与之相对应的另一变量的值虽然不确定，但它仍按某种规律在一定的范围内变化，变量间的这种相互关系，称为具有不确定性的相关关系。例如，劳动生产率与工资水平的关系、投资额和国民收入的关系、商品流转规模与流通费用的关系等等都属于相关关系。

变量之间的函数关系和相关关系，在一定条件下是可以互相转化的。本来具有函数关系的变量，当存在观测误差时，其函数关系往往以相关的形式表现出来。而具有相关关系的变量之间的联系，如果我们对它们有了深刻的规律性认识，并且能够把影响因变量变动的因素全部纳入方程，这时的相关关系也可能转化为函数关系。另外，相关关系也具有某种变动规律性，所以，相关关系经常可以用一定的函数形式去近似地描述。客观现象的函数关系可以用数学分析的方法去研究，而研究客观现象的相关关系则必须借助于统计学中的相关与回归分析方法。

二、相关关系的种类

客观现象的相关关系可以按不同的标志加以区分。

（一）按相关的程度区分

按相关的程度可分为完全相关、不完全相关和不相关。当一个现象的数量变化完全由另一个现象的数量变化所确定时，称这两种现象间的关系为完全相关。例如，在价格不变的条件下，某种商品的销售总额与其销售量总是成正比例关系。在这种场合，相关关系便成为函数关系，因此也可以说函数关系是相关关系的一个特例。当两个现象彼此互不影响，其数量变化各自独立时，称为不相关现象。例如，通常认为股票价格的高低与气温的高低是不相关的。两个现象之间的关系介于完全相关和不相关之间，称为不完全相关，一般的相关现象都是指这种不完全相关。

（二）按相关的方向区分

按相关的方向可分为正相关和负相关。当一个现象的数量增加（或减少），另一个现象的数量也随之增加（或减少）时，称为正相关。例如，消费水平随收入的增加而提高。当一个现象的数量增加（或减少），而另一个现象的数量向相反方向变动时，称为负相关。例如商品流转的规模愈大，流通费用水平则愈低。

（三）按相关的形式区分

按相关的形式可分为线性相关和非线性相关。当两种相关现象之间的关系大致呈现为线性关系时，称之为线性相关。例如人均消费水平与人均收入水平通常成线性关系。如果两种相关现象之间，并不表现为直线的关系，而是近似于某种曲线方程的关系，则这种相关关系称为非线性相关。例如产品的平均成本与产品总产量就是一种非

线性相关。

(四)按研究的变量区分

按所研究的变量多少可分为单相关、复相关和偏相关。两个变量之间的相关,称为单相关。当所研究的是一个变量对两个或两个以上其他变量的相关关系时,称为复相关。例如,某种商品的需求与其价格水平以及收入水平之间的相关关系便是一种复相关。在某一现象与多种现象相关的场合,假定其他变量不变,专门考察其中两个变量的相关关系称为偏相关。例如,在假定人们的收入水平不变的条件下,某种商品的需求与其价格水平的关系就是一种偏相关。

三、相关分析与回归分析

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。所谓相关分析,就是用一个指标来表明现象间相互依存关系的密切程度。所谓回归分析,就是根据相关关系的具体形态,选择一个合适的数学模型,来近似地表达变量间的平均变化关系。

相关分析和回归分析有着密切的联系,它们不仅具有共同的研究对象,而且在具体应用时,常常必须互相补充。相关分析需要依靠回归分析来表明现象数量相关的具体形式,而回归分析则需要依靠相关分析来表明现象数量变化的相关程度。只有当变量之间存在着高度相关时,进行回归分析寻求其相关的具体形式才有意义。由于上述原因,回归分析和相关分析在一些统计学的书籍中被合称为相关关系分析。

但是,应当指出相关分析与回归分析之间在研究目的和方法上是有明显区别的。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,确定一个相关的数学表达式,根据这个数学方程式可以从已知量来推测未知量,从而为估算和预测提供一个重要的方法。因此,相关分析可以不必确定变量中哪个是自变量,哪个是因变量,其所涉及的变量可以都是随机变量。而回归分析则必须事先研究确定具有相关关系的变量中哪个为自变量,哪个为因变量。一般地说,回归分析中因变量是随机的,而把自变量作为研究时给定的非随机变量。

相关与回归分析可以加深人们对客观现象之间相关关系的认识,因而是对客观现象进行分析的有效方法。但是,相关分析和回归分析只是定量分析的手段。通过相关与回归分析虽然可以从数量上反映现象之间的联系形式及其密切程度,但是无法准确地判断现象内在联系的有无,也无法单独以此来确定何种现象为因,何种现象为果。只有以实质性科学理论为指导,并结合实际经验进行分析研究,才能正确判断事物的内在联系和因果关系。对没有内在联系的事物进行相关和回归分析,不但没有意义,反而会

导致荒谬的结论。因此,在应用这两种方法对客观现象进行研究时,一定要始终注意把定性分析和定量分析结合起来,在定性分析的基础上开展定量分析。

对于相关分析和回归分析,既可以从描述统计的角度也可以从推断统计的角度来说明。本章采用了后一种方式,这不仅是因为后者可以将前者的内容包含在内,而且还因为在有关现实经济和管理问题的定量分析中,作为推断统计的相关与回归分析更加具有广泛的应用价值。

四、相关图

相关图又称散点图。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。相关图是研究相关关系的直观工具,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、形式和密切程度作大致的判断。

例如,我国城镇居民"十五"期间人均收入和人均消费支出资料如表 7-1 所示。

表 7-1 我国城镇居民人均消费和收入情况 单位:元

	1996 年	1997 年	1998 年	1999 年	2000 年
人均消费支出	4 845	5 189	5 458	5 889	6 317
人均收入	3 920	4 186	4 332	4 616	4 998

资料来源:国家统计局.中国统计摘要.北京:中国统计出版社,2001

根据表 7-1 的资料绘制的相关图,如图 7-1 所示。

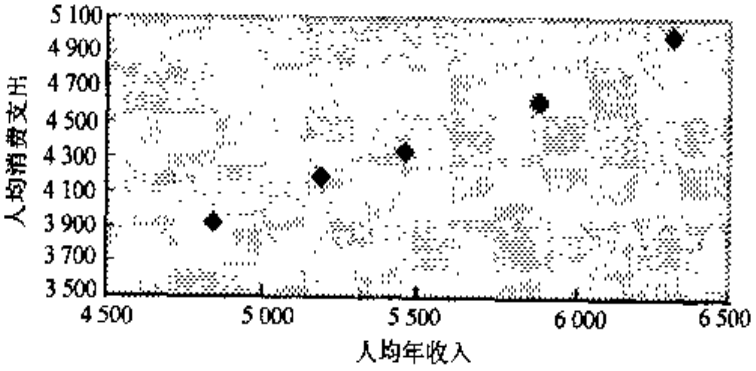


图 7-1 居民消费支出与收入的相关图

从图 7-1 可以看出,我国居民的人均消费支出和人均收入之间呈现正线性相关关系。

第二节 简单线性相关与回归分析

一、相关系数及其检验

(一)相关系数的定义

单相关分析是对两个变量之间的线性相关程度进行分析。单相关分析所采用的尺度为单相关系数,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:

$$\rho=\frac{Cov(X,Y)}{Var(X)Var(Y)} \tag{7.1}$$

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:

$$r=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} \tag{7.2}$$

式中,$\bar{X}$ 和 $\bar{Y}$ 分别是 X 和 Y 的样本平均数。

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。容易证明,样本相关系数是总体相关系数的一致估计量。

(二)相关系数的特点

样本相关系数 r 有以下特点:

1. r 的取值介于 -1 与 1 之间。

2. 当 $r=0$ 时,X 与 Y 的样本观测值之间没有线性关系。

3. 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,X 与 Y 为正相关,当 $r<0$ 时,X 与 Y 为负相关。

4. 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。

5. r 是对变量之间线性相关关系的度量。$r=0$ 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关系。对于二者之间可能存在的非线性相关关系,需要利用其他指标去进行分析。关于这一问题,我们将在后面作进一

步讨论。

(三)相关系数的计算

具体计算样本相关系数时,通常利用以下公式:

$$r=\frac{n\sum X_tY_t-\sum X_t\sum Y_t}{\sqrt{(n\sum X_t^2-(\sum X_t)^2)(n\sum Y_t^2-(\sum Y_t)^2)}} \tag{7.3}$$

上式是由样本相关系数的定义式推导而来的。

【例7-1】2000年我国部分省市自治区城镇居民人均消费性支出和人均可支配收入的有关资料如表7-2所示。试利用表7-2中的数据计算消费性支出与可支配收入的样本相关系数。

表7-2 2000年我国部分省市城镇居民人均消费支出和收入情况

	可支配收入 X(元)	消费支出 Y(元)	X^2	Y^2	XY
北京	10 350	8 494	107 122 500	72 148 036	87 912 900
天津	8 141	6 121	66 275 881	37 466 641	49 831 061
河北	5 661	4 349	32 046 921	18 913 801	24 619 689
山西	4 724	3 942	22 316 176	15 539 364	18 622 008
内蒙古	5 129	3 928	26 306 641	15 429 184	20 146 712
辽宁	5 358	4 356	28 708 164	18 974 736	23 339 448
吉林	4 810	4 021	23 136 100	16 168 441	19 341 010
黑龙江	4 913	3 824	24 137 569	14 622 976	18 787 312
上海	11 718	8 868	137 311 524	78 641 424	103 915 224
江苏	6 800	5 323	46 240 000	28 334 329	36 196 400
浙江	9 279	7 020	86 099 841	49 280 400	65 138 580
安徽	5 294	4 233	28 026 436	17 918 289	22 409 502
福建	7 432	5 639	55 234 624	31 798 321	41 909 048
合计	89 609	70 118	682 962 377	415 235 942	532 168 894

资料来源:国家统计局.中国统计摘要.北京:中国统计出版社,2001

解:

将表7-2中的有关数据代入(7.3)式,可得:

$$r=\frac{13\times 532\ 168\ 894-89\ 609\times 70\ 118}{\sqrt{(13\times 682\ 962\ 377-89\ 609^2)(13\times 415\ 235\ 942-70\ 118^2)}}=0.993\ 3$$

(四)相关系数的检验

在实际的客观现象分析研究中,相关系数一般都是利用样本数据计算的,因而带有一定的随机性。样本容量越小其可信程度就越差,因此也需要进行检验。相关系数的显著性检验问题可分为两类: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。限于篇幅,这里只介绍对总体相关系数 ρ 是否等于0进行检验。

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

首先,计算相关系数 r 的 t 值:

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.4}$$

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

【例7-2】假设根据6对样本观测数据计算出某公司的股票价格与气温的样本相关系数 $r=0.50$,试问是否可以根据5%的显著水平认为该公司的股票与气温之间存在一定程度的线性相关关系?

解:

$H_0:\rho=0$; $H_1:\rho\neq0$

将以上数据代入(7.4)式,计算 r 的 t-检验值

$$t=\frac{0.5\sqrt{(6-2)}}{\sqrt{1-0.5^2}}=1.1547$$

查表可知:显著水平为5%,自由度为4的临界值 $t_{\alpha/2}=2.776$,上式中的 t 值小于2.776,因此,r 不能通过显著性检验。这就是说,尽管根据样本观测值计算的 r 达到0.5,但是由于样本单位过少,这一结论并不可靠,它不足以证明该公司的股票与气温之间存在一定程度的线性相关关系。

二、标准的一元线性回归模型

(一)总体回归函数

当变量之间存在着显著的相关关系时,可以利用一定的数学模型对其进行回归分析。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:

$$Y_t=\beta_1+\beta_2X_t+u_t \tag{7.5}$$

上式被称为总体回归函数。式中的 β_1 和 β_2 是未知的参数,又叫回归系数;Y_t 和 X_t 分别是 Y 和 X 的第 t 个观测值;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。

为了便于大家的理解,我们举一个具体例子,作进一步说明。在许多经济学的入门教科书中,都可以看到以下形式的消费函数:

$$Y_t = \beta_1 + \beta_2 X_t \tag{7.6}$$

式中,Y 代表消费支出,X 代表可支配收入;β_1 是被称为基础消费水平的常数项,代表不受可支配收入影响的消费支出;β_2 是边际消费倾向,它表明可支配收入每增加一个单位时,消费支出所增加的数量。这种类型的消费函数认为,可支配收入是决定消费支出的主要因素,而且它们之间的关系是线性关系。如果以 Y 为纵轴,X 为横轴,则(7.6)式可以用一条直线来表示。这种确定型的消费函数作为理论分析的一种抽象是允许的。但是,在现实经济生活中,这种确定型的消费函数却是很难成立的。试想一下,在全国亿万个家庭中,考察一下具有相同收入的家庭,他们用于消费的部分会完全相同吗?也就是说,在给定的 X 值下,Y 是否都会得到相同的结果?显而易见,这是不可能的,因为,除了收入之外,还有各种影响消费的因素。一些家庭虽然收入相同,但是,其消费习惯、家庭所处的地理位置和气候条件等千差万别,这都会使他们用于消费的部分发生差异。所以,我们只能说,平均来看消费支出与可支配收入的关系能够用直线反映。如果用数学形式表示,可以表示为:

$$E(Y_t) = \beta_1 + \beta_2 X_t \tag{7.7}$$

上式表明:在 X 的值给定的条件下,Y 的期望值是 X 的严密的线性函数。(7.7)式所反映的直线被称为总体回归直线。Y 的实际观测值并不一定位于该直线上,只是散布在该直线的周围。我们把各实际观测点与总体回归线垂直方向的间隔,称为随机误差项(参见图 7-2),也就是定义:

$$u_t = Y_t - E(Y_t) \tag{7.8}$$

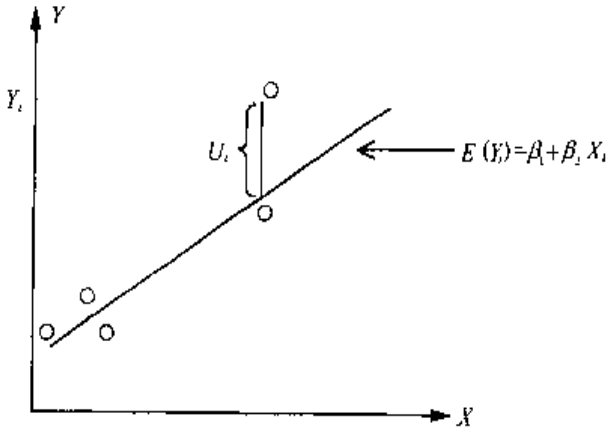


图 7-2 总体回归线与随机误差项

在反映 Y 的实际观测值与给定的 X 值的线性关系式中，引进随机误差项，就得到总体回归函数(7.5)式。

(二)样本回归函数

在现实问题研究中，由于所要研究的现象的总体单位数一般是很多的，在许多场合甚至是无限的，因此无法掌握因变量 Y 总体的全部取值。也就是说，总体回归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。显然，样本回归线的函数形式应与总体回归线的函数形式一致。一元线性回归模型的样本回归线可表示为：

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t \tag{7.9}$$

式中的 $\hat{Y}_t$ 是样本回归线上与 X_t 相对应的 Y 值，可视为 $E(Y_t)$ 的估计；$\hat{\beta}_1$ 是样本回归函数的截距系数，$\hat{\beta}_2$ 是样本回归函数的斜率系数，它们是对总体回归系数 β_1 和 β_2 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：

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X + e_t \qquad (t = 1, 2, \ldots, n) \tag{7.10}$$

上式称为样本回归函数。式中 e_t 称为残差，在概念上，e_t 与总体误差项 u_t 相互对应；n 是样本的容量。

样本回归函数与总体回归函数之间的联系显而易见。这里需要特别指出的是它们之间的区别。第一，总体回归线是未知的，它只有一条；而样本回归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线。第二，总体回归函数中的 β_1 和 β_2 是未知的参数，表现为常数；而样本回归函数中的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是随机变量，其具体数值随所抽取的样本观测值不同而变动。第三，总体回归函数中的 u_t 是 Y_t 与未知的总体回归线之间的纵向距离，它是不可直接观测的；而样本回归函数中的 e_t 是 $\hat{Y}_t$ 与样本回归线之间的纵向距离，当根据样本观测值拟合出样本回归线之后，可以计算出 e_t 的具体数值。

综上所述，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要采用适当的方法，充分利用样本所提供的信息，使得样本回归函数尽可能地接近于真实的总体回归函数。

(三)误差项的标准假定

随机误差项 u_t 是无法直接观测的。为了进行回归分析，通常需要对其概率分布提出一些假定。这些假定有：

假定 1：误差项的期望值为 0，即对所有的 t 总有

$$E(u_t)=0 \tag{7.11}$$

假定 2:误差项的方差为常数,即对所有的 t 总有

$$Var(u_t)=E(u_t^2)=\sigma^2 \tag{7.12}$$

假定 3:误差项之间不存在序列相关关系,其协方差为零,即当 $t\neq s$ 时有:

$$Cov(u_t u_s)=E(u_t u_s)=0 \tag{7.13}$$

假定 4:自变量是给定的变量,与随机误差项线性无关。

假定 5:随机误差项服从正态分布。

满足以上标准假定的一元线性模型,称为标准的一元线性回归模型。

应当指出,在现实生活中,由于各种原因,上述标准假定常常不能得到满足。那么学习以标准假定为基础的回归分析理论与方法是否会失去意义呢?当然不会。同其他一切科学研究一样,对相关现象的分析方法的研究,也可以从标准的理想状态出发,首先研究这一状态下的基本方法与规律,然后再以此为规范,进一步研究现实存在的非理想状态下可以采用的方法。关于非标准条件下的分析方法,属于计量经济学研究的内容,对此本书不作进一步的讨论。

三、一元线性回归模型的估计

(一)回归系数的点估计

如前所述,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。在根据样本资料确定样本回归方程时,一般总是希望 Y 的估计值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,因此为了便于处理,通常采用残差平方和 $\sum e_t^2$ 作为衡量总偏差的尺度。所谓最小二乘法就是根据这一思路,通过使残差平方和为最小来估计回归系数的一种方法。

$$\begin{aligned} Q &= \sum e_t^2=\sum(Y_t-\hat{Y}_t)^2 \\ &= \sum(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)^2 \end{aligned} \tag{7.14}$$

很明显,残差平方和 Q 的大小将依赖于 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的取值。根据微积分中求极小值的原理,可知 Q 存在极小值,同时欲使 Q 达到最小,Q 对 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的偏导数必须等于零。

将 Q 对 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 求偏导数,并令其等于零,可得:

$$-2\sum(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)=0 \tag{7.15}$$

$$-2\sum X_t(Y_t-\hat{\beta}_1-\hat{\beta}_2X_t)=0 \tag{7.16}$$

加以整理后有:

$$n\hat{\beta}_1+\hat{\beta}_2\sum X_t=\sum Y_t \tag{7.17}$$

$$\hat{\beta}_1\sum X_t+\hat{\beta}_2\sum X_t^2=\sum X_tY_t \tag{7.18}$$

以上方程组称为正规方程组或标准方程组,式中的 n 是样本容量。求解这一方程组可得:

$$\hat{\beta}_2=\frac{n\sum X_tY_t-\sum X_t\sum Y_t}{n\sum X_t^2-(\sum X_t)^2} \tag{7.19}$$

$$\hat{\beta}_1=\sum Y_t/n-\hat{\beta}_2\sum X_t/n=\bar{Y}-\hat{\beta}_2\bar{X} \tag{7.20}$$

以上两式是估计总体回归系数 β_1 和 β_2 公式。

【例 7-3】我们利用例 7-1 的表 7-2 中已给出我国部分省市城镇居民人均消费支出和人均可支配收入的数据,来估计我国城镇居民的基础消费水平与边际消费倾向。

解:

$$Y_t=\beta_1+\beta_2X_t+u_t$$

将表 7-2 中合计栏的有关数据代入(7.18)和(7.19)式,可得:

$$\hat{\beta}_2=\frac{13\times532\ 168\ 894-89\ 609\times70\ 118}{13\times682\ 962\ 377-89\ 609^2}=0.748\ 2$$

$$\hat{\beta}_1=70\ 118\div13-0.748\ 16\times89\ 609\div13=236.63$$

样本回归方程为:

$$\hat{Y}_t=236.63+0.748\ 2X_t$$

上式中,0.748 2 表示人均可支配收入每增加 1 元,人均消费支出会增加 0.748 2 元;236.63 元则为人均基础消费水平。

(二)总体方差的估计

除了 β_1 和 β_2 外,一元线性回归模型还包括了另一个未知参数,那就是总体随机误差项的方差 σ^2。σ^2 可以反映理论模型误差的大小,它是检验模型时必须利用的一个重要参数。由于随机误差项本身是不能直接观测的,因此,需要用最小二乘残差代替随机误差项来估计 σ^2。数学上可以证明,σ^2 的无偏估计 S 可由下式给出:

$$S^2=\frac{\sum e_t^2}{n-2} \tag{7.21}$$

上式中,分子是残差平方和;分母是自由度,其中 n 是样本观测值的个数,2 是一元线性回归方程中回归系数的个数。在一元线性回归模型中,残差 e_t 必须满足以下两个可由(7.15)式和(7.16)式导出的约束条件:

$$\sum e_t=0;\qquad \sum e_tX_t=0 \tag{7.22}$$

因而失去了两个自由度,所以其自由度为 $n-2$。

S^2 的正平方根又叫做回归估计的标准误差。S 越小表明实际观测点与所拟合的样本回归线的离差程度越小,即回归线具有较强的代表性。反之,S 越大表明实际观测

点与所拟合的样本回归线的离差程度越大，即回归线的代表性较差。

直接根据(7.21)式计算 S^2，需要逐一计算出残差平方的数值。这样手续十分麻烦，而且计算误差较大。因此，一般采用以下公式计算残差平方和：

$$\sum e_t^2 = \sum Y_t^2 - \hat{\beta}_1 \sum Y_t - \hat{\beta}_2 \sum X_t Y_t \tag{7.23}$$

上式的推导过程如下：

$$\begin{aligned}\sum e_t^2 &= \sum (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t) e_t \\ &= \sum Y_t e_t = \sum Y_t (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t) \\ &= \sum Y_t^2 - \hat{\beta}_1 \sum Y_t - \hat{\beta}_2 \sum X_t Y_t\end{aligned}$$

【例 7－4】根据例 7－1 中给出的有关数据和例 7－3 中已得到的回归系数估计值，计算我国城镇居民消费函数的总体方差 S^2 和回归估计标准差 S。

解：

将有关数据代入(7.23)式，得：

$$\begin{aligned}\sum e_t^2 &= 415\ 235\ 942 - 236.625\ 4 \times 70\ 118 - 0.748\ 16 \times 532\ 168\ 894 \\ &= 496\ 762.5\end{aligned}$$

将以上结果代入(7.21)式，可得：

$$S^2 = 496\ 762.5 \div (13-2) = 45\ 160.22$$

进而有 $$S = \sqrt{45\ 160.22} = 212.509\ 3$$

（三）最小二乘估计量的性质

最小二乘法是多种估计方法中的一种。按照最小二乘法求得的估计总体回归系数的数学公式(7.19)和(7.20)式是样本观测值的函数，通常称之为最小二乘估计量。最小二乘估计量的形式是不变的，但根据所选取的样本不同，其具体数值即回归系数的估计值却会随之变化，因此，它是一种随机变量。可以证明，在标准假定能够得到满足的条件下，回归系数的最小二乘估计量的期望值等于其真值，即有：

$$E(\hat{\beta}_1) = \beta_1 ; E(\hat{\beta}_2) = \beta_2 \tag{7.24}$$

其方差为：

$$Var(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}}{\sum (X_t - \bar{X})^2} \right) \tag{7.25}$$

$$Var(\hat{\beta}_2) = \sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum (X_t - \bar{X})^2} \tag{7.26}$$

$\hat{\beta}_2$ 和 $\hat{\beta}_1$ 的期望值与方差的推导过程基本类似。这里只就 $\hat{\beta}_2$ 进行证明。

为了便于讨论，将 $Y_t = \beta_1 + \beta_2 X_t + u_t$ 代入估计量 $\hat{\beta}_2$，并作以下变形：

$$\hat{\beta}_2 = \frac{n \sum X_t Y_t - \sum X_t \sum Y_t}{n \sum X_t^2 - (\sum X_t)^2} = \frac{\sum (X_t - \bar{X})(Y_t - \bar{Y})}{\sum (X_t - \bar{X})^2} = \frac{\sum (X_t - \bar{X}) Y_t}{\sum (X_t - \bar{X})^2}$$

$$= \beta_2 + \frac{\sum(X_t - \bar{X})u_t}{\sum(X_t - \bar{X})^2} \tag{7.27}$$

为了推导上式,利用了以下恒等式:

$$\sum(X_t - \bar{X}) = 0$$

$$\sum(X_t - \bar{X})^2 = \sum(X_t - \bar{X})X_t$$

$$\sum(X_t - \bar{X})(Y_t - \bar{Y}) = \sum(X_t - \bar{X})Y_t$$

这样,回归系数的最小二乘估计量可以表现为所要估计的参数的真值与随机误差项的线性组合。由于我们已假定 X_t 是给定的变量(不是随机变量),因此,同各期误差项相乘的权数也都是确定量。为了叙述的方便,令

$$W_t = \frac{(X_t - \bar{X})}{\sum(X_t - \bar{X})^2} \tag{7.28}$$

利用前面所述的关于随机误差项的标准假定和期望值运算的规则,可以证明 $\hat{\beta}_2$ 的期望值为:

$$\begin{aligned} E(\hat{\beta}_2) &= \beta_2 + E(\sum W_t u_t) \\ &= \beta_2 + \sum W_t E(u_t) && \text{(根据标准假定 4)} \\ &= \beta_2 + \sum W_t \times 0 && \text{(根据标准假定 1)} \\ &= \beta_2 \end{aligned} \tag{7.29}$$

$\hat{\beta}_2$ 的方差为:

$$\begin{aligned} Var(\hat{\beta}_2) &= Var(\beta_2 + \sum W_t u_t) \\ &= E(\sum W_t u_t)^2 \\ &= \sum W_t^2 E(U_t^2) && \text{(根据标准假定 4、3)} \\ &= \sigma^2 \sum W_t^2 && \text{(根据标准假定 2)} \\ &= \frac{\sigma^2}{\sum(X_t - \bar{X})^2} \end{aligned} \tag{7.30}$$

证毕。

由以上推导过程可知,最小二乘估计量是因变量观测值 Y_t 的线性函数,其期望值等于总体回归系数的真值。因此,最小二乘估计量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[①];同时随着样本容量的增大,其方差会不断缩小。也就是说,回归系数的最小二乘估计量是最优线性无偏估计量和一致估计量。

标准线性回归模型中,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

① 有兴趣的读者请参见本章习题三的解答。

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。通俗地讲，这一定理表明，在标准的假定条件下，最小二乘估计量是一种最佳的估计方式。但是应当明确，这并不意味着根据这一方式计算的每一个具体的估计值都比根据其他方式计算的具体估计值更接近真值，而只是表明如果反复多次进行估计值计算或是扩大样本的容量进行估计值计算，按最佳估计方式计算的估计值接近真值的可能性(概率)最大。

(四)回归系数的区间估计

利用以上得到的回归系数的点估计及其方差，还可以对回归系数进行区间估计。为了进行区间估计，有必要了解 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的概率分布。因为 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 均为线性估计量，是因变量 Y_t 的线性组合。根据上一节所述的标准假定，可知 Y_t 是服从正态分布的变量，所以 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 也服从正态分布。另外，以上我们已给出了 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的期望值与方差。所以有：

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}); \hat{\beta}_2 \sim N(\hat{\beta}_2, \sigma^2_{\hat{\beta}_2}) \tag{7.31}$$

在总体方差已知的情况下，利用上述正态分布便可以进行区间估计。可是，一般来说，总体方差 σ^2 是未知的，要用其无偏估计量 S^2 去代替。数学上可以证明，当样本为小样本时，回归系数估计值的标准化变换值 t 服从自由度为 $n-k$ 的 t-分布。(n 是样本容量，k 是回归系数的个数，在标准的一元线性回归模型中，$k=2$。)根据第四章中介绍的关于参数区间估计的原理，可得到以下回归系数区间估计的公式：

$$\hat{\beta}_j \pm t_{\alpha/2}{}^{(n-2)} \times S_{\hat{\beta}_j} (j=1,2) \tag{7.32}$$

式中，$S_{\hat{\beta}_j}$ 是回归系数 $\hat{\beta}_j$ 估计的样本标准误差，$t_{\alpha/2}{}^{(n-2)}$ 是显著水平为 α，自由度为($n-2$)的 t-分布双侧临界值。

$$S_{\hat{\beta}_1} = S\sqrt{\frac{1}{n} + \frac{\bar{X}}{\sum(X_t - \bar{X})^2}} \tag{7.33}$$

$$S_{\hat{\beta}_2} = \frac{S}{\sqrt{\sum(X_t - \bar{X})^2}} \tag{7.34}$$

【例 7－5】利用例 7－3 和例 7－4 的有关资料和结果，对例 7－3 中估计的我国城镇居民边际消费倾向进行置信度为 95% 的区间估计。

解：

将前面已求得的有关数据代入(7.34)式，可得：

$$S_{\hat{\beta}_2} = 212.509\,3 \div \sqrt{(682\,962\,377 - 89\,609^2/15)} = 0.026\,3$$

查 t 分布表可知：显著水平为 5%，自由度为 11 的 t-分布双侧临界值是 2.201，前面已求得 $\hat{\beta}_2 = 0.748\,16$，将其代入(7.32)式，可得：

$$0.748\,16-2.201\times0.026\,3\leqslant\beta\leqslant0.748\,16+2.201\times0.026\,3$$

即：
$$0.690\,27\leqslant\beta_2\leqslant0.806\,05$$

三、一元线性回归模型的检验

（一）回归模型检验的种类

回归模型中的参数估计出来之后，还必须对其进行检验。如果通过检验发现模型有缺陷，则必须回到模型的设定阶段或参数估计阶段。重新选择因变量和自变量及其函数形式，或者对数据进行加工整理之后再次估计参数。

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，如果它们与实质性科学的理论以及人们的实践经验不相符，就说明模型不能很好地解释现实的现象。例如，在前面所举的消费函数中，β_2 的取值区间应在 0 至 1 之间。如果根据样本数据估计的 $\hat{\beta}_2$ 大于 1 或小于 0，则不能通过经济意义检验。在对实际的社会经济现象进行回归分析时，常常会遇到经济意义检验不能通过的情况。造成这一结果的主要原因是：社会经济的统计数据无法像自然科学中的统计数据那样通过有控制的实验去取得，因而所观测的样本容量有可能偏小，不具有足够的代表性，或者不能满足标准线性回归分析所要求的假定条件。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，具体又可分为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、多重共线性检验等。二级检验对于社会经济现象的定量分析具有特别重要的意义。关于二级检验的问题在计量经济学教科书中有详细介绍。本书只讨论一级检验。

（二）拟合程度的评价

所谓拟合程度，是指样本观测值聚集在样本回归线周围的紧密程度。判断回归模型拟合程度优劣最常用的数量尺度是样本决定系数（又称决定系数）。它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（$Y_t-Y_{t_t}$），它是不能由回归直线加以解释的残差 e_t。对任一实际观测值 Y_t 总有：

$$\begin{aligned}(Y_t-\overline{Y})&=(\hat{Y}_t-\overline{Y})+(Y_t-\hat{Y}_t)\\&=(\hat{Y}_t-\overline{Y})+e_t\end{aligned}\tag{7.35}$$

对(7.35)式两边取平方并求和,得到:

$$\sum(Y_t-\bar{Y})^2=\sum(\hat{Y}_t-\bar{Y})^2+\sum(Y_t-\hat{Y}_t)^2+2(\hat{Y}_t-\bar{Y})(Y_t-\hat{Y}_t)$$

利用残差的定义与(7.22)式,可以证明:

$$\sum(\hat{Y}_t-Y)(Y_t-\hat{Y}_t)=0$$

从而有:

$$\sum(Y_t-\bar{Y})^2=\sum(\hat{Y}_t-\bar{Y})^2+\sum(Y_t-\hat{Y}_t)^2$$

即:

$$SST=SSR+SSE \tag{7.36}$$

上式中,SST 是总离差平方和;SSR 是由回归直线可以解释的那一部分离差平方和,称为回归平方和;SSE 是用回归直线无法解释的离差平方和,称为残差平方和。(7.36)式的两边同除以 SST,得:

$$1=\frac{SSR}{SST}+\frac{SSE}{SST} \tag{7.37}$$

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,可定义这一比例为决定系数,即有:

$$r^2=\frac{SSR}{SST}=1-\frac{SSE}{SST} \tag{7.38}$$

决定系数是对回归模型拟合程度的综合度量,决定系数越大,模型拟合程度越高。决定系数越小,则模型对样本的拟合程度越差。

决定系数 r^2 具有如下特性:

1.决定系数 r^2 具有非负性。[①]

由决定系数的定义式可知,r^2 的分子分母均是不可能为负值的平方和,因此其比值必大于零。

2.决定系数的取值范围为 $0\leqslant r^2\leqslant 1$。

由 r^2 的计算公式可以看出:当所有的实际观测值都位于回归直线上时,$SSE=0$,这时 $r^2=1$,说明总离差可以完全由所估计的样本回归直线来解释;当实际观测值并不是全部位于回归直线上时,但又大致分布在其附近时,$SSE>0$,这时 $r^2<1$;当回归直线没有解释任何离差,即模型中解释变量 X 与因变量 Y 完全无关时,Y 的总离差可全部归于残差平方和,即 $SSE=SST$,这时 $r^2=0$。

3.决定系数是样本观测值的函数,它也是一个统计量。

① 但是在回归模型中不包括截距项的场合,由于总离差平方和的分解公式不成立,按该式计算的 r^2 有可能小于0。

4.在一元线性回归模型中,决定系数是单相关系数的平方。

【例7-6】利用例7-4中计算的残差平方和,计算例7-3所拟合的样本回归方程的决定系数。

解:

$$r^2=1-\frac{SSE}{SST}=1-\frac{496\ 762.5}{37\ 041\ 025}=0.986\ 6$$

上式中的 SST 是利用表7-2中给出的数据按下式计算的:

$$SST=\sum Y_i^2-(\sum Y_i)^2/n$$
$$=415\ 235\ 942-(70\ 118)^2\div 13=37\ 041\ 025$$

(三)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。在一元线性回归模型中,由于只有一个解释变量 X,对 β_2 的检验与对整个方程的显著性检验是等价的。因此这里只介绍对回归系数的显著性检验,对整个回归方程的显著性检验将在下一节中再作介绍。

所谓回归系数的显著性检验,就是根据样本估计的结果对总体回归系数的有关假设进行检验。β_1 与 β_2 的检验方法是相同的,但 β_2 的检验更为重要,因为它表明自变量对因变量线性影响的程度。

下面我们以 β_2 的检验为例,介绍回归系数显著性检验的基本步骤:

1. t-检验。

(1)提出假设。对回归系数进行显著性检验,所提出的假设的一般形式是:

$$H_0:\ \beta_2=\beta_2^*,\ H_1:\ \beta_2\neq\beta_2^* \tag{7.39}$$

式中,H_0 表示原假设;H_1 表示备选假设;β_2^* 是假设的总体回归系数的真值。在许多回归分析的计算机程序里,常常令 $\beta_2^*=0$。这是因为 β_2是否为0,可以表明 X 对 Y 是否有显著的影响。

(2)确定显著水平 α。显著水平的大小应根据犯哪一类错误可能带来损失的大小确定。一般情况下取0.05。

(3)计算回归系数的 t 值。

$$t_{\beta_2}=\frac{\hat{\beta}_2-\beta_2^*}{S_{\beta_2}} \tag{7.40}$$

上式中,S_{β_2}是回归系数 $\hat{\beta}_2$ 估计的标准误差。

(4)确定临界值。t-检验的临界值是由显著水平和自由度 d_f 决定的。这时应该注意,原假设和备选假设设定的方式不同,据以判断的接受域和拒绝域也不相同。例如,对 H_0: $\beta_2=0$,H_1: $\beta_2\neq0$,进行的是双侧 t 检验;而对 H_0: $\beta_2=0.9$,H_1: $\beta_2<0.9$,进

行的是单侧 t-检验。对此，在双侧检验的场合，依据 α 和 d_f，查 t-分布表所确定的临界值是$(-t_{\alpha/2})$和$(t_{\alpha/2})$；而在单侧检验的场合，所确定的临界值是(t_α)。

(5)作出判断。如果 t_{β_2}的绝对值大于临界值的绝对值，就拒绝原假设，接受备择假设；反之，如果 t_{β_2}的绝对值小于临界值的绝对值，则接受原假设。

2. p-检验。回归系数的显著性检验还可以采用 p 检验。其前三步与 t-检验相同，但 t-值计算出来之后，并不与 t-分布的临界值进行对比，而是直接计算自由度为$n-2$的 t 统计量大于或小于根据样本观测值计算的 t_{β_2}的概率即 p 值。然后将其与给定的显著水平 α 对比，如果 p 小于α，则拒绝原假设，反之则接受原假设。利用 Excel 进行回归分析时，计算机将直接给出回归系数估计的 p 值。

【例 7－7】利用例 7－3 和例 7－4 的有关资料和结果，对例 7－3 中估计的我国城镇居民边际消费倾向进行显著性检验。

(1)以 5% 的显著水平检验可支配收入是否对消费支出有显著影响。

(2)对 H_0：$\beta_2=0.8$，H_1：$\beta_2<0.8$ 进行检验。

解：

(1) 首先，提出假设 H_0：$\beta_2=0$，H_1：$\beta_2\neq0$。

其次，利用(7.40)式计算 t 值

$$t_{\beta_2}=0.748\ 16/0.026\ 3=28.45$$

上式中的 S_{β_2}是利用例 7－3 和例 7－4 的有关数据和结果按(7.34)式求得的。

$$S_{\beta_2}=212.509\ 3\div\sqrt{(682\ 962\ 377-89\ 609^2/15)}=0.026\ 3$$

查 t－分布表可知：显著水平为 5%，自由度为 11 的双侧 t 检验的临界值是 2.201。以上计算的 t 值远远大于 2.201，所以拒绝原假设，接受备择假设，即认为可支配收入对消费支出的影响是非常显著的。

(2) 利用(7.40)式计算 t 值

$$t_{\beta_2}=(0.748\ 16-0.8)/0.026\ 3=-1.971$$

查 t－分布表可知：显著水平为 5%，自由度为 11 的单侧 t 检验的临界值是 1.796。因为计算的 t 值的绝对值大于 1.796，所以否定 $\beta_2=0.8$ 的原假设，接受备选假设，认为我国城镇居民的平均消费倾向小于 0.8。

四、一元线性回归模型预测

(一)回归预测的基本公式

建立回归模型的重要目的之一是进行预测。如果所拟合的样本回归方程经过检验，被认为具有经济意义，同时被证明有较高的拟合程度，就可以利用其来进行预测。

简单回归预测的基本公式如下：

$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f \tag{7.41}$$

式中，X_f 是给定的 X 的具体数值；$\hat{Y}_f$ 是 X_f 给定时 Y 的预测值；$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是已估计出的样本回归系数。回归预测是一种有条件的预测，在进行回归预测时，必须先给出 X_f 的具体数值。当给出的 X_f 属于样本内的数值时，利用该式去计算 $\hat{Y}_f$ 称为内插检验或事后预测。而当给出的 X_f 在样本之外时，利用该式去计算 $\hat{Y}_f$ 称为外推预测或事前预测。通常所说的预测是指事前预测。

（二）预测误差

$\hat{Y}_f$ 是根据样本回归方程计算的，它是样本观测值的函数，因而也是一个随机变量。$\hat{Y}_f$ 与所要预测的 Y 的真值之间必然存在一定的误差。在实际的回归模型预测中，发生预测误差的原因可以概括为以下四个：

1. 模型本身的误差因素所造成的误差。由于总体回归函数并未将所有影响 Y 的因素都纳入模型，同时其具体的函数形式也只是实际变量之间数量联系的近似反映，因此必然存在误差。这一误差可以用总体随机误差项的方差来评价。

2. 由于回归系数的估计值同其真值不一致所造成的误差。如前所述，样本回归系数是根据样本估计的，它与总体回归系数之间总是有一定的误差。这一误差可以用回归系数的最小二乘估计量的方差来评价。

3. 由于自变量 X 的设定值同其实际值的偏离所造成的误差。当给出的 X_f 在样本之外时，其本身也需要利用某种方法去进行预测。如果 X_f 与未来时期 X 的实际值不符，将其代入(7.41)式求得的 Y 的预测值当然也会与其实际值有所不同。

4. 由于未来时期总体回归系数发生变化所造成的误差。在研究客观经济现象的总体回归方程中，总体回归系数是一定时期内经济结构的数量特征，随着社会经济运行机制和经济结构的变化，它也会有所变动。这时，如果仍沿用根据样本期数据拟合的样本回归方程去进行预测，也会造成误差。

在以上造成预测误差的原因中，3,4 两项不属于回归方程本身的问题，而且也难以事先予以估计和控制。因此，在下面的讨论中，假定只存在 1,2 两种误差。

设 X_f 给定时 Y 的真值为 Y_f，

$$Y_f = \beta_1 + \beta_2 X_f + u_f \tag{7.42}$$

则有

$$\begin{aligned} e_f &= Y_f - \hat{Y}_f \\ &= (\beta_1 + \beta_2 X_f + u_f) - (\hat{\beta}_1 + \hat{\beta}_2 X_f) \\ &= (\beta_1 - \hat{\beta}_1) + (\hat{\beta}_2 - \beta_2) X_f + u_f \end{aligned} \tag{7.43}$$

式中，e_f 是预测的残差。利用期望值与方差的运算规则以及前面给出的回归系数最小

二乘估计量的期望值和方差，可以证明：

$$E(e_f)=0 \tag{7.44}$$

$$Var(e_f)=\sigma^2\left[1+\frac{1}{n}+\frac{(X_f-\bar{X})^2}{\sum(X_t-\bar{X})^2}\right] \tag{7.45}$$

在此基础上，还可以进一步证明 $\hat{Y}_f$ 是 Y_f 的最优线性无偏预测，即在标准假定能够满足的情况下，(7.41)式是 Y_f 的最佳预测方式。

（三）区间预测

(7.41)式给出的是对 Y_f 的点估计，在许多场合，人们更为关心的是对 Y_f 的区间估计。在标准假定条件下，e_f 服从于正态分布，即

$$e_f \sim N(0, Var(e_f)) \tag{7.46}$$

由于 $Var(e_f)$ 中的 σ^2 是未知的，通常用其无偏估计 S^2 来代替。若用 S_{e_f} 来表示预测标准误差的估计值，

$$S_{e_f}=S\sqrt{1+\frac{1}{n}+\frac{(X_f-\bar{X})^2}{\sum(X_t-\bar{X})^2}} \tag{7.47}$$

则数学上可以证明：

$$(Y_f-\hat{Y}_f)/S_{e_f} \tag{7.48}$$

服从于自由度为$(n-2)$的 t-分布。按照确定置信区间的方法，可以得出 Y_f 的$(1-\alpha)$的置信区间为：

$$Y_f \pm t_{\alpha/2}^{(n-2)} \times S_{e_f} \tag{7.49}$$

式中，$t_{\alpha/2}^{(n-2)}$是置信度为$(1-\alpha)$、自由度为$(n-2)$的 t-分布的临界值。

对于每一个给定的 X 值，计算相应的 Y 的置信区间，并将连接各点的曲线描绘在平面图上，便可得到图 7－3。

从置信区间和 S_{e_f} 的计算公式以及图 7－3，可以得到以下结论：

第一，置信区间的上下限对称地落在样本回归直线两边，呈中间小两头大的喇叭型。当 $X_f=\bar{X}$ 时的置信区间最窄，而当 X_f 远离 $\bar{X}$ 时，其置信区间逐渐增大。这就是说，在用回归模型进行预测时，X_f 的取值不宜离开 $\bar{X}$ 过远，否则预测精度将会降低，有可能使预测失效。

第二，在样本容量 n 保持不变时，$t_{\alpha/2}^{(n-2)}$的值，随置信度$(1-\alpha)$的提高而增加，因此，要求预测值的概率保证程度增加，在其他条件不变时，也就意味着预测精度的降低。

第三，当其他条件不变时，$t_{\alpha/2}^{(n-2)}$和 S_{e_f} 的值均为样本容量 n 的减函数，即随着 n 的增加，这二者将逐渐减少。这说明随着样本容量的增加，预测精度将会提高，而样本容量过小，预测的精度就较差。

第四，当 n 足够大时，S_{e_f}会趋近于 S；$t_{\alpha/2}^{(n-2)}$会趋近于 $z_{\alpha/2}$。［$z_{\alpha/2}$是置信度为(1

$-\alpha$)的标准正态分布的临界值]。这时,可以用 S 和 $z_{\alpha/2}$ 取代 S_{e_f} 和 $t_{\alpha/2}$ 来确定预测区间。即样本容量充分大时,Y_f 的$(1-\alpha)$的置信区间为:

$$Y_f \pm z_{\alpha/2} \times S \tag{7.50}$$

按上式确定的预测区间的上、下限在平面图上呈两条直线(参见图 7-3 中与样本回归线平行的两条虚线)。

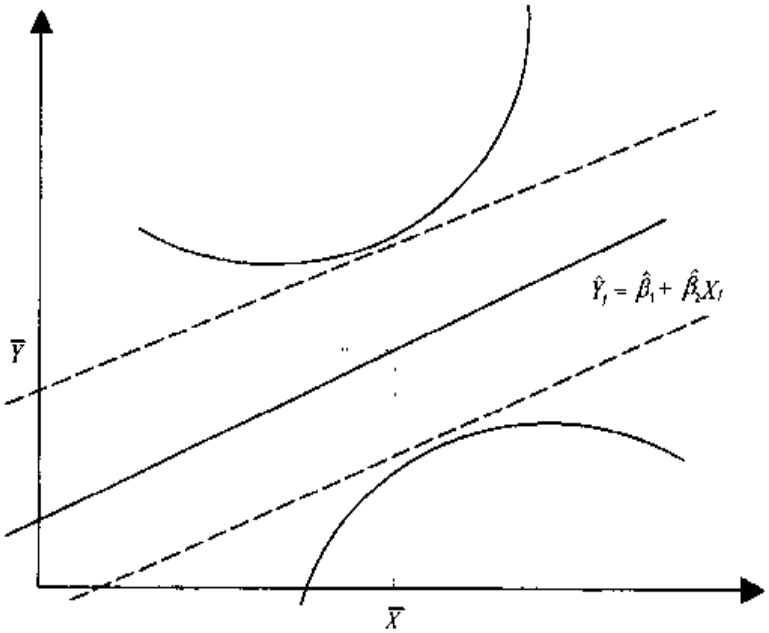


图 7-3 回归预测的置信区间

【例 7-8】假定已知某居民家庭的年人均可支配收入为 6 000 元,要求利用例 7-3 中拟合的样本回归方程与有关数据,计算置信度为 95% 的年人均消费支出的预测区间。

解:

将有关数据代入拟合好的样本回归方程,可得:

$$\hat{Y}_t = 236.625\,4 + 0.748\,16X_f$$
$$= 236.625\,4 + 0.748\,16 \times 6\,000 = 4\,725.59(\text{元})$$

从前面几例的结果可知:$S = 212.509\,3$

$$\sum X_t = 89\,609, \quad \sum(X_t - \bar{X})^2 = 682\,962\,377 - 89\,609^2/13 = 65\,287\,540$$

$n = 13$,将其代入求预测标准误差估计值的公式(7.47)式,有

$$S_{e_f} = 212.509\,3\sqrt{1 + 1/13 + (6\,000 - 89\,609/13)^2/65\,287\,540} = 221.78\ (\text{元})$$

查 t-分布表可知:显著水平为 5%,自由度为 11 的双侧 t 检验的临界值是 2.201。因此,当人均可支配收入为 6 000 元时,置信度为 95 % 的消费支出的预测区间如下:

$$4\,725.59 - 2.201 \times 221.78 \leqslant Y_f \leqslant 4\,725.59 + 2.201 \times 221.78$$

$$4\,237.45(\text{元}) \leqslant Y_f \leqslant 5\,213.73\ (\text{元})$$

第三节　多元线性相关与回归分析

一、标准的多元线性回归模型

上一节介绍的一元线性回归分析所反映的是一个因变量与一个自变量之间的关系。但是，在现实中，某一现象的变动常受多种现象变动的影响。例如，消费除了受本期收入水平的影响外，还会受以往消费和收入水平的影响；一个工业企业利润额的大小除了与总产值多少有关外，还与成本、价格等有关。这就是说，影响因变量的自变量通常不是一个，而是多个。在许多场合，仅仅考虑单个变量是不够的，还需要就一个因变量与多个自变量的联系来进行考察，才能获得比较满意的结果。这就产生了测定与分析多因素之间相关关系的问题。

研究在线性相关条件下，两个和两个以上自变量对一个因变量的数量变化关系，称为多元线性回归分析，表现这一数量关系的数学公式，称为多元线性回归模型。多元线性回归模型是一元线性回归模型的扩展，其基本原理与一元线性回归模型相类似，只是在计算上比较麻烦一些而已。限于本书的篇幅和程度，本节对于多元回归分析中与一元回归分析相类似的内容，仅给出必要的结论，不作进一步的论证。只对某些多元回归分析所特有的问题作比较详细的说明。

多元线性回归模型总体回归函数的一般形式如下：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t \tag{7.51}$$

上式假定因变量 Y 与 $(k-1)$ 个自变量之间的回归关系可以用线性函数来近似反映。式中，Y_t 是变量 Y 的第 t 个观测值；X_{jt} 是第 j 个自变量 X_j 的第 t 个观测值 $(j=2,\cdots,k)$；u_t 是随机误差项；$\beta_1,\beta_2,\cdots,\beta_k$ 是总体回归系数。β_j 表示在其他自变量保持不变的情况下，自变量 X_j 变动一个单位所引起的因变量 Y 平均变动的数额，因而又叫做偏回归系数。该式中，总体回归系数是未知的，必须利用有关的样本观测值来进行估计。

假设已给出了 n 个观测值，同时 $\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k$ 为总体回归系数的估计，则多元线性回归模型的样本回归函数如下：

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \cdots + \hat{\beta}_k X_{kt} + e_t \qquad (t=1,2,\cdots,n) \tag{7.52}$$

式中，e_t 是 Y_t 与其估计 $\hat{Y}_t$ 之间的离差，即残差。与一元线性回归分析相类似，为了进行多元线性回归分析也需要提出一些必要的假定。多元线性回归分析的标准假定除了包括上一节中已经提出的关于随机误差项的假定外，还要追加一条假定。这就是回归模型所包含的自变量之间不能具有较强的线性关系，同时样本容量必须大于所要估计的回归系数的个数即 $n>k$。我们称这条假定为标准假定 6。

二、多元线性回归模型的估计

(一)回归系数的估计

多元线性回归模型中回归系数的估计同样采用最小二乘法。设

$$\begin{aligned} Q &= \sum e_t^2 = \sum (Y_t - \hat{Y}_t)^2 \\ &= \sum (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \cdots - \hat{\beta}_k X_{kt})^2 \end{aligned} \tag{7.53}$$

根据微积分中求极小值的原理，可知残差平方和 Q 存在极小值，欲使 Q 达到最小，Q 对 $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 的偏导数必须等于零。将 Q 对 $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 求偏导数，并令其等于零，加以整理后可得到以下 k 个方程式：

$$\begin{aligned} n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2t} + \cdots + \hat{\beta}_k \sum X_{kt} &= \sum Y_t \\ \hat{\beta}_1 \sum X_{2t} + \hat{\beta}_2 \sum X_{2t}^2 + \cdots + \hat{\beta}_k \sum X_{2t} X_{kt} &= \sum X_{2t} Y_t \\ &\cdots \\ \hat{\beta}_1 \sum X_{kt} + \hat{\beta}_2 \sum X_{2t} X_{kt} + \cdots + \hat{\beta}_k \sum X_{kt}^2 &= \sum X_{kt} Y_t \end{aligned} \tag{7.54}$$

以上 k 元一次方程组称为正规方程组或标准方程组，通过求解这一方程组便可以得到 $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$。

求解多元回归方程，用矩阵形式来表达较为简便①。记

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix} \quad \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$\boldsymbol{B} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} \quad \hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

① 这里给出的矩阵形式具有一般性，对于一元线性回归模型也同样适用。对于尚未学过矩阵代数的读者，可以不必掌握这一部分内容。

则总体回归函数(7.51)式可以写为：

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U} \tag{7.55}$$

样本回归函数(7.52)式可以写为：

$$\boldsymbol{Y} = \hat{\boldsymbol{X}}\boldsymbol{B} + e \tag{7.56}$$

标准方程组(7.54)式可以写为：

$$(X'X)\hat{B} = X'Y \tag{7.57}$$

式中，$\boldsymbol{X}'$表示$\boldsymbol{X}$的转置矩阵。$(\boldsymbol{X}'\boldsymbol{X})$是一个$\boldsymbol{k} \times \boldsymbol{k}$的对称矩阵，根据标准假定6，$\boldsymbol{k}$个自变量之间不存在高度的线性相关，因此其逆矩阵存在。在(7.57)式的两边同时左乘$(\boldsymbol{X}'\boldsymbol{X})^{-1}$，可以得到：

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{7.58}$$

上式是回归系数最小二乘估计的一般形式。

实际求解多元回归方程中的回归系数的估计值，通常需要依靠电子计算机。在电子计算机技术十分发达的今天，多元回归分析的计算已经变得相当简单。利用现成的软件包如Excel等，只要将有关数据输入电子计算机，并指定因变量和相应的自变量，立刻就能得到计算结果。因此，对于从事应用研究的人们来说，更为重要的是要能够理解输入和输出之间相互对应的关系，以及对电子计算机输出的结果作出正确的解释。

限于篇幅，这里不给出具体的数值计算实例。而在下一节中，我们将结合实际的例子，讲解如何利用EXCEL进行多元线性回归分析。

(二)总体方差的估计

除了回归系数以外，多元线性回归模型中还包含了另一个未知参数，那就是随机误差项的方差σ^2。与一元回归分析相类似，多元线性回归模型中的σ^2也是利用残差平方和除以其自由度来估计的。即有：

$$S^2 = \frac{\sum e_i^2}{n-k} \tag{7.59}$$

上式中，n是样本观测值的个数；k是方程中回归系数的个数；在$(k-1)$元回归模型中，标准方程组有k个方程式，残差必须满足k个约束条件，因此其自由度为$(n-k)$。数学上可以证明，S^2是σ^2的无偏估计。S^2的正平方根S又叫做回归估计的标准误差。S越小表明样本回归方程的代表性越强。

在编制计算机程序时，残差平方和一般不是按照其定义式计算，而是利用以下公式计算：

$$\sum e_i^2 = e'e = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{B}}'\boldsymbol{X}'\boldsymbol{Y} \tag{7.60}$$

上式是残差平方和的矩阵形式。式中的“′”表示求转置；Y是因变量样本观测值向

量；X 是自变量样本观测值矩阵；$\hat{B}'$是回归系数估计值向量的转置向量。

（三）最小二乘估计量的性质

与一元线性回归模型类似，多元线性回归模型中回归系数的最小二乘估计量也是随机变量。数学上可以证明，在标准假定条件可以得到满足的情况下，多元回归模型中回归系数最小二乘估计量的期望值同样等于总体回归系数的真值，即有：

$$E(\hat{B})=B \tag{7.61}$$

回归系数最小二乘估计量的方差、协方差矩阵为：

$$\begin{aligned} Var(\hat{B}) &= E(B-B)(\hat{B}-B)' \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{7.62}$$

该矩阵主对角元素是各回归系数估计量的方差 $E(\hat{\beta}_j-\beta_j)^2$，其他元素是各回归系数估计量之间的协方差 $E(\hat{\beta}_j-\beta_j)(\hat{\beta}_i-\beta_i)$ $(i\neq j)$。在此基础上，还可以进一步证明回归系数的最小二乘估计量是最优线性无偏估计量和一致估计量。也就是说，在标准的多元线性回归模型中，高斯·马尔可夫定理同样成立。

三、多元线性回归模型的检验和预测

（一）拟合程度的评价

在多元线性回归分析中，总离差平方和的分解公式依然成立。因此也可以用上一节所定义的决定系数作为评价模型拟合程度的一项指标。不过，为了避免混淆，多元回归的决定系数用 R^2 表示。

$$R^2=1-\frac{\sum e_t^2}{\sum(Y_t-\overline{Y})^2} \tag{7.63}$$

利用 R^2 来评价多元线性回归方程的拟合程度，必须注意以下问题。由决定系数的定义可知，R^2 的大小取决于残差平方和$\sum e_t^2$ 在总离差平方和$\sum(Y_t-\overline{Y})^2$ 中所占的比重。在样本容量一定的条件下，总离差平方和与自变量的个数无关，而残差平方和则会随着模型中自变量个数的增加不断减少，至少不会增加。因此，R^2 是自变量个数的非递减函数。在一元线性回归模型中，所有模型包含的变量数目都相同，如果所使用的样本容量也一样，决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，各回归模型所含的变量的数目未必相同，以 R^2 的大小作为衡量拟合优劣的尺度是不合适的。因此，在多元回归分析中，人们更常用的评价指标是所谓的修正自由度的决定系数 $\bar{R}^2$。该指标的定义如下：

$$\bar{R}^2=1-\frac{\sum e_t^2/(n-k)}{\sum(Y_t-\overline{Y})^2/(n-1)} \tag{7.64}$$

$$=1-\frac{(n-1)}{(n-k)}(1-R^2)^{①} \tag{7.65}$$

式中，n 是样本容量；k 是模型中回归系数的个数。$(n-1)$和$(n-k)$实际上分别是总离差平方和与残差平方和的自由度。

修正自由度的决定系数 $\bar{R}^2$ 具有以下特点：

1. $\bar{R}^2 \leqslant R^2$。因为 $k \geqslant 1$，所以根据 $\bar{R}^2$ 和 R^2 各自的定义式可以得出这一结论。对于给定的 R^2 值和 n 值，k 值越大 $\bar{R}^2$ 越小。在进行回归分析时，一般总是希望以尽可能少的自变量去达到尽可能高的拟合程度。$\bar{R}^2$ 作为综合评价这两方面情况的一项指标显然比 R^2 更为合适。

2. $\bar{R}^2$ 小于 1，但未必都大于 0。在拟合极差的场合，$\bar{R}^2$ 有可能取负值。

【例 7-9】假设有 7 年的年度统计资料，现利用其对同一因变量拟合了两个样本回归方程。方程一中，$k=6$，$R^2=0.82$；方程二中，$k=2$，$R^2=0.80$。试对这两个回归方程的拟合程度作出评价。

解：

如果仅从 R^2 考察，似乎方程一的拟合程度更佳。但是，由于两个方程选用的自变量个数不同，这一结论是不正确的。将上列数据代入(7.65)式，可得：

方程一的 $\bar{R}^2=1-[(7-1)/(7-6))(1-0.82]=-0.08$

方程二的 $\bar{R}^2=1-[(7-1)/(7-2))(1-0.80]=0.76$

由此可见，方程二的实际拟合程度远远优于方程一。

（二）显著性检验

多元线性回归模型的显著性检验同样包括两方面的内容，即回归系数的显著性检验与回归方程的显著性检验。现分述如下：

1. 回归系数的显著性检验。多元回归中进行这一检验的目的主要是为了检验与各回归系数对应的自变量对因变量的影响是否显著，以便对自变量的取舍作出正确的判断。一般来说，当发现某个自变量的影响不显著时，应将其从模型中删除。这样才能够做到以尽可能少的自变量去达到尽可能高的拟合优度。

多元模型中回归系数的检验同样采用 t 检验，其原理和基本步骤与一元回归模型中的 t 检验基本相同，这里不再赘述。下面仅给出回归系数显著性检验 t 统计量的一般计算公式。

$$t_{\beta_j}=\frac{\hat{\beta}_j}{S_{\beta_j}} \qquad (j=1,2,\cdots,k) \tag{7.66}$$

① 对于不包含常数项的回归方程，该公式不适用。

式中，$\hat{\beta}_j$是回归系数的估计值，$S_{\hat{\beta}_j}$是$\hat{\beta}_j$的标准差的估计值。$S_{\hat{\beta}_j}$按下式计算：

$$S_{\hat{\beta}_j}=\sqrt{S^2\times\Psi_{jj}} \tag{7.67}$$

式中，Ψ_{jj}是$(\boldsymbol{X}'\boldsymbol{X})^{-1}$的第 j 个对角线元素，S^2 是随机误差项方差的估计值。(7.66)式的 t- 统计量背后的原假设是H_0：$\beta_j=0$，因此 t 的绝对值越大表明β_j 为 0 的可能性越小，即表明相应的自变量对因变量的影响是显著的。

2. 回归方程的显著性检验。多元线性回归模型包含了多个回归系数，因此对于多元回归模型，除了要对单个回归系数进行显著性检验外，还要对整个回归模型进行显著性检验。由离差平方和的分解公式可知，回归模型的总离差平方和等于回归平方和与残差平方和的和。回归模型总体函数的线性关系是否显著，其实质就是判断回归平方和与残差平方和之比值的大小问题。由于回归平方和与残差平方和的数值会随观测值的样本容量和自变量个数的不同而变化，因此不宜直接比较，而必须在方差分析的基础上利用 F 检验进行。其具体的方法步骤可归纳如下：

(1)假设总体回归方程不显著，即有

$$H_0:\ \beta_2=\beta_3=\cdots=\beta_k=0$$

(2)进行方差分析，列出回归方差分析表，如表 7－3 所示。

表 7－3 回归模型方差分析表

离差名称	平方和	自由度	方差
回归平方和	$SSR=\sum(\hat{Y}_t-\bar{Y})^2$	$k-1$	$SSR/(k-1)$
残差平方和	$SSE=\sum e_t^2$	$n-k$	$SSE/(n-k)$
总离差平方和	$SST=\sum(Y_t-\bar{Y})^2$		

表中，回归平方和的取值受 k 个回归系数估计值的影响，同时又要服从$\sum\hat{Y}_t/n=\bar{Y}$的约束条件，因此其自由度是 $k-1$。残差平方和取决于 n 个因变量的观测值，同时又要服从 k 个正规方程式的约束，因此其自由度是 $n-k$。回归平方和与残差平方和各除以自身的自由度得到的是样本方差。

(3)根据方差分析的结果求 F 统计量，即

$$F=\frac{SSR/(k-1)}{SSE/(n-k)} \tag{7.68}$$

数学上可以证明，在随机误差项服从正态分布同时原假设成立的条件下，F 服从于自由度为$(k-1)$和$(n-k)$的 F 分布。

(4)根据自由度和给定的显著性水平 α，查 F 分布表中的理论临界值 F_α。当 $F>F_\alpha$时，拒绝原假设，即认为总体回归函数中各自变量与因变量的线性回归关系显

著。当$F<F_\alpha$时，接受原假设，即认为总体回归函数中，自变量与因变量的线性关系不显著，因而所建立的回归模型没有意义。

（三）多元线性回归预测

在通过各种检验的基础上，多元线性回归模型可以用于预测。多元线性回归预测与一元线性回归预测的原理是一致的，其基本公式如下：

$$\hat{Y}_f=\hat{\beta}_1+\hat{\beta}_2X_{2f}+\cdots+\hat{\beta}_kX_{kf} \tag{7.69}$$

式中，$X_{jf}(j=2,3,\cdots,k)$是给定的 X_j 在预测期的具体数值；$\hat{\beta}_j$ 是已估计出的样本回归系数；$\hat{Y}_f$ 是 X_j 给定时 Y 的预测值。

该方程的矩阵形式为：

$$\hat{\boldsymbol{Y}}_f=\boldsymbol{X}'_f\hat{\boldsymbol{\beta}} \tag{7.70}$$

式中：

$$\boldsymbol{X}_f=\begin{pmatrix}1\\X_{2f}\\\vdots\\X_{kf}\end{pmatrix},\hat{\boldsymbol{\beta}}=\begin{pmatrix}\hat{\beta}_1\\\hat{\beta}_2\\\vdots\\\hat{\beta}_k\end{pmatrix}$$

多元线性回归预测标准误差的计算公式如下：

$$S_{e_f}=S\sqrt{1+\boldsymbol{X}'_f(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_f} \tag{7.71}$$

式中，S 是回归方程估计的标准误差。

多元线性回归预测 Y_f 的$(1-\alpha)$的置信区间可由下式给出：

$$Y_f\pm t_{\alpha/2}\times S_{e_f} \tag{7.72}$$

式中，$t_{\alpha/2}$是显著水平为 α 的 t-分布双侧临界值。

四、复相关系数和偏相关系数

在多变量的情况下，变量之间的相关关系是很复杂的，需要计算复相关系数与偏相关系数。

（一）复相关系数

样本复相关系数（以下简称复相关系数）的定义式如下：

$$R=\frac{\sum(Y_t-\bar{Y})(\hat{Y}_t-\bar{Y})}{\sqrt{\sum(Y_t-\bar{Y})^2\sum(\hat{Y}_t-\bar{Y})^2}} \tag{7.73}$$

上式与单相关系数的定义式十分类似，不同之处仅在于用根据 $X_2,X_3,\cdots,X_k$ 等计算的回归估计值 $\hat{Y}_t$ 代替了单相关系数定义式中的 X_t。在所涉及的变量只有两个时，因为

$\hat{Y}_t$ 是 X_t 的严密函数，所以(7.73)式完全等价于单相关系数的定义式。而在多元分析的场合，以上定义的复相关系数的平方实际上就是多元线性回归方程的决定系数。

实际计算复相关系数时，一般不直接根据其定义式，而是先计算出决定系数，然后再求决定系数的平方根。应当指出：在多个变量的情况下，Y 与其他多个变量之间既可能有正相关又可能有负相关，所以复相关系数也就只取正值。因此，复相关系数只是反映一个变量 Y 与其他多个变量 $X_2, X_3, \cdots, X_k$ 之间线性相关程度的指标，而不能反映其相互之间线性相关的方向。

复相关系数的取值区间为：$0 \leqslant R \leqslant 1$。复相关系数为 1 表明 Y 与 $X_2, X_3, \cdots, X_k$ 之间存在严密的线性关系，复相关系数为 0 则表明 Y 与 $X_2, X_3, \cdots, X_k$ 之间不存在任何线性相关关系。一般情况下，复相关系数的取值在 0 和 1 之间，表明变量之间存在一定程度的线性相关关系。

（二）偏相关系数

在对其他变量的影响进行控制的条件下，衡量多个变量中某两个变量之间的线性相关程度和相关方向的指标称为偏相关系数。偏相关系数不同于上一节中所介绍的单相关系数。在计算单相关系数时，只需要掌握两个变量的观测数据，并不考虑其他变量对这两个变量可能产生的影响。而在计算偏相关系数时，需要掌握多个变量的数据，一方面考虑多个变量相互之间可能产生的影响，一方面又采用一定的方法控制其他变量，专门考察两个特定变量的净相关关系。在多变量相关的场合，由于变量之间存在错综复杂的关系，因此偏相关系数与单相关系数在数值上可能相差很大，有时甚至符号都可能相反。单相关系数受其他因素的影响，反映的往往是表面的非本质的联系，而偏相关系数则较能说明现象之间真实的联系。例如，一种商品的需求既受收入水平的影响又受其价格的影响。按照经济学理论，在一定的收入水平下，该商品的价格越高，商品的需求量就越小。也就是说，需求与价格之间应当是负相关。可是，在现实经济生活中，由于收入和价格常常都有不断提高的趋势，如果不考虑收入对需求的影响，仅仅利用需求和价格的时间序列数据去计算单相关系数，就有可能得出价格越高需求越大的错误结论。

在明确偏相关系数与单相关系数区别的基础上，我们再来讨论偏相关系数的定义公式。在上一节中，我们已经给出了样本单相关系数的定义公式：

$$r = \frac{\sum(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2 \sum(Y_t - \bar{Y})^2}} \tag{7.74}$$

样本相关系数的定义还可以从另一个角度给出。在进行相关分析时，对于所涉及的两个变量 X 和 Y 是同等看待的。若设

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t \tag{7.75}$$

$$\hat{Y}_t = \hat{a}_1 + \hat{a}_2 Y_t \tag{7.76}$$

则样本单相关系数也可定义为两个样本回归系数的乘积的开方，即：

$$r = \pm\sqrt{\hat{\beta}_2 \hat{\alpha}_2} \tag{7.77}$$

上式中 r 的符号应与回归系数的符号一致。回归系数为正数时，r 取正值；回归系数为负数时，r 取负值。容易证明(7.74)式与(7.77)式是完全等价的。也就是说，单相关系数可以表现为两个回归系数的几何平均数。样本偏相关系数也可以按照类似的形式来定义，即偏相关系数等于两个相应的偏回归系数的几何平均数。

为简明起见，下面以3变量的偏相关分析为例。设有3个变量 X_1，X_2 和 X_3。3个变量各自以另两个变量为自变量拟合的样本回归方程如下：

$$\hat{X}_{1t} = \hat{\beta}_{1.23} + \hat{\beta}_{12.3} X_{2t} + \hat{\beta}_{13.2} X_{3t} \tag{7.78}$$

$$\hat{X}_{2t} = \hat{\beta}_{2.13} + \hat{\beta}_{21.3} X_{1t} + \hat{\beta}_{23.1} X_{3t} \tag{7.79}$$

$$\hat{X}_{3t} = \hat{\beta}_{3.12} + \hat{\beta}_{31.2} X_{1t} + \hat{\beta}_{32.1} X_{2t} \tag{7.80}$$

以上各式中的第1项均为截距系数，表示当模型中的自变量取零值时因变量的平均值。例如，$\hat{\beta}_{1.23}$表示 X_2，X_3 为0时 X_1 的平均值。式中其他的回归系数称为偏回归系数，它们都有三个下标，在小圆点左边的下标为主下标，小圆点右边的下标为次下标。主下标表示所要考察的两个变量；次下标表示在考察上述两个变量的关系时，使其保持不变的变量。不难理解，偏回归系数表示：当其他自变量保持不变时，某一自变量变化一个单位而使因变量平均变化的数值。例如，$\hat{\beta}_{12.3}$表示 X_3 保持不变时，X_2 变化一单位而引起的 X_1 平均变化的数值；$\hat{\beta}_{32.1}$表示 X_1 保持不变时，X_2 变化一单位而引起的 X_3 平均变化的数值。

利用以上偏回归系数，3个变量之间的偏相关系数可定义如下：

$$r_{12.3} = \pm\sqrt{\hat{\beta}_{12.3}\hat{\beta}_{21.3}} \tag{7.81}$$

$$r_{13.2} = \pm\sqrt{\hat{\beta}_{13.2}\hat{\beta}_{31.2}} \tag{7.82}$$

$$r_{23.1} = \pm\sqrt{\hat{\beta}_{23.1}\hat{\beta}_{32.1}} \tag{7.83}$$

偏相关系数的取值范围与单相关系数一样也是在 -1 至 $+1$ 之间，其符号与相应的偏回归系数相同。

以上偏相关系数的定义可以推广到 k 个变量的场合。在进行实际的客观现象的定量分析时，人们所关心的通常是某一个因变量 Y 与多个自变量之间的偏相关程度。这时若令 Y 为 X_1，则 Y 与各自变量的偏相关系数的一般形式可表现为：

$$r_{1j.2,3,\cdots,(j-1),(j+1),\cdots,k} = \pm\sqrt{\hat{\beta}_{1j.2,3,\cdots,(j-1),(j+1),\cdots,k}\hat{\beta}_{j1.2,3,\cdots,(j-1),(j+1),\cdots,k}} \quad (j=2,3,\cdots,k) \tag{7.84}$$

式中：$\hat{\beta}_{1j.2,3,\cdots,(j-1),(j+1),\cdots,k}$是 Y 对 X_j 的偏回归系数；$\hat{\beta}_{j1.2,3\cdots,(j-1),(j+1),\cdots,k}$是 X_j 对 Y 的偏回归系数；$r_{1j.2,3,\cdots,(j-1),(j+1),\cdots,k}$表示 k 个变量情况下 Y 与 X_j 的偏相关系数，它反映其他自变量保持不变时 Y 与 X_j 的净相关程度。

第四节 Excel 在相关与回归分析中的运用

一、数据资料

这里我们给出一个具体实例，说明如何利用 Excel 进行相关与回归分析。

【例 7－10】我国城镇居民鲜蛋消费量及其影响因素的有关资料（已输入到 Excel），如图 7－4 所示。

	A	B	C	D
1	年份	城镇居民人均购买鲜蛋量（千克）	城镇居民实际人均收入（元）	肉禽蛋相对价格指数
2		Y	X_2	X_3
3	1985	6.84	748.92	100
4	1986	7.08	850.3941	102.99
5	1987	6.56	869.6369	109.26
6	1988	6.87	848.7136	122.04
7	1989	7.05	849.6538	120.27
8	1990	7.25	920.5336	116.62
9	1991	8.26	985.418	109.08
10	1992	9.45	1075.8929	105.86
11	1993	8.86	1178.3145	105.03
12	1994	9.68	1277.8954	114.35
13	1995	9.74	1339.5292	121.49
14	1996	9.64	1391.0346	118.59
15	1997	11.13	1444.9099	116.51

图 7－4 中国城镇居民鲜蛋消费量与有关资料

注：①本图表中数据根据《中国统计年鉴》的基础资料整理

②城镇居民人均实际收入＝城镇居民人均收入/城镇居民消费价格指数

③肉禽蛋相对价格指数＝肉禽蛋价格指数/城镇居民消费价格指数

④价格指数均以 1985 年为 100

要求利用上述资料进行以下分析：

1. 画出人均鲜蛋消费量与人均实际收入水平的相关图；

2. 计算人均鲜蛋消费量与人均实际收入水平及肉禽蛋相对价格指数的相关系数；

3. 建立我国城镇居民鲜蛋消费需求的多元线性回归方程，并进行统计检验；

4. 计算人均鲜蛋消费量与肉禽蛋相对价格指数之间的偏相关系数；

5. 假定 2002 年我国城镇居民的人均实际收入为 2 200 元，肉禽蛋相对价格指数为 102%，试预测人均鲜蛋的需求量，给出置信度为 95% 的预测区间。

二、绘制相关图与计算相关系数

（一）绘制相关图

在将有关数据已经输入到工作表的基础上，可按如下步骤绘制散点图：

1. 拖动鼠标选定数值区域，该区域不包括数据的标志。本例中的数值区域为“$B$3:$C$15”。

2. 执行菜单命令“插入”→“图表”，进入图表向导。

3. 选择“图表类型”为“散点图”，然后单击“下一步”按钮。

4. Excel 自动将前面所选的数值区的地址放入图表的数据区内，同时自动将排在前面一列的数据作为 X，排在后面的一列数据作为 Y。本例中，Excel 自动把人均鲜蛋购买量作为 X，人均实际收入作为 Y。如果要将纵轴与横轴对调，可点击“系列”，修改其中 X 与 Y 的数值区域后，单击“下一步”。

5. 填写图表标题为“人均鲜蛋购买量与人均实际收入的相关图”，X 轴坐标名称为“人均实际收入”，Y 轴坐标名称为“人均鲜蛋购买量”，单击“下一步”按钮。

6. 选择图表输出的位置，单击“完成”按钮，得到的结果如图 7-5 所示。

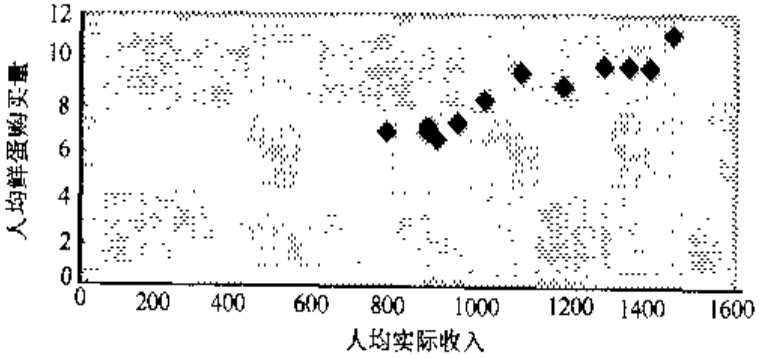


图 7-5　人均鲜蛋购买量与人均实际收入相关图

（二）计算相关系数

利用 Excel 中的相关系数分析工具可以方便地计算相关系数。操作步骤如下：

1. 调出“相关系数”分析工具对话框，其填写如图 7-6 所示。

2. 填写完“相关系数”对话框，单击“确定”即可得到各个变量的相关系数矩阵，结果如表 7-4 所示。

表 7－4

	Y	X2	X3
Y	1		
X2	0.949 724	1	
X3	0.231 866	0 371 948	1

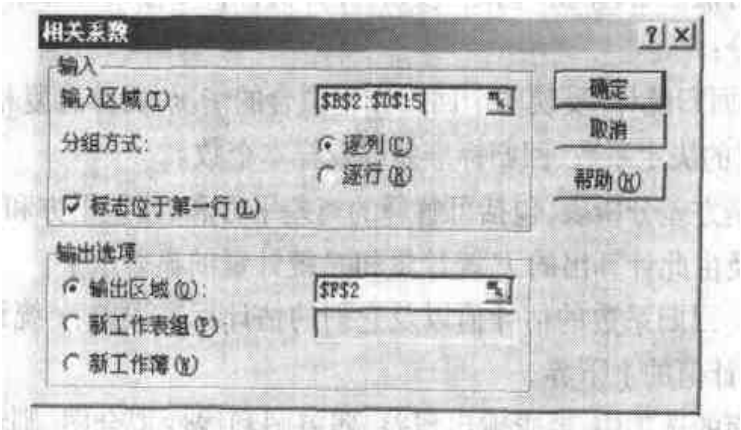


图 7－6

由上表可以看出，我国城镇居民的鲜蛋购买量与人均实际收入之间存在显著的正相关，其单相关系数为 0.949 7，与肉禽蛋相对价格指数之间则存在较弱的正相关。

三、线性回归分析

根据经济理论的分析，某种商品的需求主要取决于人们的实际收入水平和该商品的相对价格。因此，可设定以下鲜蛋需求的回归模型：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (t = 1, 2, \cdots, n) \tag{7.85}$$

利用 Excel 中配备的回归分析工具，可以方便地进行回归模型的估计和检验。具体操作步骤如下：

1. 调出"回归"分析对话框，其中主要选项的含义如下：

Y 值输入区域：在此输入因变量数据区域，该区域只能由单列数据组成，本例为"$B$2:$B$15"

X 值输入区域：在此输入自变量数据区域，本例为"$C$2:$D$15"，Excel 将此区域中的自变量从左到右排列，自变量的个数最多可达 16 个。

标志：当输入的数值区域包括变量名时，选择该复选项。

置信度：如果需要在汇总输出表中包含附加的置信度信息，则选中此复选框，然后输入所要使用的置信度，95% 为默认值。

常数为零：如果回归方程中不想包含常数项，则选中此复选框。

残差:如果需要查看残差,则选中此复选框。

标准残差:如果需要在残差输出表中包含标准残差,则选中此复选框。

残差图:如果需要生成一张图表,绘制每个自变量及其残差,则选中此复选框。

线形拟合图:如需要为预测值和观察值生成一个图表,则选中此复选框。

要注意,输出结果至少需要占有7列宽度。

2.填写完对话框后,单击"确定"按钮,计算机将自动输出计算结果。其基本内容包括以下三个部分:

第一部分"回归统计"反映整个回归方程拟合的情况,具体有复相关系数、决定系数R^2、调整自由度的决定系数、回归标准差以及样本个数。

第二部分是方差分析表,包括可解释的离差平方和、残差平方和、总离差平方和、它们的自由度以及由此计算出的F统计量和F统计量的显著水平。

第三部分是回归系数的估计值以及它们的估计标准误差、t统计量、t统计量的P值、回归系数估计值的上下界。

如果在前面的选项中,要求输出残差、残差图和线性拟合图,则计算机输出结果除了以上三个基本部分外,还包括所要求的内容。本例的输出结果如表7-5所示。

表7-5

SUMMARY OUTPUT

回归统计	
Multiple R	0.958 684
R Square	0.919 075
Adjusted R Square	0.902 89
标准误差	0.465 335
观测值	13

方差分析

	df	SS	MS	F	Significance F
回归分析	2	24.592 33	12.296 16	56.785 66	3.47E-06
残差	10	2.165 364	0.216 536		
总计	12	26.757 69			

	Coefficients	标准误差	t-Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	4.871 979	2.026 114	2.404 592	0.037 024	0.357 514	9.386 443	0.357 514	9.386 443
X2	0.006 248	0.000 604	10.340 59	1.17E-06	0.004 901	0.007 594	0.004 901	0.007 594
X3	-0.028 06	0.019 302	-1.453 59	0.176 712	-0.071 06	0.014 95	-0.071 06	0.014 95

从计算结果可知，本例所拟合的样本回归方程为：

$$\hat{Y}=4.872\,0+0.006\,2X_2-0.028\,06X_3$$
$$t=(2.40)\quad(10.34)\quad(-1.45)$$

回归系数的符号与经济理论分析的结果相符。X_2 的 t 检验值较大、P 值较低，而 X_3 的 t 检验值较小、P 值较高，这表明人均实际收入对鲜蛋需求量的影响较大，相对价格指数的影响则不够显著。另外，整个方程的 F 统计量也较大达 56.78，修正自由度的决定系数达 0.902 9，这表明该回归方程有较好的拟合程度，能够通过统计检验。

四、计算偏相关系数

前面已经得到了人均鲜蛋购买量对肉禽蛋价格指数的偏回归系数，即 $\hat{\beta}_{13.2}=-0.002\,8$。按照类似的步骤，以肉禽蛋价格指数为因变量，人均鲜蛋购买量和人均实际收入为自变量，再次进行回归，可得到肉禽蛋价格指数对人均鲜蛋购买量的偏回归系数即 $\hat{\beta}_{31.2}=-6.217\,2$。将其代入计算偏相关系数的公式，可得：

$$r_{13.2}=-\sqrt{(-0.002\,8)(-6.217\,2)}=-0.131\,9$$

将其与前面已求得的单相关系数（$r_{13}=0.231\,8$）比较，可以看出，本例中需求与价格的单相关系数和偏相关系数有较大差别，连符号也不相同。偏相关分析的结果更符合客观事物的内在联系。

五、预测

本例需要利用多元线性回归模型区间预测的矩阵公式和 Excel 中的矩阵计算函数，步骤如下：

1. 构造如图 7－7 所示的工作表，该表可通过对图 7－4 改造后得到。图中，A1:A14输入年份，B2:D14 为矩阵 X，B16:D16 为矩阵 X_f，即自变量在预测期的取值，F2:F4 存放的是 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的估计值，F5:F10 存放的是一些中间变量及最终计算结果。

为了便于输写公式，在此定义某些单元格区域的名称。定义 F6，F7，F8 的名称，选定 E6:F8，执行菜单命令“插入”→“名称”→“指定”，在调出的对话框中，选中“最左列”，单击“确定”按钮即可；定义 B2:D14 的名称，选定该区域，执行菜单命令“插入”→“名称”→“定义”，调出“定义名称”对话框，输入名称“X”，单击“确定”按钮即可。同样，将 B16:D16 定义为“Xf”，F2:F4 定义为“B”。

2. 计算点预测值 Y_f。在 F6 中输入公式“=MMULT(Xf,B)”即可。

3. 计算 t - 临界值。在 F7 中输入公式“=TINV(1－0.95,13－2－1)”即可，其中 0.95为置信度，13 为样本个数。样本个数也可通过 COUNT 函数求得。

4. 计算预测估计误差的估计值 Se_f。

先计算 $X_f(X'X)^{-1}X'_f$，在 F5 中输入公式："=MMULT(MMULT(Xf,MINVERSE(MMULT(TRANSPOSE(X),X))),TRANSPOSE(Xf))"。

然后按"Ctrl+Shift+Enter"组合键即可，表示输入的是数组公式。

再计算 Se_f，在 F8 中输入公式"=0.465 3*SQRT(1+F5)"即可，其中，0.465 3 是回归估计标准差。

5. 计算置信区间上下限。在 F9，F10 中分别输入公式"=Yf-t 临界值*Sef"和"=Yf+t 临界值*Sef"。最终结果见图 7-7。

	A	B	C	D	E	F
1	年份		X2	X3	区间预测	
2	1985	1	748.92	100	β1	4.872
3	1986	1	850.3941	102.99	β2	0.0062
4	1987	1	869.6369	109.26	β3	-0.028
5	1988	1	848.7135	122.04		2.934115
6	1989	1	849.6538	120.27	Yf	15.656
7	1990	1	920.5336	116.62	t临界值	2.228139
8	1991	1	985.418	109.08	Sef	0.922904
9	1992	1	1076.893	105.86	置信区间下限	13.69964
10	1993	1	1178.315	105.03	置信区间上限	17.71236
11	1994	1	1277.895	114.35		
12	1995	1	1339.529	121.49		
13	1996	1	1391.035	118.59		
14	1997	1	1444.91	116.51		
15						
16	2002	1	2200	102		

图 7-7

第五节　非线性相关与回归分析

一、非线性回归分析的意义

在前几节中，我们一直假定因变量和自变量之间的相关关系可以用线性方程来近似地反映。但是，在现实生活中，非线性关系是大量存在的。在许多场合，非线性的回归函数比线性回归函数更能够正确地反映客观现象之间的相互联系。例如，在考虑建立生产函数时，线性假定就不大合适。因为，线性生产函数实际上是假定各生产要素的

边际生产率不变，即资本投入每增加1单位，产出总是增加β_1，劳动投入每增加1单位，产出总是增加β_2。在这种类型的生产函数中，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，产出还会继续增加。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但又不能完全替代这样一种更符合客观现实的生产函数，就必须考虑采用非线性回归模型。

非线性回归分析必须着重解决以下两个问题：第一，如何确定非线性函数的具体形式。与线性回归分析的场合不同，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最小二乘估计法。但需要根据函数的不同类型，作适当的处理。

二、非线性函数形式的确定

在对实际的客观现象进行定量分析时，选择回归方程的具体形式应遵循以下原则：

首先，方程形式应与有关实质性科学的基本理论相一致。例如，采用幂函数的形式，能够较好地表现生产函数；采用多项式方程能够较好地反映总成本与总产量之间的关系等等。

其次，方程有较高的拟合程度。因为只有这样，才能说明回归方程可以较好地反映现实经济的运行情况。

最后，方程的数学形式要尽可能简单。如果几种形式都能基本符合上述两项要求，则应该选择其中数学形式较简单的一种。一般来说，数学形式越简单，其可操作性就越强。

为了帮助读者选择合适的函数形式，下面我们扼要介绍实际分析中较常用的几种非线性函数的特点。

（一）抛物线函数

抛物线方程的具体形式为：

$$Y = a + bX + cX^2 \tag{7.86}$$

式中a，b和c为待定参数。

判断某种现象是否适合应用抛物线，可以利用“差分法”。其步骤如下：

首先将样本观察值按X的大小顺序排列，然后按以下两式计算X和Y的一阶差分ΔX_t、ΔY_t以及Y的二阶差分ΔY_{2t}。

$$\Delta X_t = X_t - X_{t-1};\ \Delta Y_t = Y_t - Y_{t-1} \tag{7.87}$$

$$\Delta Y_{2t} = \Delta Y_t - \Delta Y_{t-1} \tag{7.88}$$

当ΔX_t接近于一常数，而ΔY_{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。

（二）双曲线函数

假如 Y 随着 X 的增加而增加（或减少），最初增加（或减少）很快，以后逐渐放慢并趋于稳定，则可以选用双曲线来拟合。双曲线的方程式是：

$$Y = a + \frac{b}{X} \tag{7.89}$$

（三）幂函数

幂函数方程的一般形式是：

$$Y = aX_1^{b1}X_2^{b2}\cdots X_k^{bk} \tag{7.90}$$

这类函数的优点在于：方程中的参数可以直接反映因变量 Y 对于某一个自变量的弹性。所谓 Y 对于 X_j 的弹性，是指在其他情况不变的条件下，X_j 变动 1% 时所引起 Y 变动的百分比。弹性是一个无量纲的数值，它是定量分析中常用的一个尺度。其一般定义如下：

$$E_{Y,X_j} = \frac{\partial Y/Y}{\partial X_j/X_j} = \frac{\partial Y}{\partial X_j}\cdot\frac{X_j}{Y} \tag{7.91}$$

利用求偏导数的规则，容易证明：在幂函数中

$$E_{Y,X_j} = b_j aX_1^{b1}X_2^{b2}\cdots X_j^{bj-1}\cdots X_k^{bk}\cdot\frac{X_j}{Y} = b_j \tag{7.92}$$

由于幂函数所具有的上述优点，它在生产函数分析和需求函数分析中，得到了广泛的应用。

（四）指数函数

指数曲线的函数为：

$$Y = ab^x \tag{7.93}$$

式中有两个待定参数 a 和 b。当 $a>0$，$b>1$ 时，曲线随 X 值的增加而弯曲上升，趋于 $+\infty$；当 $a>0$，$0<b<1$ 时，曲线随 X 值的增长而弯曲下降趋于 0。

这种曲线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；如成本、原材料消耗按一定比例降低，就符合第二种形式曲线。

（五）对数函数

对数函数的方程形式为：

$$Y = a + b\ln X \tag{7.94}$$

式中，ln 表示取自然对数。对数函数的特点是随着 X 的增大，X 的单位变动对因变量 Y 的影响效果不断递减。

（六）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：

$$Y=\frac{L}{1+ae^{-bx}} \quad (L,a,b>0) \tag{7.95}$$

逻辑曲线具有以下性质：Y 是 X 的非减函数，开始时随着 X 的增加，Y 的增长速度也逐渐加快，但是 Y 达到一定水平之后，其增长速度又逐渐放慢，最后无论 X 如何增加，Y 只会趋近于 L，而永远不会超过 L。由于逻辑曲线的这一特点，它常被用来表现耐用消费品普及率的变化。

（七）多项式方程

多项式方程在非线性回归分析中占有重要的地位。因为根据数学上级数展开的原理，任何曲线、曲面、超曲面的问题，在一定的范围内都能够用多项式任意逼近。所以，当因变量与自变量之间的确实关系未知时，可以用适当幂次的多项式来近似反映。

当所涉及的自变量只有一个时，所采用的多项式方程称为一元多项式，其一般形式如下：

$$Y=b_0+b_1X+b_2X^2+\cdots+b_kX^k \tag{7.96}$$

前面介绍过的简单线性函数、抛物线函数和双曲线函数都是一元多项式的特例。

当所涉及的自变量在两个以上时，所采用的多项式称为多元多项式。例如，二元二次多项式的形式如下：

$$Y=b_0+b_1X_1+b_2X_2+b_3X_1X_2+b_4X_1^2+b_5X_2^2 \tag{7.97}$$

一般来说，涉及的变量越多，变量的幂次越高，计算量就越大。因此，在实际的定量分析中，一般尽量避免采用多元高次多项式。

三、非线性回归模型估计

不少具有实用价值的非线性函数，可以通过适当的变换，转化为线性函数，然后再利用线性回归分析的方法进行估计和检验。

常用的非线性函数的线性变换方法有以下几种：

（一）倒数变换

倒数变换是用新的变量来替换原模型中变量的倒数，从而使原模型变成线性模型的一种方法。例如，对于双曲线函数，令 $X^*=1/X$ 代入原方程式，可有：$Y=a+bX^*$。

（二）半对数变换

这种方法主要应用于对数函数的线性变换。对于对数函数，令 $X^*=\ln X$，代入原方程，同样可得：$Y=a+bX^*$。

（三）双对数变换

这种方法通过用新变量替换原模型中变量的对数，从而使原模型变换为线性模型。例如，对幂函数的两边求对数，可得：

$$\ln Y = \ln a + b_1 \ln X_1 + b_2 \ln X_2 + \cdots + b_k \ln X_k$$

令 $Y^* = \ln Y$；$b_0 = \ln a$；$X_1^* = \ln X_1, \cdots, X_k^* = \ln X_k$，代入上式可得：

$$Y^* = b_0 + b_1 X_1^* + b^2 X_2^* + \cdots + b_k X_k^*$$

（四）多项式变换

这种方法适用于多项式方程的变换。例如，对于二元二次多项式，可令 $X_2^* = X_1$，$X_3^* = X_2$，$X_4^* = X_1 X_2$，$X_5^* = X_1^2$，$X_6^* = X_2^2$，代入原方程，可得：

$$Y = b_0 + b_1 X_2^* + b_2 X_3^* + b_3 X_4^* + b_4 X_5^* + b_5 X_6^*$$

以上所述的线性变换的方法具有简便易行的优点。但是，在实际应用时要注意以下几个问题：

第一，对于一些比较复杂的非线性函数，常常需要综合利用上述的几种方法。例如，对于逻辑曲线函数，若假定 $L = 100$，则可以采用以下方式进行线性变换。首先，(7.95)式两边同时取倒数，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：$\ln(100/Y - 1) = \ln a - bX$

令 $Y^* = \ln(100/Y - 1)$；$b_1 = \ln a$；$b_2 = -b$，代入上式，可得：$Y^* = b_1 + b_2 X$

在以上变换过程中，就综合利用了倒数变换和对数变换。

第二，为了能够根据样本观测值，对通过变换得到的线性回归方程式进行估计，该方程中的所有变量都不允许包含未知的参数。例如，如果 L 未知，上面所述的逻辑曲线方程的线性变换就是不正确的。这是因为这种情况下 Y^* 包含了未知的 L，因而是不可观测的。

第三，在以上的讨论中，为了叙述方便，我们省略了非线性回归函数中包含的随机误差项。但事实上与线性回归分析的场合一样，非线性回归分析也要考虑随机误差项的问题。只有当变换后的新模型中包含的误差项能够满足各种标准假定时，新模型中回归系数最小二乘估计量的各种理想性质才能成立。

第四，严格地说，上述的各种线性变换方法只是适用于变量为非线性的函数。对于参数为非线性或参数与变量均为非线性的函数来说，即使有可能进行线性变换和回归估计，也无法得到原方程中非线性参数的无偏估计量。

最后，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。在遇到这种情况时，还需要利用其他一些方法，如泰勒级数展开法等去进行估计。

这些方法比较复杂,超出了本书的程度,这里不作进一步的介绍。

【例 7－11】利用例 7－10 中给出的资料

(1)拟合幂函数形式的我国城镇居民鲜蛋的需求函数。

(2)利用以上建立的样本回归方程,试预测人均实际收入为 2 200 元,肉禽蛋相对价格指数为 102% 时的人均鲜蛋需求量。

解:

(1)幂函数形式的需求函数如下:

$$Y_t = aX_{2t}^{b2}X_{3t}^{b3}$$

利用双对数变换法,同时加入随机误差项,可得以下线性回归函数:

$$Y_t^* = \beta_1 + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + u_t$$

式中,$Y^* = \ln Y$;$\beta_1 = \ln a$;$\beta_2 = b_2$;$\beta_3 = b_3$;$X_2^* = \ln X_2$; $X_2^* = \ln X_3$。

利用 Excel 中的自然对数函数,对例 7－10 中给出的人均鲜蛋购买量 Y,城镇居民人均实际收入 X_1 和肉禽蛋价格指数 X_2 求自然对数。例如,人均鲜蛋购买量放在 B 列,所要求解的相应自然对数放在 E 列 ,则先令 E2＝ln(B2),然后,复制其他单元格,便可得到 $\ln Y$。

然后,利用上一节中介绍的求解一般线性回归模型的方法,对变换后的模型求解,可以得到以下结果:(具体运算过程省略)

$$\hat{Y}_t^* = -1.602\,0 + 0.799\,3X_{2t}^* - 0.390\,1X_{3t}^* \tag{7.98}$$

$$t = (-1.45),(10.53),(-1.55);$$

$$F = 33.60;\bar{R}^2 = 0.906\,0$$

因为,$\hat{a} = e^{\beta_1} = e^{-1.602\,0} = 0.201\,493$,所以与上式相对应的幂函数形式的样本回归方程为:

$$\hat{Y}_t = 0.201\,493X_{2t}^{0.793\,3}X_{3t}^{-0.390\,1} \tag{7.99}$$

由上式可知:居民人均实际收入的需求弹性约为 0.8,而相对价格的需求弹性约为 －0.4。也就是说,在其他情况不变的条件下,居民人均实际收入每增加 1% 会使人均鲜蛋的需求增加 0.8% ,相对价格每提高 1% 会使人均鲜蛋需求减少 0.4% 。

(2)将以上给出的居民人均实际收入和相对价格指数代入(7.99)式,可得:

$$\hat{Y}_t = 0.201\,493(2\,200)^{0.793\,3}(103)^{-0.390\,1} = 14.81(\text{公斤/人})$$

四、相关指数

变量之间存在的非线性相关的强弱,难以用单相关系数去判断。在这种场合,可以利用相关指数作为判断变量之间是否显著存在某种类型的非线性相关关系的尺度。所

谓相关指数，也就是对非线性回归模型进行拟合时所得到的决定系数。

【例 7－12】假设变量 Y 与变量 X 的样本观测值如表 7－6 所示。

表 7－6

X	0	1	2	3	4	5	6
Y	9	5	2	1	2	5	9.2

试计算 Y 与 X 的单相关系数和以 Y 为因变量、X 为自变量的抛物线方程的相关指数，判断 Y 与 X 之间是否存在某种相关关系？

解：

利用求单相关系数的公式可得：Y 与 X 的单相关系数 $r_{xy}=0.013\,8$，可以认为两者之间线性关系很不密切。但是，拟合抛物线方程可得：

$$Y=15.371\,4-7.121\,4X+0.892\,8X^2;R^2=0.997\,02;F=669.77$$

因此，可以认为 Y 与 X 之间存在非常显著的抛物线形式的相关关系。

本章小结

1. 客观现象之间的数量联系存在着两种不同的类型：一种是确定型的函数关系；另一种是非确定型的相关关系。

2. 相关关系可以按不同的标志加以区分。按相关的程度可分为完全相关、不完全相关和不相关。按相关的形式可分为线性相关和非线性相关。按所研究的变量多少可分为单相关、复相关和偏相关。

3. 单相关分析所采用的尺度为相关系数。样本相关系数的计算公式是：

$$r=\frac{n\sum X_tY_t-\sum X_t\sum Y_t}{\sqrt{[n\sum X_t^2-(\sum X_t)^2][n\sum Y_t^2-(\sum Y_t)^2]}}$$

4. 满足一定标准假定的一元线性模型，称为标准的一元线性回归模型。其总体回归函数为：$Y_t=\beta_1+\beta_2X_t+u_t$。

5. 回归系数估计可利用最小二乘法，其计算公式为：

$$\hat{\beta}_2=\frac{n\sum X_tY_t-\sum X_t\sum Y_t}{n\sum X_t^2-(\sum X_t)^2}$$

$$\hat{\beta}_1=\sum Y_t/n-\hat{\beta}_2\sum X_t/n=\bar{Y}-\hat{\beta}_2\bar{X}$$

它们是总体回归系数的最优线性无偏估计量。

6. 总体随机误差项的方差 σ^2 的无偏估计 $S^2=\sum e_t^2/(n-2)$

S^2 的正平方根又叫做回归估计的标准误差。

7. 判断一元回归模型拟合程度最常用的数量尺度是决定系数。

$$r^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

$$0\leqslant r^2\leqslant 1$$

8. 回归系数的显著性检验，可用 t 检验。

$$t_{\hat{\beta}_j}=\frac{\hat{\beta}_j-\hat{\beta}_j^*}{S_{\hat{\beta}_j}}\qquad (j=1,2)$$

9. 简单回归预测的基本公式：$\hat{Y}_f=\hat{\beta}_1+\hat{\beta}_2X_f$

$$\text{预测标准误差 } S_{e_f}=S\sqrt{1+\frac{1}{n}+\frac{(X_f-\bar{X})^2}{\sum(X_t-\bar{X})^2}}$$

Y_f 的$(1-a)$的置信区间为：$Y_f\pm t_{a/2}^{(n-2)}\times S_{e_f}$

10. 多元线性回归模型总体回归函数的一般形式为：

$$Y_t=\beta_1+\beta_2X_{2t}+\cdots+\beta_kX_{kt}+u_t\qquad (t=1,2,\cdots,n)$$

其矩阵形式为：

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{U}$$

11. 最小二乘估计量的矩阵形式为：$\hat{B}=(X'X)^{-1}X'Y$，总体方差的估计量 $S^2=\frac{\sum e_t^2}{n-k}$。

12. 多元回归分析中，利用修正自由度的决定系数 $\bar{R}^2$ 对方程的拟合程度进行评价，更为恰当。

$$\bar{R}^2=1-\frac{\sum e_t^2/(n-k)}{\sum(Y_t-\bar{Y})^2/(n-1)}=1-\frac{(n-1)}{(n-k)}(1-R^2)$$

13. 多元线性回归模型中回归系数的显著性检验与一元模型类似，同样采用 t 检验，如表 7－7 所示。整个回归方程的显著性检验，则要在方差分析的基础上，采用 F 检验。

表 7－7 回归模型方差分析表

离差名称	平方和	自由度	方差
回归平方和	$SSR=\sum(\hat{Y}_t-\bar{Y})^2$	$k-1$	$SSR/(k-1)$
残差平方和	$SSE=\sum e_t^2$	$n-k$	$SSE/(n-k)$
总离差平方和	$SST=\sum(Y_t-\bar{Y})^2$		

$$F=\frac{SSR/(k-1)}{SSE/(n-k)}$$

14. 多元线性回归预测与一元线性回归预测的原理是一致的，其矩阵形式为：$\hat{Y}_j=$

$X'_f\hat{\beta}$。多元线性回归预测标准误差的计算公式如下：

$$S_{e_f}=S\sqrt{1+\boldsymbol{X}'_f(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_f}$$

15.复相关系数是反映一个变量 Y 与其他多个变量 $X_2,X_3,\cdots,X_k$ 之间线性相关程度的指标，它是回归方程决定系数的正的平方根。偏相关系数是在对其他变量的影响进行控制的条件下，衡量多个变量中某两个变量之间的线性相关程度和相关方向的指标，它可定义为两个相应的偏回归系数的乘积的平方根，其符号取值则与偏回归系数相同。

16.常用的非线性回归函数有：抛物线函数、双曲线函数、幂函数、指数函数、对数函数、S 形曲线函数和多项式方程。

17.相当一部分非线性函数，可以通过适当的变换，转化为线性函数，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：倒数变换、半对数变换、双对数变换和多项式变换等。

18.相关指数是对非线性回归模型进行拟合时所得到的决定系数，可用来衡量变量之间某种非线性相关关系的强弱。

19.在相关与回归分析中，Excel 的应用主要有以下几个方面：

(1)利用“图表”工具绘制相关散点图，利用“相关系数”分析工具计算相关系数；

(2)利用“回归”分析工具进行回归分析；

(3)根据回归结果，利用“矩阵计算函数”进行点预测和区间预测；

(4)利用各种函数进行非线性函数的线性变换，然后进行回归分析。

思考与练习题

一、判断题：

1.产品的单位成本随着产量增加而下降，这种现象属于函数关系。(　　)

2.相关系数为 0 表明两个变量之间不存在任何关系。(　　)

3.单纯依靠相关与回归分析，无法判断事物之间存在的因果关系。(　　)

4.圆的直径越大，其周长也越大，两者之间的关系属于正相关关系。(　　)

5.总体回归函数中的回归系数是常数，样本回归函数中的回归系数的估计量是随机变量。(　　)

6.当抽取的样本不同时，对同一总体回归模型估计的结果也有所不同。(　　)

二、选择题：

1. 变量之间的关系按相关程度分可分为：(　　)。

a. 正相关　b. 不相关　c. 完全相关　d. 不完全相关

2. 复相关系数的取值区间为：(　　)。

a. $0 \leqslant R \leqslant 1$　b. $-1 \leqslant R \leqslant 1$　c. $-\infty \leqslant R \leqslant 1$　d. $-1 \leqslant R \leqslant \infty$

3. 修正自由度的决定系数为：(　　)。

a. $\bar{R}^2 \leqslant R^2$　b. 有时小于 0

c. $0 \leqslant \bar{R}^2 \leqslant 1$　d. 比 R^2 更适合作为衡量回归方程拟合程度的指标

4. 回归预测误差的大小与下列因素有关：(　　)。

a. 样本容量　b. 自变量预测值与自变量样本平均数的离差

c. 自变量预测误差　d. 随机误差项的方差

三、问答题：

1. 请举一实例说明什么是单相关和偏相关？以及它们之间的差别。

2. 试证明(7.20)式给出的 S^2 是标准一元线性回归模型中随机误差项的方差 σ^2 的无偏估计量。

3. 试证明最小二乘估计量 $\hat{\beta}_2$ 是标准一元线性回归模型中总体回归系数 β_2 的最优线性无偏估计量。

4. 试根据表 7－8 的资料：

表 7－8　我国的国内生产总值与最终消费　　单位：亿元

年份	国内生产总值	消费	年份	国内生产总值	消费
1978	3 605.6	2 239.1	1990	18 319.5	11 365.2
1979	4 074.0	2 619.4	1991	21 280.4	13 145.9
1980	4 551.3	2 976.1	1992	25 863.6	15 952.1
1981	4 901.4	3 309.1	1993	34 500.7	20 182.1
1982	5 489.2	3 637.9	1994	46 690.7	26 796.0
1983	6 076.3	4 020.5	1995	58 510.5	33 635.0
1984	7 164.4	4 694.5	1996	68 330.4	40 003.9
1985	8 792.1	5 773.0	1997	74 894.2	43 579.4
1986	10 132.8	6 542.0	1998	79 003.3	46 405.9
1987	11 784.0	7 451.2	1999	82 673.1	49 722.7
1988	14 704.0	9 360.1	2000	89 112.5	54 617.2
1989	16 466.0	10 556.5			

资料来源：国家统计局. 中国统计年鉴. 北京：中国统计出版社，2001

要求：

(1) 以消费为因变量，国内生产总值为自变量，拟合线性回归方程。

(2) 计算回归估计的标准误差和决定系数。

(3) 对回归系数进行显著水平为 5% 的显著性检验。

(4) 假定 2001 年我国的国内生产总值为 104 880 亿元，利用拟合的回归方程预测该年可能达到的消费额，给出置信度为 95% 的预测区间。

5. 设销售收入 X 为自变量，销售成本 Y 为因变量。现已根据某百货公司 12 个月的有关资料计算出以下数据(单位：万元)：

$$\sum(X_t-\bar{X})^2=425\,053.73;\qquad \bar{X}=647.88$$

$$\sum(Y_t-\bar{Y})^2=262\,855.25;\qquad \bar{Y}=549.8$$

$$\sum(Y_t-\bar{Y})(X_t-\bar{X})=334\,229.09$$

利用以上数据，要求：

(1) 拟合简单线性回归方程，并对方程中回归系数的经济意义作出解释。

(2) 计算决定系数和回归估计的标准误差。

(3) 对 β_2 进行显著水平为 5% 的显著性检验。

(4)假定明年 1 月销售收入为 800 万元，利用拟合的回归方程预测相应的销售成本，并给出置信度为 95% 的预测区间。

6. 讨论以下几种场合回归方程：$Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+U_t$ 中回归系数的经济意义和应取的符号。

(1) Y_t 为商业利润率；X_{2t} 为人均销售额；X_{3t} 为流通费用率。

(2) Y_t 为粮食销售量；X_{2t} 为人口数；X_{3t} 为人均收入。

(3) Y_t 为工业总产值；X_{2t} 为占用的固定资产；X_{3t} 为职工人数。

(4) Y_t 为国内生产总值；X_{2t} 为工业总产值；X_{3t} 为农业总产值。

7. 利用习题 4 中给出的我国 GDP 和消费的资料，要求：

(1) 拟合以下形式的消费函数：

$$C_t=\beta_1+\beta_2Y_t+\beta_3C_{t-1}+U_t$$

式中，C_t 是 t 期的消费；C_{t-1} 是 $t-1$ 期的消费；Y_t 是 t 期的 GDP。

(2) 计算随机误差项的方差估计值。

(3) 计算修正自由度的决定系数。

(4) 计算各回归系数的 t 统计量。

(5) 对整个回归方程进行显著性检验。

(6) 假设 2001 年的国内生产总值为 95 350 亿元，试利用拟合的消费函数预测当年的消费总额，并给出置信度为 95% 的预测区间。

8.设有以下资料,见表7-9。

表7-9 某企业近年来总成本与产量

年份	总成本 Y	产量 X	年份	总成本 Y	产量 X
1	32 900	400	7	86 300	900
2	52 400	600	8	139 000	1 200
3	42 400	500	9	115 700	1 100
4	62 900	700	10	154 800	1 300
5	74 100	800	11	178 700	1 400
6	100 000	1 000	12	203 100	1 500

要求:

(1) 试拟合以下总成本函数:

$$Y_t = \beta_t + \beta_2 X_t + \beta_3 X_t^2 + \beta_4 X_t^3 + u_t$$

(2) 根据总成本函数推导出平均成本函数,并描出平均成本函数的图形。

(3) 试根据以上结果推算总产量为1 550时的单位产品平均成本。

第八章

时间序列分析

★ 本章学习要点与要求 ★

本章学习常规的时间序列分析方法，内容包括：水平分析、速度分析、长期趋势分析、季节变动分析与循环波动分析。通过本章学习，要求：①理解时间序列的含义、编制的要求；②了解各类时间序列分析的意义和基本内容，掌握有关指标和数学模型的计算方法；③能运用 Excel 和所学知识对有关经济现象的时间序列进行分析和预测。

第一节　时间序列分析概述

一、时间序列的概念

社会经济现象随着时间推移不断发展变化，关于社会经济现象的统计指标是在不同时间观察记录的。所谓时间序列也称动态数列或时间数列，它就是将各个不同时间的社会经济统计指标，按时间先后顺序排列而形成的。

表 8－1 是我国 1991 年至 2000 年若干统计指标的时间序列。不难看出，时间序列有两个基本要素：一是统计指标所属的时间，二是统计指标在特定时间的具体指标值。

在统计研究中，时间序列的分析和研究具有很重要的地位。开展时间序列分析的主要目的如下。

表 8-1 我国若干国民经济指标时间序列

指标 / 年份	GDP(亿元)	年末人口(万人)	人均 GDP(元)	第三产业所占比重(%)
1991	21 618	115 823	1 879	33.4
1992	26 638	117 171	2 287	34.3
1993	34 634	118 517	2 939	32.7
1994	46 759	119 850	3 923	31.9
1995	58 478	121 121	4 854	30.7
1996	67 885	122 389	5 576	30.1
1997	74 663	123 626	6 079	30.9
1998	78 345	124 810	6 307	32.1
1999	81 911	125 909	6 534	33.0
2000	89 404	126 583	7 082	33.2

(1)通过计算各种水平指标和速度指标,了解和分析社会经济现象发展变化的历史过程。

(2)通过长期趋势分析、季节变动分析、循环变动分析等了解和分析社会经济现象发展变化的规律性。

(3)通过建立数学模型预测现象的未来发展趋势。

二、时间序列的种类

时间序列按照其指标的性质,可以分为总量指标、相对指标和平均指标三种类型。总量指标时间数列也称绝对数时间序列,它是基本的时间序列,相对指标和平均指标时间序列是在其基础上派生出来的。

(一)总量指标时间序列

总量指标按其所反映的内容的不同,有总体单位总量和总体标志总量两种,用以反映社会经济现象达到的绝对水平。所谓总量指标时间序列就是把一系列总量指标值按时间的先后顺序排列起来所形成的时间序列。根据总量指标反映的社会经济现象性质的不同,又可分为时期指标时间序列和时点指标时间序列。在表 8-1 中,GDP 是总体标志总量指标,年末人口数是总体单位总量指标;GDP 是时期指标,年末人口数是时点指标。

下面讨论时期指标和时点指标的特点。

1.时期指标的特点

(1)可加性。不同时期的总量指标可以相加,所得数值表明现象在更长一个时期的

指标值。例如,月度GDP、季度GDP和年度GDP指标所属的时间长短不同,把1月份、2月份、3月份的GDP加总,得到第一季度的GDP,把一年四个季度的GDP加总,则得到年度的GDP。

(2)指标值的大小与所属时间的长短有直接关系。这是由时期指标的可加性特点所决定的。一般指标所属时期越长,指标值越大。如上面所说的季度GDP总是大于月度GDP,年度GDP也总是大于季度GDP。

(3)指标值采用连续统计的方式获得。由于时期指标是反映现象在一段时间内的发展过程总量,因而必须在这段时间内把现象发生的数量逐一登记,并进行累计得到指标值。

2.时点指标的特点

(1)不可加性。不同时点的总量指标不可相加,这是因为把不同时点的总量指标相加后,无法解释所得数值的时间状态。如果把我国1990年末人口数114 333万人和表8-1中的我国1991年末人口数115 823万人相加得到230 156万人,但230 156万人这个数据属于哪个具体时间是无法说明的,因此也没有意义。

(2)指标数值的大小与时点间隔的长短一般没有直接关系。在时点数列中,相邻两个指标所属时间的差距为时点间隔。因为时点指标的时间单位是瞬间,因而许多现象时间间隔的长短与指标值的大小没有直接联系。如企业年底的库存不一定比各月底的库存量大;企业12月底的职工人数也未必比11月底的职工人数多。但如果现象本身存在着长期变化趋势,如呈现长期增长或长期下降趋势,则指标数值的大小与时点间隔的长短就有一定关系了。例如,我国总人口变动呈现长期增长趋势,因此,时点间隔越长,指标的数值就越大。

(3)指标值采用间断统计的方式获得。时点指标具有不连续统计的特点。因为时点指标是反映现象在某一时刻上状况的数量,只需要在某一时点上进行统计,取得该时点资料,不必连续统计。例如,我国历次的人口普查就是根据联合国的有关建议和国家的有关规定,每间隔一定时间进行一次。

(二)相对指标和平均指标时间序列

前面已经指出,相对指标和平均指标是由总量指标派生出来的。相对指标和平均指标反映社会经济现象达到的相对水平和平均水平,把一系列相对指标或平均指标按时间先后顺序排列起来,就得到相对指标或平均指标时间序列。例如,表8-1的人均GDP和第三产业比重时间序列就属此类。

相对指标或平均指标时间序列反映了社会经济现象之间相互联系的发展过程。在经济统计分析中,往往把总量指标、相对指标和平均指标时间序列结合起来,以便从多方位对社会经济现象进行分析。

三、时间序列的编制原则

编制时间序列的目的是要通过数列中各指标的比较，来研究社会经济现象的发展及其规律。因此，保证数列中各个指标数值的可比性，是编制时间序列的基本原则。具体有以下几点要求：

（一）时间一致

对于时期指标时间序列，各指标值所属时期长短应一致。对于时点指标时间序列，各指标的时点间隔长短应一致。

（二）口径一致

包含几个方面：一是现象总体范围应一致。无论是时期指标时间序列还是时点指标时间序列，指标值的大小都与现象总体范围有密切关系，若指标的总体范围不一致，则失去比较意义。二是计算价格应一致。价值指标有不变价、现行价，而不变价又有不同时期的不变价，编制价值指标的时间序列要保证各指标的计算价格相同，才具有比较意义。三是计量单位一致。实物量指标度量单位有吨、公斤以及标准实物量和混合实物量等，编制实物量指标时间序列要保证各指标的计量单位相同。四是经济内容要一致。例如新中国成立以来，我国曾经采用过工农业总产值、社会总产值、国民收入和国内生产总值等指标反映我国的经济活动总量。这些指标都有不同的经济内容，在编制新中国成立以来的经济活动总量时间序列时，就需要对这些指标加以区别和调整，才能符合可比性的要求。

（三）计算方法一致

指标名称、总体范围、计算价格和计量单位以及经济内容都一致的指标，有时因计算方法不一致，也会导致数值上的差异。如 GDP 指标，可以用生产法、分配法和支出法来计算，从理论上讲，三种方法的计算结果应一致，但由于资料来源的渠道不同，这三种方法计算的结果往往存在差异。因此，在编制时间序列时，应注意各指标的计算方法是否统一，以确保指标可比。

第二节 时间序列的水平分析

一、发展水平与平均发展水平

(一)发展水平

时间序列中,各指标数值就是该指标所反映的社会经济现象在所属时间的发展水平。它们是受长期趋势、季节变动、循环变动和不规则变动等因素共同作用的结果,反映社会经济现象在具体时间下所达到的规模和发展程度。表 8-1 中资料显示,2000 年我国 GDP 为 89 404 亿元、人口数为 12.66 亿、人均 GDP 为 7 082 元、第三产业比重为 33.2%,就说明 2000 年我国这些社会经济现象的发展水平以及所达到的规模和发展程度。

分析现象的发展水平,要区分几个有关的概念。在时间序列中,各个指标值用 y_i 表示,则时间序列可表示为 $y_0, y_1, y_2, \cdots, y_{n-1}, y_n$,通常把首项 y_0 称为最初水平,末项 y_n 称为最末水平,作为对比基准的水平称为基期水平,被研究考察时间的水平称为报告期水平。

(二)平均发展水平

在对时间序列进行分析时,为了综合说明现象在一段时期内的发展水平,需要计算平均发展水平指标。就是对各不同时间上的指标数值求平均数,将指标在各时间上表现的差异加以抽象,以一个数值来代表现象在这一段时间上的一般发展水平,称为序时平均数。

计算序时平均数的方法要根据时间序列指标的性质来确定。在前面我们已经知道,有由总量指标、相对指标和平均指标形成的时间序列,计算这三种时间序列序时平均数的方法不同。但由于相对指标和平均指标是由总量指标派生的,所以根据总量指标时间序列计算序时平均数的方法是最基本的方法。

1.由总量指标计算序时平均数。总量指标时间数列分为时期和时点两种,二者的计算方法有所不同,下面分别加以说明。

(1)时期指标的序时平均数。这种序时平均数的计算方法比较简单,通常采用简单算术平均数方法计算。用公式表示为:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{8.1}$$

式中，$\bar{y}$ 代表序时平均数，y_i 代表各期发展水平，n 代表时期项数。

(2)时点指标的序时平均数。从理论上说，要准确计算时点指标序时平均数，应当掌握现象在每一时点上的数据。这时，时点数列序时平均数的计算公式就是：$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$。但时点指标大多数是间断统计的。有的是每隔一定时间统计一次，如每月末、每季末、每半年末、每年末统计一次；有的是当现象发生变动时才登记一次；有的则是不定期统计。对这些不同的资料情况，时点指标时间序列的序时平均数计算有相应的方法，下面逐一介绍：

①间隔相等的时点数列序时平均数的计算。间隔相等的时点数列序时平均数的计算采用"首末折半法"，公式如下：

$$\bar{y}=\frac{\frac{y_1}{2}+y_2+\cdots+y_{n-1}+\frac{y_n}{2}}{n-1} \tag{8.2}$$

【例 8 1】根据表 8-1 的数据计算"九五"期间我国年平均人口数。

解：

表 8-1 人口数资料是各年年末人口数据，用公式 8.2 计算 1996～2000 年平均人口，首先要考虑一下它的首项应该是哪一年的数据。显然，首项不应是 1996 年年末人口数，而应是 1995 年年末人口数。这是因为 1996 年年末人口数是年末时点的人口数，它不能作为 1996 年人口数的代表值，而应以 1996 年年平均人口数作为代表值。我们可以把 1995 年年末的人口数看做是 1996 年年初的人口数，那么，1996 年的年平均人口就是年初和年末人口数的简单平均数(当然，这又是建立在假设两个时点间人口呈均匀变动的基础上)。即：

$$1996\text{ 年平均人口}=\frac{121\ 121+122\ 389}{2}=121\ 755(\text{万人})$$

类似地可以计算 1997～2000 年各年平均人口数。不难推知，公式 8.2 的实质就是对各时点的平均数求平均数，即：

$$\bar{y}=\frac{\frac{y_1+y_2}{2}+\frac{y_2+y_3}{2}+\cdots+\frac{y_{n-1}+y_n}{2}}{n-1}$$

$$=\frac{\frac{y_1}{2}+y_2+\cdots+y_{n-1}+\frac{y_n}{2}}{n-1}$$

因此，1996～2000 年年平均人口数为：

$$y=\frac{\frac{121\ 121}{2}+122\ 389+123\ 626+124\ 810+125\ 909+\frac{126\ 583}{2}}{6-1}=124\ 117.2(\text{亿人})$$

②间隔不等的时点数列序时平均数的计算。时点间隔不等的时间序列计算序时平均数的方法,其思路与时点间隔相等的时点序列相同。但因为这时时点间隔不同,所以要用时点间隔为权数来加权计算。下面以一例子说明间隔不等的时点数列序时平均数的计算。

【例 8－2】某省 2000 年有关人口统计数据资料如表 8－2 所示,求年平均人口。

表 8－2　某省 2000 年有关人口统计数据　　单位:万人

	1 月 1 日	3 月 31 日	6 月 1 日	12 月 31 日
人口数	2 704	2 716	2 698	2 702

解:

$$\bar{y}=\frac{\frac{2\,704+2\,716}{2}\times 3+\frac{2\,716+2\,698}{2}\times 2+\frac{2\,698+2\,702}{2}\times 7}{12}$$

$$=\frac{32\,444}{12}=2\,703.7(\text{万人})$$

由此,不难归纳出间隔不等的时点数列序时平均数的计算公式:

$$\bar{y}=\frac{\sum \bar{y}_i f_i}{\sum f_i}$$

$$=\frac{\frac{1}{2}(y_1+y_2)f_1+\frac{1}{2}(y_2+y_3)f_2+\cdots+\frac{1}{2}(y_{n-1}+y_n)f_{n-1}}{f_1+f_2+\cdots+f_{n-1}} \tag{8.3}$$

③当现象发生变动时登记一次的时点数列序时平均数的计算。仍以一例子来说明这种情况的时点数列序时平均数的计算方法。

【例 8－3】某企业 2000 年 5 月某种工具库存量资料如表 8－3 所示,求 5 月份工具平均库存量。

表 8－3　某企业 2000 年 5 月某种工具库存量资料　　单位:件

	1 日	10 日	12 日	20 日	25 日	31 日
库存量	85	5	105	50	20	100

解:

$$\bar{y}=\frac{85\times 9+5\times 2+105\times 8+50\times 5+20\times 6+100\times 1}{9+2+8+5+6+1}=\frac{2\,085}{31}=67.3(\text{件})$$

这个例子的情况与上面两例有所不同。在上面两例中,后一时点的数据是前一时点数据逐渐变化的结果,所以,当我们假设这种变动是均匀变动时,可计算两个时点数据的简单平均数作为两个时点之间的代表值。而本例中后一时点的数据并非前一时点数据逐渐变化的结果,而是前一时点现象一直维持到后一时点的前一天,在后一时点的那一天库存量才发生变化。因此,在两个时点之间的代表值就是前一时点的数据。由

此，可归纳出计算公式：

$$\bar{y}=\frac{y_1f_1+y_2f_2+\cdots+y_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}y_if_i}{\sum_{i=1}^{n}f_i} \tag{8.4}$$

2.以相对指标或平均指标时间序列计算序时平均数。计算相对指标或平均指标时间序列，一般不能就时间序列中的相对指标或平均指标直接计算，而要分别计算出相对数或平均数分子和分母的平均数后，再进行对比。用公式表示如下：

$$y=\frac{\bar{a}}{\bar{b}} \tag{8.5}$$

公式中，$\bar{a}$ 代表作为分子的时间序列序时平均数；$\bar{b}$ 代表作为分母的时间序列序时平均数。

【例8－4】某商厦2000年第一季度各月份有关商品销售资料如表8－4所示，求第一季度商品月平均流转次数。

表8－4 第一季度有关商品销售资料

	1月	2月	3月
商品销售额(a)	120	143	289
平均库存额(b)	60	65	85
商品流转次数(y)	2.0	2.2	3.4

解：

$$商品流转次数=\frac{商品销售额}{平均库存额}$$

$$\bar{y}=\frac{\bar{a}}{\bar{b}}=\frac{(120+143+289)\div 3}{(60+65+85)\div 3}=\frac{184}{70}=2.63(次)$$

计算结果表明，该商厦第一季度商品月平均流转次数为2.63次。

二、增长水平与平均增长水平

(一)增长水平

1.增长水平。增长水平也称增长量，是报告期发展水平与基期发展水平之差。用公式表示就是：增长量＝报告期水平－基期水平。

增长量有逐期增长量和累积增长之分。逐期增长量是报告期水平与前一期水平之差，说明报告期比前一期增长的绝对数量；累积增长量是报告期水平与某一固定时间的水平(通常为最初水平)之差，说明报告期比某一固定时间增长的绝对数量，即某一段较长时期内的总增长量。这两个指标用公式表示如下：

逐期增长量为

$$y_1-y_0,y_2-y_1,y_3-y_2,\cdots,y_n-y_{n-1} \tag{8.6}$$

累积增长量为

$$y_1 - y_0, y_2 - y_0, y_3 - y_0, \cdots, y_n - y_0 \tag{8.7}$$

不难看出,逐期增长量和累积增长量存在以下数量依存关系:

累积增长量等于相应的时期的逐期增长量之和,即:

$$y_n - y_0 = y_1 - y_0 + y_2 - y_1 + y_3 - y_2 + \cdots + y_n - y_{n-1} \tag{8.8}$$

相邻两期的累积增长量之差等于相应时期的逐期增长量:

$$(y_i - y_0) - (y_{i-1} - y_0) = y_i - y_{i-1} \tag{8.9}$$

【例 8-5】根据表 8-1 我国年末人口数资料,计算人口逐期增长量和累积增长量,如表 8-5 所示。

表 8-5 我国年末人口数 单位:万人

	1995	1996	1997	1998	1999	2000
年末人口数(y_i)	121 121	122 389	123 626	124 810	125 909	126 583
逐期增长量($y_i - y_{i-1}$)	—	1 268	1 237	1 184	1 099	674
累积增长量($y_i - y_0$)	—	1 268	2 505	3 689	4 788	5 462

在实际统计分析中,为了消除季节变动的影响,也常常计算年距增长量指标。

年距增长量是本期发展水平与上年同期发展水平之比,其公式如下:

年距增长量 = 本期发展水平 - 上年同期发展水平

【例 8-6】某市 2002 年春节旅游黄金周接待游客 47.59 万人次,2001 年同期接待 43 万人次,求春节旅游黄金周接待游客年距增长量。

解:

年距增长量 = 47.59 - 43 = 4.59(万人次)

(二)平均增长水平

平均增长水平也称平均增长量,它是逐期增长量的序时平均数。计算平均增长量可以将各逐期增长量相加除以逐期增长量个数,用简单算术平均法计算;也可以将累积增长量除以时间序列项数减 1。用公式表示如下:

$$\text{平均增长量} = \frac{\text{逐期增长量之和}}{\text{逐期增长量个数}} = \frac{\text{累计增长量}}{\text{时间序列项数} - 1} \tag{8.10}$$

根据例 8-5 资料,计算我国 1996~2000 年年平均增加人口:

$$\text{平均增长量} = \frac{1\,268 + 1\,237 + 1\,184 + 1\,099 + 674}{6 - 1} = \frac{5\,462}{5} = 1\,092.4(\text{万人})$$

第三节 时间序列的速度分析

一、发展速度与增长速度

(一)发展速度

发展速度是反映社会经济发展快慢的相对指标。计算公式如下：

$$发展速度=\frac{报告期水平}{基期水平}\times 100\% \tag{8.11}$$

在通常情况下，报告期水平总是大于0，因此发展速度指标值也总是表现为正数。当发展速度指标值大于0小于1时，表明报告期水平低于基期水平；当发展速度指标值等于1或大于1时，表明报告期水平达到或超过基期水平。由于采用的基期不同，发展速度有环比发展速度和定基发展速度之分。所谓环比发展速度也称逐期发展速度，是报告期水平与前一期水平之比。所谓定基发展速度则是报告期水平与某一固定时期水平(通常为最初水平或特定时期水平)之比，表明现象在较长时期内总的发展速度，也称为总速度。环比发展速度和定基发展速度的计算公式如下：

环比发展速度为

$$\frac{y_1}{y_0},\frac{y_2}{y_1},\cdots,\frac{y_n}{y_{n-1}} \tag{8.12}$$

定基发展速度为

$$\frac{y_1}{y_0},\frac{y_2}{y_0},\cdots,\frac{y_n}{y_0} \tag{8.13}$$

不难看出，环比发展速度和定基发展速度存在以下两个数量关系：

1. 环比发展速度的连乘积等于相应时期的定基发展速度。

$$\frac{y_1}{y_0}\times\frac{y_2}{y_1}\times\cdots\times\frac{y_n}{y_{n-1}}=\frac{y_n}{y_0} \tag{8.14}$$

2. 相邻时期的定基发展速度之商等于相应时期的环比发展速度。

$$\frac{y_i}{y_0}\div\frac{y_{i-1}}{y_0}=\frac{y_i}{y_{i-1}} \tag{8.15}$$

根据以上两个数量关系式，可以进行相互推算。

【例8-7】某产品外贸进出口量各年环比发展速度资料如下，1996年为103.9%，1997年为100.9%，1998年为95.5%，1999年为101.6%，2000年为108%，试计算

2000 年以 1995 年为基期的定基发展速度。

解：

$$\frac{y_{2000}}{y_{1995}}=\frac{y_{1996}}{y_{1995}}\times\frac{y_{1997}}{y_{1996}}\times\frac{y_{1998}}{y_{1997}}\times\frac{y_{1999}}{y_{1998}}\times\frac{y_{2000}}{y_{1999}}$$

$$=103.9\%\times100.9\%\times95.5\%\times101.6\%\times108\%=109.86\%$$

在实际统计分析工作中，为了消除季节变动的影响，类似于前面的年距增长量指标，也常计算年距发展速度，用以说明报告期水平与上年同期水平对比达到的相对程度。用公式表示如下：

$$\text{年距发展速度}=\frac{\text{报告期发展水平}}{\text{上年同期发展水平}}\times100\% \tag{8.16}$$

根据例 8－6 资料计算，该市春节旅游黄金周接待游客人次的年距发展速度为：47.59÷43＝110.67%。

（二）增长速度

增长速度是反映增长程度的相对指标。增长速度与发展速度有密切关系，两者相差一个基数（1 或 100%）。增长速度的计算公式如下：

$$\begin{aligned}\text{增长速度}&=\text{发展速度}-100\%\\&=\frac{\text{报告期水平}}{\text{基期水平}}-100\%\\&=\frac{\text{报告期水平}-\text{基期水平}}{\text{基期水平}}\\&=\frac{\text{增长量}}{\text{基期水平}}\times100\%\end{aligned} \tag{8.17}$$

上面几种计算方法可根据实际掌握资料的情况选择应用。不难看出，增长速度指标值有可能为正数，也有可能为负数。

与发展速度类似，由于采用的基期不同，也可以有环比增长速度和定基增长速度之分。前者表示现象的逐期增长速度，后者表示在较长时期内总的增长速度。与发展速度有所不同的是，环比发展速度和定基发展速度之间可以相互推算，环比增长速度和定基增长速度之间则不能直接相互推算。要进行环比增长速度和定基增长速度之间的推算，要先把它们还原成发展速度后，才能进行推算。环比增长速度和定基增长速度的计算公式如下：

$$\begin{aligned}\text{环比增长速度}&=\text{环比发展速度}-100\%\\&=\frac{\text{逐期增长量}}{\text{前期水平}}\times100\%\end{aligned} \tag{8.18}$$

$$\begin{aligned}\text{定基增长速度}&=\text{定基发展速度}-100\%\\&=\frac{\text{累积增长量}}{\text{最初水平}}\times100\%\end{aligned} \tag{8.19}$$

【例8-8】某企业几年来产量不断增长。已知1996年比1995年增长20%，1997年比1995年增长50%，1998年比1997年增长25%，1999年比1998年增长15%，2000年比1995年增长132.5%。试计算表8-6的空缺数字。

解：

将已知资料列入表8-6。

表8-6 某企业1996~2000年产量增长速度已知资料

年 份	1996	1997	1998	1999	2000
环比增长速度(%)	20	②	25	15	⑤
定基增长速度(%)	①	50	③	④	132.5

表中：

①1996年定基增长速度=20%；

②1997年环比增长速度=[(1+50%)÷(1+20%)]-1=25%；

③1998年定基增长速度=[(1+20%)×(1+25%)×(1+25%)]-1=87.5%；

④1999年定基增长速度=[(1+87.5%)×(1+15%)]-1=115.6%；

⑤2000年环比增长速度=[(1+132.5)÷(1+115.6%)]-1=7.8%。

二、平均发展速度与平均增长速度

平均发展速度是各个时期环比发展速度的序时平均数，说明社会经济现象在较长时期内速度变化的平均程度。平均发展速度与平均增长速度指标的数量关系是：平均增长速度=平均发展速度-1。

平均发展速度是各个时期环比发展速度的平均数。在实际工作中有两种计算平均发展速度的方法，即几何平均法和方程式法。

(一)几何平均法(水平法)

现象发展的平均速度一般用几何平均法计算，计算公式如下：

$$\bar{x}=\sqrt[n]{x_1 \cdot x_2 \cdot x_3 \ldots x_n}=\sqrt[n]{\prod x} \tag{8.20}$$

式中，$\bar{x}$ 表示平均发展速度，x 表示各年环比发展速度，n 表示环比发展速度的项数。

用几何平均法公式计算平均发展速度的数学依据是：现象发展的总速度不等于各年发展速度之和，而等于各年环比发展速度的连乘积。用 y_i 表示各期的发展水平，公式(8.20)可以表现为下面的形式：

$$\bar{x}=\sqrt[n]{\frac{y_1}{y_0}\times\frac{y_2}{y_1}\times\frac{y_3}{y_2}\ldots\frac{y_n}{y_{n-1}}}=\sqrt[n]{\frac{y_n}{y_0}} \tag{8.21}$$

不难看出，几何平均法名义上是各个环比发展速度的几何平均数，但实际上也是由

最初和最末两期的水平所决定的。只要最末水平决定后，中间各期的水平变化对平均发展速度的计算结果并没有影响。所以，平均发展速度的几何平均法也称为水平法。

以上两个计算公式，可以根据所掌握资料的情况来选择应用。

【例 8－9】根据例 8－7 的数据，按水平法计算平均发展速度。

解：

$$\bar{x}=\sqrt[n]{x_1\cdot x_2\cdot x_3\cdots x_n}=\sqrt[n]{\prod x}$$
$$=\sqrt[5]{103.9\%\times100.9\%\times95.5\%\times101.6\%\times108\%}$$
$$=\sqrt[5]{109.86\%}=101.90\%$$

如果现象的发展过程划分了几个时期，又具有各时期的平均速度指标，要计算全期平均发展速度，则要以各时期持续的年数为权数，按加权几何平均法计算。计算公式如下：

$$\bar{x}=\sqrt[\Sigma f]{\prod(\bar{x})^f} \tag{8.22}$$

【例 8－10】某地区 GDP“九五”前三年平均发展速度为 112%，后两年平均发展速度为 109%，求该地区“九五”期间 GDP 年平均发展速度和平均增长速度。

解：

$$\text{平均发展速度 }\bar{x}=\sqrt[\Sigma f]{\prod(\bar{x})^f}=\sqrt[3+2]{(112\%)^3(109\%)^2}=110.79\%$$
$$\text{平均增长速度}=\text{平均发展速度}-1=110.79\%-1=10.79\%$$

（二）方程式法（累计法）

方程式法是按照这样的要求来计算的，即时间序列中各年发展水平的总和等于全期的总水平，而各年发展水平是基期水平与各该年定基发展速度的乘积。根据定基发展速度等于环比发展速度连乘积的关系，各年发展水平也是基期水平和有关各年环比发展速度的乘积。这样，把各年环比发展速度加以平均化，列出方程式，求解即得出年平均发展速度。将以上的文字表述用数学关系式来表达，即为：

$$y_1+y_2+y_3+\cdots+y_n=\sum_{i=1}^{n}y_i$$

$$y_0\frac{y_1}{y_0}+y_1\frac{y_2}{y_1}+y_2\frac{y_3}{y_2}+\cdots+y_{n-1}\frac{y_n}{y_{n-1}}=\sum_{i=1}^{n}y_i$$

用 z_i 表示各期的发展速度，则上式可表示为：

$$y_0z_1+y_0z_2+y_0z_3+\cdots+y_0z_n=\sum_{i=1}^{n}y_i$$

把环比发展速度 x_i 代入上式：

$$y_0x_1+y_0x_1x_2+y_0x_1x_2x_3+\cdots+y_0x_1x_2x_3\cdots x_n=\sum_{i=1}^{n}y_i$$

把式中的各期环比发展速度平均化，用平均发展速度 $\bar{x}$ 取代各环比发展速度 x_i，

即：

$$y_0\bar{x} + y_0\bar{x}\cdot\bar{x} + y_0\bar{x}\cdot\bar{x}\cdot x + \cdots + y_0\bar{x}\cdot x\cdot\bar{x}\cdots\bar{x} = \sum_{i=1}^{n} y_i$$

$$y_0[(\bar{x}) + (\bar{x})^2 + (x)^3 + \cdots + (\bar{x})^n] = \sum_{i=1}^{n} y_i$$

$$(\bar{x}) + (\bar{x})^2 + (\bar{x})^3 + \cdots + (x)^n = \frac{\sum_{i=1}^{n} y_i}{y_0} \tag{8.23}$$

这个方程式的正根就是所求的年平均发展速度。用方程式法计算平均发展速度的特点，是着眼于各期水平累计之和$\sum_{i=1}^{n} y_i$，所以它也称为累计法。由于(8.23)式是一个高次方程，求解比较复杂。实际应用中，可以通过查找事先编好的累计方程法《平均增长速度查对表》(以下简称《查对表》)，去求平均增长速度。其一般步骤如下：

第一，计算$\frac{\sum y_i}{y_0}$的值。由于$\frac{y_i}{y_0} = z_i$，所以这个数值可以根据各年发展水平计算，也可以根据各年定基发展速度 z_i 计算。

第二，判断现象的发展类型，并从《查对表》中的相应部分找出所需的数值。当$\frac{\sum y_i}{y_0} > n$时，判断现象为递增型，在表中的递增部分"n"所在栏找出$\frac{\sum y_i}{y_0}$的值，与这个数值相对应的左边栏内的百分比，即为所求的年平均增长速度。当$\frac{\sum y_i}{y_0} < n$ 时，判断现象为递减型，则在表中递减部分查找，方法同上。

要注意的是，如果表中没有确切的平均增长速度与$\frac{\sum y_i}{y_0}$相对应，则找出$\frac{\sum y_i}{y_0}$的上、下界所对应的平均增长速度，然后按比例推算出$\frac{\sum y_i}{y_0}$所对应的平均增长速度。

【例 8－11】某地区"九五"期间固定资产投资额资料如表 8－7 所示，用方程式法计算各年平均发展速度。

表 8－7　　单位：百万元

	1995	1996	1997	1998	1999	2000
固定资产投资额	1 074	1 176	1 343	1 574	1 554	1 702

解：

$$\frac{\sum_{i=1}^{n} y_i}{y_0} = \frac{1\,176 + 1\,343 + 1\,574 + 1\,554 + 1\,702}{1\,074} = \frac{7\,349}{1\,074} = 6.842\,6 = 684.26\%$$

由于 684.26% >5，所以为递增型。查表，684.26% 介于 683.33% 和 685.28% 之间，对应的平均增长速度是 10.6% 和 10.7%。按比例推算，该地区"九五"期间固定资

产投资额平均增长速度为：

$$10.6\% + \frac{684.26\% - 683.33\%}{685.28\% - 683.33\%} \times (10.7\% - 10.6\%) = 10.65\%$$

（三）应用几何平均法与方程式法计算平均发展速度时要注意的问题

几何平均法和方程式法的数理论据、计算方法和应用条件有所不同。几何平均法的侧重点是从最末水平出发来进行研究，按照几何平均法所确定的平均发展速度推算的最末一年发展水平，与实际资料最末一年的发展水平相同。方程式法的侧重点则是从各年发展水平的累计总和出发来进行研究，按照方程式法所确定的平均发展速度推算的全期各年发展水平的总和，与全期各年的实际发展水平的总和相同。

我国制定国民经济发展长期计划，大致也有两种规定指标数值的方法。一种是以长期计划的最后一年应达到的水平来规定，如人口数、国内生产总值、工业主要产品产量、社会消费品零售总额等等；另一种是以整个计划期应达到的累计数来规定，如固定资产投资额等。在计算平均发展速度时，前者应采用几何平均法，后者应采用方程式法。

三、速度分析与水平分析的结合与应用

时间序列的速度指标是由水平指标对比计算而来的，以百分数表示的抽象化指标。速度指标把现象的具体规模或水平抽象掉了，因此不能反映现象的绝对量差别。在应用速度指标进行分析时，要注意以下几个问题：

第一，要结合具体研究目的适当选择基期，并注意其所依据的基本指标在整个研究时期的同质性。如果资料中有几年的环比增长速度特别快，而有几年又是负增长，出现显著的悬殊和不同的发展方向，以及所选择的最初水平和最末水平受特殊因素的影响过高或过低，用这样的资料来计算平均发展速度，就会降低甚至失去指标的代表意义和实际分析意义。

第二，要联系各个时期的环比发展速度来补充说明平均发展速度。如几何平均法名义上是各个时期环比发展速度的平均数，但实际上只计算最末水平和最初水平两个数字，把中间各个时期的具体变动抽象掉了，所以有必要补充各期的环比速度加以分析。

第三，要结合基期水平进行分析。因为发展速度是由报告期水平除以基期水平得来的，从数量关系来看，基期水平低，速度就容易达到较高；基期水平高，就难以达到高速度。因此，速度高可能掩盖低水平，速度低可能隐藏着高水平。例如，1949 年我国的钢产量为 15.8 万吨，1950 年为 61 万吨，则 1950 年我国钢产量的环比发展速度为 386.08%。1999 年我国钢产量为 12 426 万吨，2000 年为 12 850 万吨，2000 年钢产量的

环比发展速度为103.41%。从发展速度分析,前者大于后者。但从发展速度指标后面的发展水平指标分析,却是后者远远大于前者。2000年我国钢产量每增长1%的绝对量为124.26万吨;而1950年钢产量每增加1%的绝对量仅为0.16万吨。

增长1%的绝对量即前期水平的百分之一,是一个既考察速度又兼顾水平的指标,用公式表示为:

$$增长1\%的绝对量=\frac{a_t-a_{t-1}}{\left(\frac{a_t}{a_{t-1}}-1\right)\times 100}=\frac{a_{t-1}}{100} \tag{8.24}$$

第四,平均速度指标应结合其所依据的各个基本指标,如发展水平、增长量、环比发展速度、定基发展速度等进行分析研究,才能深入了解现象的全面发展、具体过程和特点,从而对研究现象具有比较确切和完整的认识。

第四节 长期趋势分析

一、时间序列的构成因素

在时间序列中,各时期的发展水平是受各种因素共同影响的结果。归纳起来,这些影响因素可以分为四类:

第一,长期趋势。它也称趋势变动,指时间序列在较长时期中所表现出来的总态势。长期趋势变动是受某种根本性的支配因素影响,而呈现出各时期的发展水平不断递增或不断递减或水平变动的基本趋势。例如,我国人口数时间序列呈现长期递增趋势。

第二,季节变动。季节变动的最基本的含义是指受自然界季节更替影响而发生的年复一年的有规律的变化。例如,农产品的生产、某些消费品的销售具有很典型的季节变动表现。在实际分析中,季节变动概念已有了扩展,一年内由于社会、政治、经济、自然因素影响形成的有规律的周期性的重复变动都称为季节变动。例如,学校寒暑假制度所带来的客运部门客流量在一年中的规律性变化,上下班制度对城市市内公交所带来的一天中客流量高峰的规律性变化等都称之为季节变动。

第三,循环变动。循环变动是指变动周期大于一年的有一定规律性的重复变动,如商业周期的繁荣、衰退、萧条、复苏四个阶段的循环变动。循环变动的规律不如季节变动明显,一般较难识别。

第四,不规则变动。它也称随机变动,指现象受众多偶然因素的影响而出现的不规则变动。

时间序列分析的任务之一,就是对时间序列中这几个影响因素进行测定,从而揭示现象变动的规律和特征,为认识和预测事物的发展提供依据。

上述四种因素,按照它们的影响方式的不同,可以设定不同的组合模型,其中最常见的有乘法模型和加法模型。乘法模型是假定四个因素对现象发展的影响是相互的,加法模型则假定各因素的影响是独立的。

乘法模型:$Y = T \cdot S \cdot C \cdot I$

加法模型:$Y = T + S + C + I$

公式中,Y 表示时间序列指标数值;T 表示长期趋势;S 表示季节变动;C 表示循环变动;I 表示不规则变动。

二、时距扩大法

时距扩大法是测定长期趋势最原始、最简单的方法。它是将原来时间序列中较小时距单位的若干个数据加以合并,得出较大时距单位的数据。扩大了时距单位的数据可以使较小时距单位数据所受到的偶然因素的影响相互抵消,而显示出现象变动的基本趋势。

【例 8-12】1999 年到 2001 年某城市旅游人数资料如表 8-8 所示。

表 8-8 某风景旅游城市旅游人数资料

年份	旅游人数(万人)			
	第一季	第二季	第三季	第四季
1999	32	40	61	28
2000	41	51	74	36
2001	57	65	93	57

从表 8-8 中 1999~2001 年 12 个季度的资料来看,该城市旅游人数有升有降,既有明显的季节变动,也有着逐年增长的趋势。为了显示长期趋势,必须将其他因素剔除。本例的变动周期为四季,把以季为单位的数据合并为以年为单位的数据,这时旅游人数有升有降的周期变化就不见了,显示出来的是长期增长的趋势。如表 8-9 所示。

表 8-9 时距扩大法计算表 单位:万人

年份	旅游人数	季平均旅游人数
1999	161	40
2000	202	51
2001	272	68

时距扩大法的优点是简便直观。但它的缺点也很突出，表现在时距扩大之后，所形成的新数列包含的数据减少，信息量大大流失，不便于做进一步的分析。

三、移动平均法

移动平均法是对时距扩大法的一种改良。它是采用逐期递推移动的办法对原数列按一定时距扩大，得出一系列扩大时距的序时平均数。

移动平均法有多种形式，下面介绍简单移动平均法和加权移动平均法。

(一)简单移动平均法

该方法依次计算若干项时间序列数据的移动平均数，并将其作为移动平均中项的趋势测定值，所以这种方法也称为中心化移动平均法。

表8-10是根据表8-8数据计算的四季移动平均数列：

表8-10 移动平均计算表

年份	季度	序号	y_t(万人)	四季移动平均(T)(万人)
1999	1	1	32	—
	2	2	40	—
	3	3	61	41.4
	4	4	28	43.9
2000	1	5	41	46.9
	2	6	51	49.5
	3	7	74	52.5
	4	8	36	56.3
2001	1	9	57	60.4
	2	10	65	65.4
	3	11	93	—
	4	12	57	—

由表8-10数据可以看到，该城市旅游人数的变动趋势呈稳步上升态势。

应用简单移动平均法分析长期趋势时，有几点需要注意：

1.凡采用奇数项求移动平均数，只需移动平均一次。若采取偶数项移动平均，则必须进行两次移动平均。

设时间序列的指标值为 y，其指标值顺次为 $y_1, y_2, y_3, \cdots, y_n$。如取四项移动平均，则所得的第一个移动平均数 $\bar{y}_{2-3}=\frac{y_1+y_2+y_3+y_4}{4}$，对应于原数列的第二年与第

三年之间；第二个移动平均数 $\bar{y}_{3-4}=\frac{y_2+y_3+y_4+y_5}{4}$，对应于第三年与第四年之间；依次类推。因而尚需再进行一次移动平均，即将第一个移动平均数与第二个移动平均数再平均，求出一个新的平均数 $\bar{y}_3=\frac{\bar{y}_{2-3}+\bar{y}_{3-4}}{2}=\left(\frac{1}{2}y_1+y_2+y_3+y_4+\frac{1}{2}y_5\right)\div 4$，对应于原数列的第三年。以后各年数字，照此类推。

2. 逐项移动平均的一个关键的问题是取多长时间为宜。假如现象变动有周期性，就应以周期长度为移动平均的项数。如上例，以一年四个季度为一周期，就应取四项移动平均。因为只有这样，才能消除周期变动，准确反映长期趋势。

3. 移动平均后所得的修匀数列，比原数列的项数更少。奇数项移动平均所形成的新数列，首尾各少 $(N-1)/2$ 项（N 为移动项数）；偶数项移动平均所形成的新数列，首尾各少 $N/2$ 项（N 为移动项数）。如二年或三年移动平均，首尾各少一项；四年或五年移动平均，首尾各少二项。平均项数越多失去的信息也越多，所以，移动平均的项数不宜过大。

（二）加权移动平均法

简单移动平均法适用于线性趋势的测定，如果社会经济现象发展是非线性的，就要考虑用加权移动平均法。所谓加权移动平均法就是对各期指标值进行加权后再进行移动平均。依照在简单移动平均中，移动平均数代表移动平均中项时期的长期趋势值的做法，在加权移动平均法中一般也采用奇数项加权移动平均，各期权数以二项展开式的系数为计算基础，使中项时期指标值的权数最大，两边对称，逐期减小。

例如，三项移动平均，应以 $(a+b)^2=a^2+2ab+b^2$ 的系数 1，2，1 进行加权。即：$y_{t-1}=(y_{t-2}+2y_{t-1}+y_t)/4$。

四、最小二乘法

（一）普通最小二乘法

普通最小二乘法在第七章回归分析中已经讨论过。运用于时间序列分析就是将自变量 x 用时间变量 t 代替。

当时间序列逐期增长量近似于常数时，现象的发展趋势就是直线型的。若各个时期的二级增长量大体相同，即各个时期的逐期增长量的增长量近似于一个常数，则趋势线近似于一条抛物线，应配合抛物线方程。若现象各年环比增长速度大体相同，则趋势线近似于一条指数曲线，应配合指数曲线方程。根据不同情况，还可以拟合时间序列的S型曲线或多项式方程。关于这些曲线函数的形式与拟合方法，第七章第五节中已有介绍，此处不再重复。

(二)加权最小二乘法

普通最小二乘法将时间序列中各项数据的重要性同等看待。而事实上各项数据对未来的影响作用是不同的。一般来说,近期数据比起远期数据对未来的影响更大。因此,较合理的方法就是使用加权的方法,对近期数据赋予较大的权数,对远期数据则赋予较小的权数。加权最小二乘法采用指数权数 W^{n-i} 的形式($0<W<1$),加权以后求得的参数估计值满足:

$$\sum_{i=1}^{n} W^{n-i}(y_i-\hat{y}_i)^2=min \quad (i=1,2,\cdots,n) \tag{8.25}$$

以直线模型为例,其加权的剩余平方和为:

$$\sum_{i=1}^{n} W^{n-i}(y_i-\hat{\beta}_1-\hat{\beta}_2 t)^2 \tag{8.26}$$

对上式分别求 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的偏导数,得到标准方程组:

$$\begin{cases} \sum W^{n-i}y_i=\hat{\beta}_1\sum W^{n-i}+\hat{\beta}_2\sum W^{n-i}t \\ \sum W^{n-i}ty=\hat{\beta}_1\sum W^{n-i}t+\hat{\beta}_2\sum W^{n-i}t^2 \end{cases} \tag{8.27}$$

对上述方程解出 $\hat{\beta}_1$ 和 $\hat{\beta}_2$,就得到加权最小二乘法直线模型。应用加权最小二乘法,W 的取值不同,解出的回归系数也不同,因此 W 值取多少,需要经研究分析后确定。

【例 8-13】某海关口岸征收关税额(百万元)资料如表 8-11 所示,试分别用简单最小二乘法和加权最小二乘法拟合线性趋势方程。

表 8-11 某海关口岸关税额普通最小二乘法计算表

年份	时间变量 t_i	关税额 y_i	t_i^2	t_iy_i
1991	1	85.6	1	85.6
1992	2	91.0	4	182.0
1993	3	96.1	9	288.3
1994	4	101.2	16	404.8
1995	5	107.0	25	535.0
1996	6	112.2	36	673.2
1997	7	119.6	49	837.2
1998	8	125.6	64	1 004.8
1999	9	136.5	81	1 228.5
2000	10	145.2	100	1 452.0
2001	11	155.7	121	1 712.7
合计	66	1 275.7	506	8 404.1

解:

(1)普通最小二乘法：

$$\hat{\beta}_2=\frac{\sum_{i=1}^{n}t_iy_i-\frac{1}{n}(\sum_{i=1}^{n}t_i)(\sum_{i=1}^{n}y_i)}{\sum_{i=1}^{n}t_i^2-\frac{1}{n}(\sum_{i=1}^{n}t)^2}=\frac{8\,404.1-\frac{1}{11}\times66\times1\,275.7}{506-\frac{1}{11}\times66^2}=6.82$$

$$\hat{\beta}_1=\bar{y}-\beta_2\bar{t}$$

$$=\frac{1\,275.7}{11}-6.82\times\frac{66}{11}=75.05$$

所拟合的直线趋势方程为：$\hat{y}_t=75.05+6.82t$。

(2)加权最小二乘法：

表 8－12　某海关口岸关税额加权最小二乘法计算表

年份	时间变量 t	关税额 y	$W=0.9$					
			$n-i$	W^{n-i}	$W^{n-i}y$	$W^{n-i}t$	$W^{n-i}ty$	$W^{n-i}t^2$
1991	1	85.6	10	0.349	29.87	0.349	29.87	0.349
1992	2	91.0	9	0.387	35.22	0.774	70.43	1.548
1993	3	96.1	8	0.43	41.32	1.29	123.97	3.87
1994	4	101.2	7	0.478	48.37	1.912	193.49	7.648
1995	5	107.0	6	0.531	56.82	2.655	284.09	13.275
1996	6	112.2	5	0.59	66.2	3.54	397.19	21.24
1997	7	119.6	4	0.656	78.46	4.592	549.2	32.144
1998	8	125.6	3	0.729	91.56	5.832	732.5	46.656
1999	9	136.5	2	0.81	110.57	7.29	995.09	65.61
2000	10	145.2	1	0.9	130.68	9	1 306.8	90
2001	11	155.7	0	1	155.7	11	1 712.7	121
合计	66	1 275.7		6.860	844.77	48.23	6 395.33	403.34

将表 8－12 的数据代入公式 8.27，得：

$$\begin{cases}844.77=\hat{\beta}_1 6.86+\beta_2 48.23\\6395.33=\beta_1 48.23+\beta_2 403.34\end{cases}$$

解得：$\hat{\beta}_1=73.24$，$\hat{\beta}_2=7.098$

所得趋势方程为：$\hat{y}=73.24+7.098t$

比较两种估计方法的结果见表 8－13。

表 8－13 两种估计方法的误差比较

年份	实际值 y	普通最小二乘法		加权最小二乘法	
		模拟值 $\hat{y}_1$	相对误差 $\frac{y-\hat{y}_1}{y}(\%)$	模拟值 $\hat{y}_2$	相对误差 $\frac{y-\hat{y}_2}{y}(\%)$
1991	85.6	81.87	0.044	80.34	0.0614
1992	91	88.69	0.0254	87.44	0.0391
1993	96.1	95.51	0.006	94.53	0.0163
1994	101.2	102.33	－0.0112	101.63	－0.0042
1995	107	109.15	－0.0201	108.73	－0.0162
1996	112.2	115.97	－0.0336	115.83	－0.0324
1997	119.6	122.79	－0.0267	122.93	－0.0278
1998	125.6	129.61	－0.032	130.03	－0.0353
1999	136.5	136.43	0.0005	137.12	0.0045
2000	145.2	143.25	0.0134	144.22	0.0067
2001	155.7	150.07	0.0362	151.32	0.0281

可以看出，加权最小二乘法的近期误差比远期误差小，这就显示了加权最小二乘法的优势。

五、趋势外推预测

测定长期趋势的一个重要目的就是要利用这一长期趋势对未来进行预测。进行趋势外推预测的常用方法有以下几种：

（一）移动平均法

与前面的中心化移动平均法不同的是，用于外推预测的移动平均数不代表移动平均中项的趋势值，而是代表预测期的趋势值。

1. 简单移动平均法。$t+1$ 期的预测值 $\hat{y}_{t+1}$ 的计算公式为：

$$\hat{y}_{t+1}=\frac{1}{N}(y_t+y_{t-1}+\cdots+y_{t-N+1})=\frac{1}{N}\sum_{i=0}^{N-1}y_{t-i} \tag{8.28}$$

式中，N 表示移动平均的项数。

这种方法的一次移动平均只有一期的预测能力。

【例 8－14】根据例 8－13 数据，用简单移动平均法作趋势外推预测，移动平均项数 $N=3$。

解：

用公式(8.28)对例 8－13 中的数据进行趋势外推预测，其结果见表 8－14。

表 8－14　某海关关税趋势外推预测计算表　　单位:百万元

年份	时间变量 t	关税额 y_t	趋势外推预测值 $\hat{y}_t$
1991			
1992	1	85.6	—
1993	2	91 0	—
1994	3	96.1	—
1995	4	101.2	90.9
1996	5	107.0	96.1
1997	6	112.2	101.4
1998	7	119.6	106.8
1999	8	125.6	112.9
2000	9	136.5	119.1
2001	10	145.2	127.2
2002	11	155.7	135.8
			145.8

2.加权移动平均法。采用加权的方法加大近期数据的权数,突出近期数据在预测中的影响作用。设 W_i 为 y_{t-i} 的权数,满足各期值对预测值的影响由近及远逐渐减小,有 $W_0>W_1>\cdots>W_{N-1}$。

$t+1$ 期的预测值 $\hat{y}_{t+1}$ 的计算公式为:

$$\hat{y}_{t+1}=\frac{W_0y_t+W_1y_{t-1}+\cdots+W_{N-1}y_{t-N+1}}{W_0+W_1+\cdots+W_{N-1}}=\frac{\sum_{i=0}^{N-1}W_iy_{t-i}}{\sum_{i=0}^{N-1}W_i} \tag{8.29}$$

（二）最小二乘法

也就是用最小二乘法拟合的趋势方程进行外推预测。

【例 8－15】根据例 8－13 中得出的普通最小二乘法趋势方程:

$\hat{y}_t=75.05+6.82t$ 和加权最小二乘法趋势方程:$\hat{y}=73.24+7.098t$,预测 2002 年与 2003 年的趋势值。

解:

(1)普通最小二乘:

$\hat{y}_{2002}=75.05+6.82\times12=156.89$(百万元)

$\hat{y}_{2003}=75.05+6.82\times13=163.71$(百万元)

(2)加权最小二乘:

$\hat{y}_{2002}=73.24+7.098\times12=158.42$(百万元)

$$\hat{y}_{2003}=73.24+7.098\times 13=165.51(\text{百万元})$$

(三)指数平滑法

指数平滑法是统计预测中广泛使用的一种方法,它可以直接用于预测,也可以用于估计模型参数。

指数平滑法也称指数修匀预测法,按修匀次数的多少有一次指数平滑、二次指数平滑、三次乃至多次指数平滑。下面讨论的是一次指数平滑法。

在时间序列中,以本期的实际数 y_t 和本期的预测数 $\hat{y}_t$ 为依据,然后赋予不同的权数,求得下一期预测数。用公式表示如下:

$$\hat{y}_{t+1}=\alpha y_t+(1-\alpha)\hat{y}_t \tag{8.30}$$

式中,$\hat{y}_{t+1}$ 表示下一期的预测数,y_t 和 $\hat{y}_t$ 分别表示 t 时期的实际数和预测数,α 为平滑系数,$0<\alpha<1$。

由于 $\alpha+(1-\alpha)=1$,所以该公式实际上是加权平均公式。指数平滑法实质上是加权移动平均法的改良。请看对(8.30)式的推演:

$$\begin{aligned}\hat{y}_{t+1}&=\alpha y_t+(1-\alpha)\hat{y}_t\\&=\alpha y_t+(1-\alpha)[\alpha y_{t-1}+(1-\alpha)\hat{y}_{t-2}]\\&=\alpha y_t+\alpha(1-\alpha)y_{t-1}+(1-\alpha)^2\hat{y}_{t-2}\\&=\alpha y_t+\alpha(1-\alpha)y_{t-1}+\alpha(1-\alpha)^2y_{t-2}+\alpha(1-\alpha)^3y_{t-3}+\cdots\end{aligned}$$

上式的各项系数是一个无穷等比数列,当 $t\to\infty$ 时,其总和 $\lim\limits_{t\to\infty}\dfrac{\alpha[1-(1-\alpha)^t]}{[1-(1-\alpha)]}=\lim\limits_{t\to\infty}[1-(1-\alpha)^t]=1$,因此可以把各项系数视为加权平均数中的相对权数。

但由于时间序列的项数是有限的,根据以上系数计算的结果只是一种近似的平均数。

不难看出,应用指数修匀法预测,α 值的确定是一个关键问题。因为权数的分配是按指数方式递减的,递减的速度决定于 α 值的大小。α 值越大时权数的递减速度越快,反之就越慢。所以 α 的取值与预测的效果有很密切的关系。确定 α 值前,通常可以先取各种值进行试算,然后再作出决定。从选择方法上看,应选取使误差最小的 α 值。如从数据的特征上看,变化呈阶梯式上升下降者取较大的 α 值;变化比较平稳的取较小的 α 值。

【例 8-16】假设某市各年彩电的销售量如表 8-15 所示。

表 8-15 中,$\alpha=0.9$,并以 1990 年、1991 年、1992 年三年实际数的简单平均数作为 1991 年的预测数,第三年的预测数为:

$$\hat{y}_3=0.9y_2+\hat{y}_2=0.9\times 536.9+0.1\times 530.2=536.2\text{ 万台}$$

表 8－15　某市各年彩电的销售量

年度	实际销售量(万台)y_t	预测销售量(万台) $\hat{y}_{t+1}=0.9y_t+(1-0.1)\hat{y}_t$
1990	450.5	—
1991	536.9	530.2
1992	603.2	536.2
1993	716.6	596.5
1994	842.6	704.6
1995	921.8	828.8
1996	995.1	912.5
1997	1 030.5	986.8
1998	1 096.8	1 026.1
1999	1 120.3	1 089.7
2000	1 220.7	1 117.2
2001	1 210.6	1 210.4
2002	—	1 210.6

其他类推。

从上例可以看出，应用指数平滑法预测下一期的数值，只需利用本期实际数和本期的预测数即可，不必知道整个长时期的时间数列。同时，为了减少误差，还可以通过 α 值的调整使得预测值适应实际值的变化。所以这种方法较适合用于短期预测。

第五节　季节变动与循环波动分析

一、季节变动分析

关于季节变动的含义，在上一节中我们已经作了介绍。季节变动具有三个明显的特征：有规律地变动，按一定的周期重复进行，每个周期变化大体相同。由于季节变动的最大周期为一年，所以以年份为单位的时间数列中不可能有季节变动。测定季节变动的方法很多，下面介绍较常用的同期平均法和趋势剔除法。

(一)同期平均法

这种方法是测定季节变动最简便的方法。它是以若干年资料数据求出同月(季)的平均水平与全年总月(季)水平,二者对比得出各月(季)的季节指数来表明季节变动的程度。按月(季)平均法可以分为直接按月(季)平均法和比率按月(季)平均法两种。

1.直接按月(季)平均法。直接按月(季)平均法将整个时间序列的趋势值视为常数。计算步骤如下:

第一,计算各年同月(季)的平均数 $\bar{y}_i$($i=1\sim12$ 月或 $i=1\sim4$ 季);

第二,计算各年所有月份(或季度)的总平均数 $\bar{y}$;

第三,计算季节指数 S_i,$S_i=\frac{\bar{y}_i}{\bar{y}}\times100\%$。 (8.31)

【例 8-17】以表 8-16 中 1999,2000 和 2001 年数据用直接按季平均法分析季节变动。

其分析过程见表 8-16 中后三行。

表 8-16 直接按月平均法季节指数计算表

年份	旅游人数(万人)				
	第一季	第二季	第三季	第四季	合计
1999	32	40	61	28	161
2000	41	51	74	36	202
2001	57	65	93	57	272
合计	130	156	228	121	635
同季平均	43.33	52	76	40.33	52.915
季节指数 S_i(%)	81.88	98.27	143.63	76.22	400

依照表 8-16 中分析所得的季节指数,做相应的人数季节变动图,如图 8-1 所示。

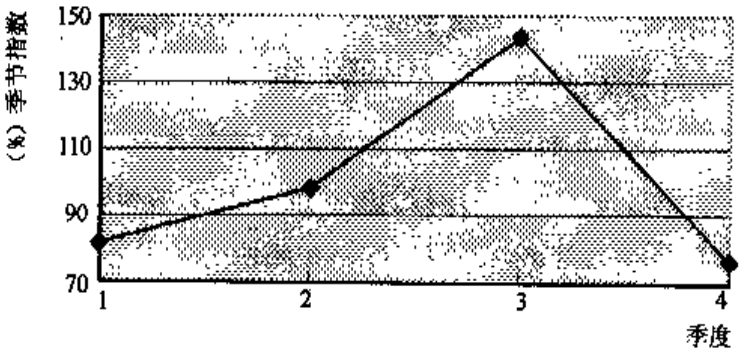


图 8-1 某风景旅游城市旅游人数季节变动

假设该城市 2002 年旅游人数将比 2001 年增长 1.5%,达到 276.08 万。利用季节

指数，可以对各季度的旅游人数进行预测。

第一季度预测值＝276.08÷4×81.88%＝56.514(万人)

第二季度预测值＝276.08÷4×98.27%＝67.826(万人)

第三季度预测值＝276.08÷4×143.63%＝99.133(万人)

第四季度预测值＝276.08÷4×76.22%＝52.607(万人)

2.比率按月(季)平均法。这种方法是在按月(季)平均之前，先将历年各月(季)的数据同其本年的月(季)平均数相比，得出说明该年度的季节比率；然后再将各年度同期(月或季)的比率进行平均，求出季节指数：

$$S_i=\frac{\sum_{i=1}^{N}\frac{y_{ij}}{\bar{y}}}{N}\times 100\% \tag{8.32}$$

【例8－18】仍以表8－16的数据，用比率按季平均法计算历年各季的年度季节比率和季节指数。

表8－17中各季数据是用各季度的数值除以年平均数而求得的。如用1999年第一季度旅游人数32万除以1999年季平均旅游人数40.25万得到0.795。

表8－17　比率按月平均法季节指数计算表

年份	第一季	第二季	第三季	第四季	合计
1999	0.795 0	0.993 8	1.515 5	0.695 7	4
2000	0.811 9	1.009 9	1.465 3	0.712 9	4
2001	0.838 2	0.955 9	1.367 6	0.838 3	4
合计	2.445 1	2.959 6	4.348 4	2.246 9	12
季节指数(%)	81.50	98.65	144.95	74.90	400

表8－17的计算结果与例8－17的计算结果类似，表中第三季度的季节指数最高，达到144.95%，表明第三季度是旅游旺季，旅游人数超过各季平均水平44.95%。

(二)移动平均趋势剔除法

在具有明显的长期趋势变动的数列中，为了测定季节变动，必须先将趋势变动因素加以剔除。假定趋势变动、季节变动、循环变动和不规则变动对时间序列影响可以用乘法模型 $Y=T\cdot S\cdot C\cdot I$ 反映，用移动平均趋势剔除法测定季节变动的步骤如下：

1.对原时间序列求移动平均数，作为相应时期的趋势值 T。

2.剔除原数列中的趋势变动 T，即将原数列各项除以移动平均数的对应时间数据：

$$\frac{T\cdot S\cdot C\cdot I}{T}=S\cdot C\cdot I \tag{8.33}$$

3.以消除趋势变动后的数列计算季节指数,测定季节变动。

【例8-19】以表8-8的数据用移动平均趋势剔除法分析季节变动。计算过程和结果见表8-18和表8-19。

表8-18 季节指数计算表(一)

年份	季度	顺序	Y_t	四季移动平均 T	$\frac{Y_t}{T}$
1999	1	1	32	—	—
	2	2	40	—	—
	3	3	61	41.4	1.473 4
	4	4	28	43.9	0.637 8
2000	1	5	41	46.9	0.874 2
	2	6	51	49.5	1.030 3
	3	7	74	52.5	1.409 5
	4	8	36	56.3	0.639 4
2001	1	9	57	60.4	0.943 7
	2	10	65	65.4	0.993 9
	3	11	93	—	—
	4	12	57	—	—

表8-19 季节指数计算表(二)

年份	第一季度	第二季度	第三季度	第四季度	合计
1999			1.473 4	0.637 8	
2000	0.874 2	1.030 3	1.409 5	0.639 4	
2001	0.943 7	0.993 9			
同季平均	0.909 0	1.012 1	1.441 5	0.638 6	1.000 3
季节指数(%)	90.87	101.18	144.11	63.84	400

二、循环变动的测定

如前所述,循环变动是指变动周期大于一年的有一定规律性的重复变动。循环变动不同于长期趋势,它所表现的不是朝着某一个方向持续上升或下降,而是从低到高,又从高到低周而复始近乎规律性地变动。循环变动也不同于季节变动,季节变动一般以一年、一季或一月等为一周期,可以预见。而循环变动没有固定的周期,一般都在数年以上,难以事先预知。因此,循环变动分析不仅要借助于统计方法,还要借助于定性

的经济分析。

从统计分析的角度来看,循环变动的测定方法有多种,如剩余法、直接法和循环平均法等。不同的方法得出的分析结论有一定的差异,这就需要对不同测定方法的基本原理、前提条件有所了解。下面介绍两种常用的方法:

(一)直接法

直接法适用于季度和月度时间序列。如果研究时间序列的目的只在于测定数列的循环波动特征,可用直接法进行分析。直接法是将每年各季或各月的数值与上年同期进行对比,即求出年距发展速度:

$$C \cdot I = \frac{y_t}{y_{t-4/12}} \tag{8.34}$$

直接法简便易行,可以大致消除趋势变动 T 和季节变动 S 的影响。它的主要局限性是在消除时间序列长期趋势的同时,相对放大了年度发展水平的影响,当某期发展水平偏低或偏高时,必然会影响 $C \cdot I$ 的数值,使之偏高或偏低,导致循环变动的振幅被拉大。

(二)剩余法

剩余法也称分解法。假定各因素对现象发展影响的组合模型为乘法模型:$Y = T \cdot S \cdot C \cdot I$,剩余法的基本思路是:利用分解分析的原理,在时间序列中剔除长期趋势和季节变动,然后再消除不规则变动,从而揭示循环变动的特征。即:

$$\frac{Y}{T \cdot S} = \frac{T \cdot S \cdot C \cdot I}{T \cdot S} = C \cdot I \tag{8.35}$$

将所得的循环变动和不规则变动的结果 $C \cdot I$ 进行移动平均,消除不规则变动 I,求得循环变动值 C。

第六节 Excel 在时间序列分析中的运用

一、季节变动、长期趋势与循环变动的测定

循环变动的测定需要在测定出季节变动和长期趋势的基础上进行,因此我们用一个测定循环变动的实例,综合说明 Excel 在时间序列分析中的运用。

【例 8-20】根据我国 1996 年到 2001 年货物周转量资料,用剩余法测定其循环变动。步骤如下:

1.输入数据,如图8-2所示。A2:A25存放时间,B2:B25存放实际值Y,A1与B1分别存放两列数据的标志"时间"和"Y"。C1至G1分别存放计算过程中各列数据的标志,分别为"S"、"T*C*I"、"T"、"C*I"和"C"。

	A	B	C	D	E	F	G
1	时间	Y	S	T*C*I	T	C*I	C
2	1996 1	6924	92.25440397	75.05332756			
3	2	7254	100.4076629	72.24548194			
4	3	7045	102.4171906	68.78728035			
5	4	7415	104.9207425	70.67239348			
6	1997 1	7346	92.25440397	79.62763493	71.6896208	111.072752	
7	2	7714	100.4076629	76.82680558	72.8331977	105.483225	105.69
8	3	7615	102.4171906	74.35275229	73.9785286	100.505854	101.15
9	4	7707	104.9207425	73.4554466	75.3698966	97.4599276	99.55
10	1998.1	7066	92.25440397	76.59254947	76.0656599	100.692677	100.14
11	2	7732	100.4076629	77.00607476	75.3068885	102.256349	106.66
12	3	9032	102.4171906	88.18832025	76.3517058	117.035599	111.78
13	4	9597	104.9207425	91.46904386	78.8105978	116.061858	117.30
14	1999.1	9131	92.25440397	98.97630473	83.3139971	118.799131	115.86
15	2	10063	100.4076629	100.2214343	88.9099359	112.722423	111.35
16	3	9946	102.4171906	97.11260332	94.7137758	102.532712	104.59
17	4	10020	104.9207425	95.50065849	96.9448466	98.5102993	98.49

图8-2

2.用直接平均法计算季节指数。在C2中输入公式:"=AVERAGE(B2,B6,B10,B14,B18,B22)/AVERAGE(B$2B$25)*100"。

公式中乘以100,是因为季节指数的单位是"%"。然后用填充命令或"填充柄"将公式复制到C3:C5单元格,此时已经计算出季节指数。再用"复制"、"选择性粘贴"命令将季节指数的值复制到其他年份。比如要将季节指数复制到1997年,先选定C2:C5单元格,执行菜单命令"编辑"→"复制",然后选中C6单元格,执行菜单命令"编辑"→"选择性粘贴",在调出的对话框中选择粘贴数值即可。

3.剔除季节变动,$T \cdot C \cdot I = Y/S$。在D2中输入公式"=B2/C2",然后将公式复制到D3:D25单元格。

4.用4项移动平均求出长期趋势值T。在E6单元格输入公式"=AVERAGE(D2:D5)",再将公式复制到E7:E25单元格。

计算移动平均也可用"移动平均"分析工具。此分析工具及其公式可以基于特定的过去某段时期中变量的均值,对未来值进行预测。移动平均值提供了由所有历史数据的简单的平均值所代表的趋势信息。

调出"移动平均"对话框后,其主要选项含义如下:

输入区域:在此输入待分析数据区域的单元格引用。该区域必须由包含四个以上数据单元格的单列组成。本例为"$D$1:$D$25"。

标志位于第一行：如果输入区域的第一行中包含标志项，请选中此复选框；如果输入区域没有标志项，请清除此复选框，Excel 将在输出表中生成适宜的数据标志。本例要选中该复选框。

间隔：在此输入用来进行移动平均计算的间隔数。默认间隔为 3。本例为“四项移动平均”，所以间隔为“4”。

输出区域：在此输入对输出表左上角单元格的引用。此分析工具的输出区域必须与输入区域在同一工作表中。因此，“新工作表组”和“新工作簿”选项均不可使用。

图表输出：选择此项可以在输出表中生成一个嵌入直方图。

标准误差：如果要在输出表的一列中包含标准误差值，请选中此复选框；如果只需要没有标准误差值的单列输出表，请清除此复选框。

本例，“移动平均”对话框的填写如图 8－3 所示。

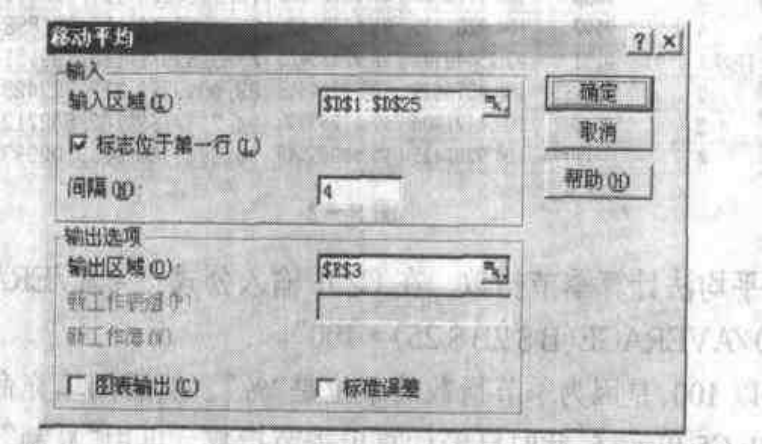


图 8－3

在此要注意输出区域的填写，因为用该分析工具计算的移动平均值直接排放在 N 个时期中的最后一期，而根据本题要求应放在 N 个时期的下一期。

5.剔除长期趋势，$C \cdot I = (T \cdot C \cdot I)/T$。在 F6 单元格中输入公式“＝D6/E6＊100”，然后将公式复制到 F7:F25。

6.进行移动平均(取 3 项)消除不规则变动。

进行三项移动平均消除不规则变动。在 G7 中输入公式“＝AVERAGE(F6:F8)”，然后将公式复制到 G8:G24。这一步也可用“移动平均”分析工具来完成。最后结果中循环变动 C 保留两位小数。结果如表 8－20 所示。

表 8-20 我国货物周转量实际值与循环变动测定值

时间	实际值 y_t（亿吨公里）	循环变动 C（%）	时间	实际值 y_t（亿吨公里）	循环变动 C（%）
1996.1	6924		1999.1	9131	115.80
2	7254		2	10063	111.36
3	7045		3	9946	104.76
4	7415		4	10020	98.60
1997.1	7346		2000.1	8533	99.02
2	7714	105.90	2	10063	99.72
3	7615	101.38	3	9946	101.25
4	7707	99.75	4	10020	106.59
1998.1	7066	100.27	2001.1	10673	109.67
2	7732	106.76	2	11237	112.59
3	9032	111.82	3	11561	106.89
4	9597	117.19	4	11734	

二、描绘循环曲线图

根据以上的结果，可作出我国货物周转量的循环变动曲线。步骤如下：

1. 执行菜单命令“插入”→“图表”，调出“图表向导－4”步骤之“1－图表类型”对话框。在对话框的“标准类型”页框的“图表类型”列表框中选择“折线图”，子图表类型选“数据点折线图”。单击“下一步”。如图 8－4 所示。

2. 进入“图表向导－4”步骤之“2－图表源数据”对话框。

在对话框的“数据区域”页框的“数据区域”文本框中，输入循环变动 C 所在区域的地址。可用鼠标拖动选择 G1：G25，则在“数据区域”文本框中相应显示“＝时间序列！$G$1：$G$25”（其中“时间序列”是该工作表的名称）。选中“列”单选框，表示系列产生在列。

单击“系列”页框，显示出“系列”页框的内容。确定分类（X）轴标志，可单击“分类（X）轴标志”文本框，再用鼠标拖动选择 A2：A25，此时在“分类（X）轴标志”文本框中显示“＝时间序列！$A$2：$A$25”。单击“下一步”。在“系列”页框中，可以修改系列。系列即是要作图的数据序列（如本题的循环变动 C）。修改或删除某系列前，要先在“系列”列表框中选中该系列。

3.进入到“图表向导－4”步骤之“3－图表选项”对话框。在该对话框中，可设置图表的标题、坐标轴的标题、网格线和图例等。一般不作修改，单击“下一步”。

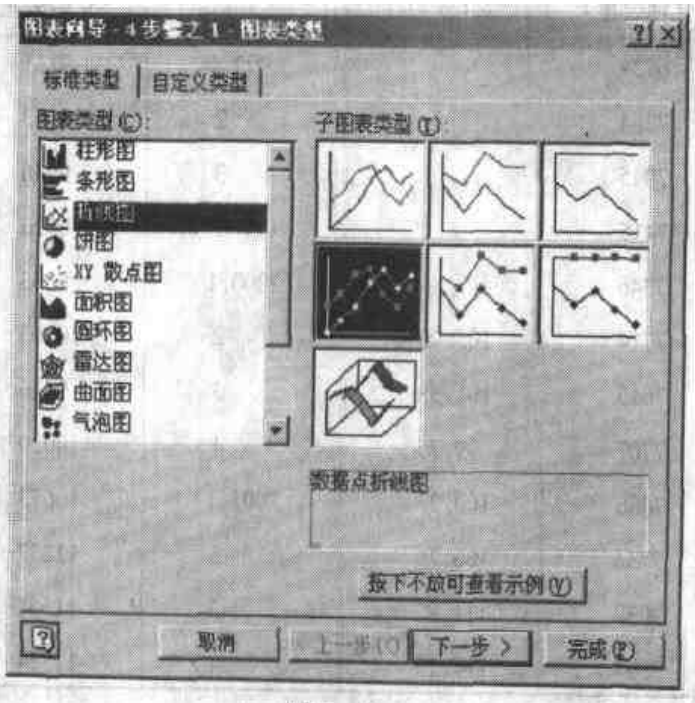


图 8－4

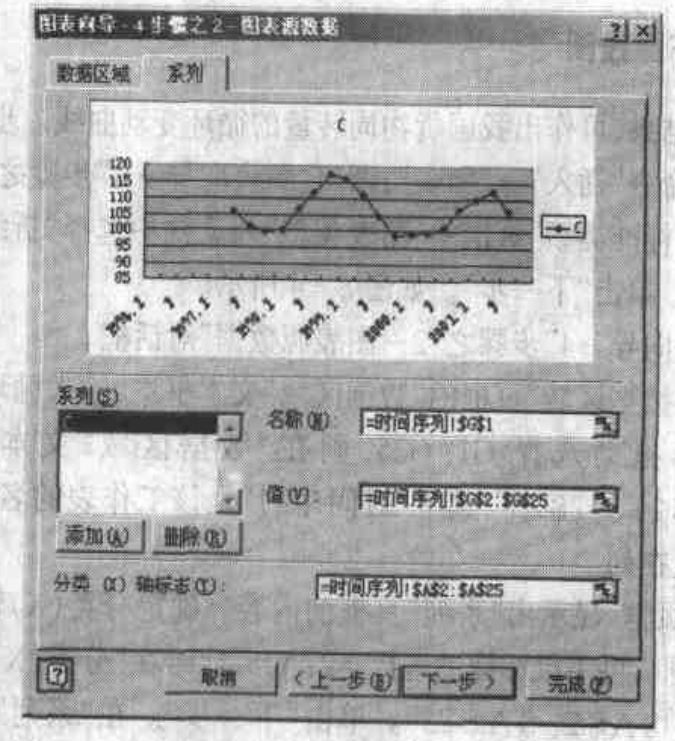


图 8－5

4.进入到“图表向导－4”步骤之“4－图表位置”对话框。可在该对话框中设置图表

存放的位置。一般不作修改,单击"完成"即可。

若在向导的第3和第4步中均不作修改,则可在填写完图表向导第2步中的对话框后,直接按"完成"按钮。

生成的图表区格式,还可以用"图表"工具栏进行修改。本例将"分类轴"的"分类轴刻度线之间的分类数"设为4,也即分类轴只显示时间列各年第一季度(如1996.1,2000.1)。图表区如图8-5所示。

关于Excel中图表应用的具体细节,请参见Excel帮助或其他专业书籍。

从循环曲线图可以看出,我国的货运量从1997年第四季度的低谷到1998年第四季度的高峰,再到1999年第四季度的低谷,历经两年,形成一个完整的循环过程。从1999年第四季度开始,进入下一个循环,其高峰已在2001年第二季度出现。

本章小结

1.总量指标时间序列是基本的时间序列,有时期数列和时点数列两种。相对指标和平均指标时间序列是由总量指标时间序列派生的。

2.水平指标包括发展水平与增长水平、平均发展水平与平均增长水平。总量指标序时平均数的计算是最基本的。相对指标或平均指标时间序列的序时平均数由分子、分母的序时平均数对比而得。增长水平是两个不同时期的发展水平之差,各逐期增长量之和等于累积增长量。平均增长水平是逐期增长量的序时平均数。

3.速度指标包括发展速度和增长速度,平均发展速度和平均增长速度。定基发展速度与环比发展速度的数量关系是:环比发展速度的连乘积等于定基发展速度,相邻两期定基发展速度之商等于相应时期的环比发展速度。平均发展速度有水平法和累计法两种计算方法。增长速度等于发展速度减1,平均增长速度等于平均发展速度减1。

4.长期趋势测定的方法有时距扩大法、移动平均法和趋势模型法。当各期的一级增长量大体为常数时,可配合直线趋势方程来测定现象变动的趋势。方程中的参数通常用最小二乘法或加权最小二乘法求得。加权最小二乘法对近期数据赋予较大的权数,对远期数据赋予较小的权数,采用指数权数 W^{n-i}的形式,$0<W<1$。

5.趋势外推预测是对现象在未来的变动趋势进行测算和推断。①移动平均法用于趋势外推是按一定的规则连续地对若干前期的数据计算平均数来估计未来的预测值。②指数平滑法是统计预测中广泛使用的一种方法。一次指数平滑法的公式是:$y_{t+1}=\alpha y_t+(1-\alpha)\hat{y}_t$,式中$0<\alpha<1$。

6.季节变动的测定。测定季节变动的主要方法是计算季节指数。季节指数越大,

说明“季节越旺”;季节指数越小,则说明“季节越淡”;季节指数等于1说明没有季节变动。计算季节指数的方法有同期平均法和趋势剔除法。可利用季节指数进行预测。

7. 循环变动是指变动周期大于一年的有一定规律性的重复变动。直接法是将每年各季或各月的数值与上年同期进行对比。剩余法的基本思路是:利用分解分析的原理,在时间序列中剔除长期趋势和季节变动,然后再消除不规则变动,从而揭示循环变动的特征。

8. 在Excel中,可综合利用公式、函数及移动平均、指数平滑和回归分析等分析工具来进行时间序列分析,还可利用图表工具来直观表示数据的变动规律。

思考与练习题

一、选择题:

1. 年末银行存款余额是:(甲)时期数列;(乙)时点数列。已知1996~2000年某银行的年末存款余额,要计算各年平均存款余额,要运用的平均数是:(丙)几何序时平均数;(丁)“首末折半法”序时平均数。(　　)

a. 甲、丙　　b. 乙、丙　　c. 甲、乙　　d. 乙、丁

2. 某地区粮食增长量1990~1995年为12万吨,1996~2000年也为12万吨。那么,1990~2000年期间,该地区粮食环比增长速度(　　)。

a. 逐年上升　　b. 逐年下降　　c. 保持不变　　d. 不能作结论

3. 某商业集团2000~2001年各季度销售资料如表8-21所示。

表8-21

	2000年				2001年			
	1	2	3	4	1	2	3	4
1. 零售额(百万)	40	42	38	44	48	50	40	60
2. 季初库存额(百万)	20	21	22	24	25	26	23	28
3. 流通费用额(百万)	3.8	3.2	2.8	3.2	3.0	3.1	3.1	4.0
4. 商品流转次数(次/季)	1.95	19.5	1.65	1.8	1.88	2.04	1.63	2.03

表8-21的资料中,是时期数列的有:(　　)

a. 1,2,3　　b. 1,3,4　　c. 2,4　　d. 1,3

4. 利用上题资料计算零售额移动平均数(简单,4项移动平均),2001年第二季度移动平均数为:(　　)

a. 47.5　　b. 46.5　　c. 49.5　　d. 48.4

5.利用第(3)题资料计算2000年商品季平均流转次数(=零售额/库存额)为()。

a. 1.885　　b. 1.838　　c. 1.832　　d. 1.829

二、判断题:

1.连续12个月逐期增长量之和等于年距增长量。(　　)

2.计算固定资产投资额的年平均发展速度应用几何平均法。(　　)

3.用移动平均法分析企业季度销售额时间序列的长期趋势时,一般应取4项来进行移动平均。(　　)

4.计算平均发展速度的水平法只适合时点指标时间序列。(　　)

5.某公司连续四个季度销售收入增长率分别为9%,12%,20%和18%,其环比增长速度为0.14%。(　　)

三、问答题:

1.某企业2000年8月几次员工数变动登记如表8-22所示。

表8-22

8月1日	8月11日	8月16日	8月31日
1 210	1 240	1 300	1 270

要求:

计算该企业8月份平均员工数。

2.某地区"九五"时期国内生产总值资料如表8-23所示。

表8-23　　单位:百万元

年份	国内生产总值	其中:		
		第一产业	第二产业	第三产业
1996	21 618	5 289	9 102	7 227
1997	26 624	5 800	11 670	9 154
1998	34 515	6 882	16 428	11 205
1999	45 006	9 438	21 259	14 309
2000	57 733	11 365	28 274	18 094

要求:

计算该地区"九五"时期国内生产总值和各产业的平均发展水平。

3.某地区"九五"期间年末居民存款余额如表8-24所示。

表 8－24

年份	1995	1996	1997	1998	1999	2000
存款余额	7 034	9 110	11 545	14 746	21 519	29 662

要求：

计算该地区“九五”期间居民年平均存款余额。

4. 某企业 2000 年产品库存量资料如表 8－25 所示。

表 8－25　　单位：件

日期	库存量	日期	库存量	日期	库存量
1 月 1 日	63	4 月 30 日	50	9 月 30 日	60
1 月 31 日	60	5 月 31 日	55	10 月 31 日	68
2 月 28 日	88	6 月 30 日	70	11 月 30 日	54
3 月 31 日	46	7 月 31 日	48	12 月 31 日	58
		8 月 31 日	49		

要求：

计算第一季度、第二季度、上半年、下半年和全年的平均库存量。

5. 某企业 1995～2000 年年底工人数和管理人员数资料如表 8－26 所示。

表 8－26　　单位：人

年份	工人数	管理人员数	年份	工人数	管理人员数
1995	1 000	40	1998	1 230	52
1996	1 202	43	1999	1 285	60
1997	1 120	50	2000	1 415	64

要求：

计算 1996～2000 年该企业管理人员数占工人数的平均比重。

6. 某地区 1995～2000 年社会消费品零售总额资料如表 8－27 所示。

表 8－27　　单位：亿元

年度	1995	1996	1997	1998	1999	2000
社会消费品零售总额	8 255	9 383	10 985	12 238	16 059	19 710

要求：

计算全期平均增长量、平均发展速度和平均增长速度，并列表计算：

(1)逐期增长量和累积增长量；

(2)定基发展速度和环比发展速度；

(3)定基增长速度和环比增长速度;

(4)增长1%的绝对值。

7.某地区2001年末人口数为2 000万人,假定以后每年以9‰的速度增长,又知该地区2001年GDP为1 240亿元,到2005年人均GDP将达到9 500元。

要求:

该地区2005年的GDP应达到多少?2002年到2004年GDP的年均增长速度应达到多少?

8.某企业1990~2000年产品产量资料如表8-28所示。

表8-28

单位:件

年份	产量	年份	产量	年份	产量
1987	344	1992	468	1997	580
1988	416	1993	486	1998	569
1989	435	1994	496	1999	548
1990	440	1995	522	2000	580
1991	450	1996	580	2001	629

要求:

(1)进行三项中心化移动平均修匀;

(2)根据修匀后的数据用最小二乘法配合直线趋势方程,并据以计算各年的趋势值;

(3)预测2002年该企业的产品产量。

9.某地区1991年到2001年人口自然增长数如表8-29所示。

表8-29

单位:万人

年份	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
增长人口	869	885	899	913	926	936	948	960	971	983	993

要求:

判断上表数列是否属于直线型,若为直线型,则运用最小二乘法拟合直线方程,并根据直线方程求各年人口增长趋势值。

10.某市集市1998~2001年各月猪肉销售量如表8-30所示。

表 8－30

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
1998	40	50	41	39	45	53	68	73	50	48	43	38
1999	43	52	45	41	48	65	79	86	64	60	45	41
2000	40	64	58	56	67	74	84	95	76	68	56	52
2001	55	72	62	60	70	86	98	108	87	78	63	58

要求：

分别用同期平均法和移动平均剔除法计算季节比例。

11. 某银行 1996～2000 年，各年年末存款余额如表 8－31 所示。

表 8－31

	年末存款余额(亿元)
1996	1.5
1997	2.0
1998	3.0
1999	3.3
2000	4.0

要求：

用普通最小二乘法和加权最小二乘法估计参数，取 $W=0.6$。并计算各年理论值，比较两种估算方法的误差。

12. 某地区 1991 年到 2000 年的 GDP 如表 8－32 所示。

表 8－32　　单位：亿元

年份	GDP	年份	GDP
1991	216	1996	679
1992	266	1997	748
1993	345	1998	816
1994	450	1999	895
1995	577	2000	1 036

要求：

选择最适合的 α 值，并用一次指数平滑模型预测 1992～2001 年的 GDP。

第九章

统计决策

★ 本章学习要点与要求 ★

本章对统计决策的基本理论、方法及其应用作扼要的介绍。通过学习，要求：①理解有关统计决策的基本概念与基本步骤，能够运用收益矩阵表与决策树形图表述所要研究的决策问题；②了解各种决策准则的特点与适用的场合，能够运用这些准则，进行完全不确定性决策与一般风险型决策；③了解贝叶斯决策的基本思想，掌握后验概率、完全信息价值与补充信息价值的计算方法，并在此基础上进行决策分析。

第一节　统计决策的基本概念

一、什么是统计决策

所谓决策，就是在占有一定信息的基础上，利用各种方法，对影响特定目标的各种因素进行计算和分析，从而选择关于未来行动的“最佳方案”或“满意方案”的过程。

不少决策问题，需要利用有关的统计信息和相应的统计分析方法。从广义上讲，所有利用统计方法和统计信息的决策都可以称为统计决策。但由于篇幅的限制，本章并不准备全面讨论这种广义的统计决策，而只是扼要介绍狭义的统计决策。

狭义统计决策的基本特点是：

第一,它研究的是非对抗型的决策问题。决策问题可分为两大类:对抗型决策问题和非对抗型决策问题。对抗型决策是由多个不同的决策主体在相互竞争和对抗中进行决策。进行对抗型决策时,必须考虑对方可能采取的策略。我国历史上著名的"田忌赛马"就是一个典型的对抗型决策的案例。对抗型决策问题属于运筹学中的博弈论(对策论)所研讨的内容。非对抗型决策只有一个决策主体,进行决策时,只要考虑可能出现的不同状态,而不必考虑对方可能采取的策略。狭义的统计决策是研究非对抗型决策问题的一种方法。

第二,它研究的是非确定型的决策问题。按照对客观条件的不同把握程度,决策可分为确定型决策和不确定型决策。在有关条件可以完全确定的情况下进行决策,称为确定型决策。求解复杂的确定型决策问题,通常运用运筹学中的数学规划方法。在有关条件不能确定的情况下进行决策,称为不确定型决策。求解不确定型的决策问题,需要应用大量概率统计的方法。这是本章讨论的重点。

按照对各种客观条件发生概率的了解程度不同,不确定型决策可分为完全不确定型决策和风险型决策。在对各种条件发生的概率一无所知的情况下进行决策,称为完全不确定型决策。在对各种条件发生的概率(可能性大小)有所了解的情况下进行决策,称为风险型决策。

第三,它是一种定量决策。统计决策是以统计分析和统计预测为基础的定量决策。利用统计决策可以把握决策问题的具体数量,便于比较、鉴别和选择。

综上所述,狭义的统计决策方法是一种研究非对抗型和非确定型决策问题的科学的定量分析方法。开展统计决策研究,有助于避免决策的盲目性,提高决策的科学性。

二、统计决策的基本步骤

一个完整的统计决策过程,包括以下几个基本步骤:

(一)确定决策目标

所谓决策目标是在一定条件制约下,决策者希望达到的结果。它是分析和研究决策问题的出发点和归宿。决策目标应根据所研究问题的具体特点确定。合理的统计决策目标应当尽可能简单明确,并且要以具有可测量性的指标来体现其内容与涵义。反映决策目标的变量,称为目标变量。

按照决策目标的多少,决策问题可分为单目标决策和多目标决策。当决策所要求达到的目标只有一个时,称为单目标决策。例如,在个人的证券投资决策中,一般将投资收益的最大化作为惟一的目标,这就是典型的单目标决策问题。当决策所要求达到的目标不止一个时,称为多目标决策。例如,对某产业项目进行投资决策,不仅要求其能够获得尽可能大的净收益,而且希望将环境污染控制在尽可能小的程度上,这就属于

多目标决策问题。求解多目标决策问题比较复杂，一般需要先利用一定的方法，将多个决策目标加权综合成一个总目标或构造一个新的综合目标函数，然后再利用单目标决策的方法求解。限于篇幅，本章只着重讨论单目标决策的问题。

（二）拟定备选方案

目标确定之后，需要分析实现目标的各种可能途径。这就是所谓的拟定备选方案。一般来说，备选方案应在两个以上。如果某一问题只有一种可能的解决方案，就不存在所谓的决策问题。备选方案是决策者可以调控的因素，备选方案中所调控的变量称为行动变量。所有备选方案的集合称为行动空间。为了拟定合适的备选方案，必须广泛收集有关的信息，认真听取各方面人员的意见，充分发挥主动性和创造性。

（三）列出自然状态

所谓自然状态简称状态，是指实施行动方案时，可能面临的客观条件和外部环境。某种状态是否出现，事先一般是无法确定的。对于同一个决策问题，各种状态不会同时出现，也就是说，它们之间是互相排斥的。例如，投资建设某种产品的生产线，该产品未来的市场销售情况可分为好、中、差三种。在这里，好、中、差就是关于市场销售的3种状态，三者都有出现的可能，但又不可能同时出现。虽然，各种状态是决策过程中客观存在的，其是否出现并不以决策者的主观意志为转移，但是，为了提高决策的科学性，人们总是尽可能地设法估计各种状态有可能出现的概率。所有可能出现的状态的集合称为状态空间，而相应的各种状态可能出现的概率的集合称为状态空间的概率分布。

（四）测算结果

为了从各种备选方案中挑选出合适的方案，需要测算不同方案在各种状态下可能实现的目标变量值，即不同方案在各种状态下的结果，所有的结果构成结果空间。

（五）选择"最佳"或"满意"的方案

在对各种方案可能产生的结果进行比较分析的基础上，决策者可按照一定的标准（准则），选择"最佳"或"满意"的方案。

（六）实施方案

方案确定之后，必须组织人力、物力和财力将其付诸实施。由于以上的决策是根据对未来的预计作出的，因此，所选择的方案是否真正合适，还需要通过实践的检验。同时，应将实施过程中的信息及时反馈给决策者。如果实施结果出乎意料，或者自然状态发生重大变化，应暂停实施，并及时修正方案，重新决策。

三、收益矩阵表

收益矩阵表是求解统计决策问题的重要工具。其基本形式如表9－1所示。

收益矩阵表由以下几部分组成：

(1)行动空间

$$A = (a_1, a_2, \cdots, a_m)$$

(2)状态空间

$$\Theta = (\theta_1, \theta_2, \cdots, \theta_n)$$

(3)状态空间的概率分布(P_j 是θ_j 出现的概率)。

$$P = (P_1, P_2, \cdots, P_n)(P_j \geqslant 0, \sum_{j=1}^{n} P_j = 1^{①})$$

(4)收益矩阵

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mn} \end{pmatrix}$$

收益矩阵的元素 q_{ij}反映在状态θ_j 下，采用行动方案 a_i 得到的收益值(结果)。这里所说的收益是广义的，凡是能作为决策目标的指标都可以称为收益。如利润、产量、销售收入等属于正的收益指标，成本、亏损额等属于负的收益指标。收益是行动方案和自然状态的函数，可用下式表示：

$$q_{ij} = Q(a_i, \theta_j) \quad (i = 1,2,\cdots,m;\ j = 1,2,\cdots,n) \tag{9.1}$$

以表格形式表示如表9-1所示。

表9-1 收益矩阵表

状态		θ_1	θ_2	$\cdots$	θ_n
概率		P_1	P_2	$\cdots$	P_n
方案	a_1	q_{11}	q_{12}	$\cdots$	q_{1n}
	a_2	q_{21}	q_{22}	$\cdots$	q_{2n}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_m	q_{m1}	q_{m2}	$\cdots$	q_{mn}

收益矩阵表的作用在于可以将各种备选方案在不同状态下的结果以及产生有关结果的可能性一目了然地列示出来，从而便于决策者从中选择最满意的方案。

【例9-1】一家酿酒厂就是否推出一种新型啤酒的问题进行决策分析。拟采取的

① 有的场合，决策问题的状态表现为连续型变量，这时其概率分布为连续型的概率分布。限于篇幅，本章不讨论连续型变量的决策问题。以下论述，均以离散型概率分布为前提。

方案有三种：一是进行较大规模的投资，年生产能力为 2 500 万瓶，其每年的固定成本费用为 300 万元；二是进行较小规模的投资，年生产能力 1 000 万瓶，其每年的固定成本费用为 100 万元；三不推出该种啤酒。假定在未考虑固定费用的前提下，每售出一瓶酒，均可获纯利 0.3 元。据预测，这种啤酒可能的年销售量为：50 万瓶、1 000 万瓶和 2 500 万瓶，这三种状况发生的概率分别为：0.2，0.3，0.5。

试编制该问题的收益矩阵表。

解：

首先，分别计算不同状态下，采用不同方案可能带来的收益。

例如，当需求量大（年销售 2 500 万瓶）时，

方案一的收益为：$0.3\times2\ 500-300=450$ 万元

方案二的收益为：$0.3\times1\ 000-100=200$ 万元

方案三的收益为：0

其他状态的收益计算方法相同，过程不一一列出。

在以上计算的基础上，可编制如下收益矩阵表，如表 9－2 所示。

表 9－2 啤酒投资的收益矩阵表 单位：万元

状态		需求大	需求中	需求小
概率		0.5	0.3	0.2
方案	方案一	450	0	－285
	方案二	200	200	－85
	方案三	0	0	0

第二节 完全不确定型决策

一、完全不确定型决策的准则

如前所述，完全不确定型决策是在对各种状态发生的概率一无所知的情况下进行决策。因此，需要先列出各种状态下可能的结果，然后根据一定的标准即所谓的决策准则去选择其满意的方案。完全不确定型决策常用的准则有以下几种：

（一）最大的最大收益值准则

该准则又称乐观准则或“好中求好”准则。其特点是决策者对未来形势比较乐观。在决策时，先选出各种状态下每个方案的最大收益值，然后再从中选择最大者，并以其相对应的方案作为所要选择的方案。该准则的数学表达式为：

$$a^* = \max_i \max_j \{q_{ij}\} \tag{9.2}$$

式中，a^*是所要选择的方案。

（二）最大的最小收益值准则

该准则又称悲观准则或“坏中求好”准则。它正好与乐观准则相反，决策者对未来形势比较悲观。在决策时，先选出各种状态下每个方案的最小收益值，然后再从中选择最大者，并以其相对应的方案作为所要选择的方案。该准则的数学表达式为：

$$a^* = \max_i \min_j \{q_{ij}\} \tag{9.3}$$

【例9－2】假设例9－1中，有关市场状态的概率完全不知道，试根据最大的最大收益值准则和最大的最小收益值准则进行决策。

解：

(1)例9－1中，方案一在各种状态下的最大收益为450万元，方案二在各种状态下的最大收益为200万元，方案三在各种状态下的最大收益为0，根据最大的最大收益值准则，应选择方案一。

(2)例9－1中，方案一在各种状态下的最小收益为－285万元，方案二在各种状态下的最小收益为－85万元，方案三在各种状态下的最小收益为0，根据最大的最小收益值准则，应选择方案三。

（三）最小的最大后悔值准则

后悔值又称机会损失值，即由于决策失误而造成的其实际收益值与最大可能的收益值的差距。方案a_i在状态θ_j下的后悔值，可按下式计算：

$$r_{ij} = \max_i Q(a_i, \theta_j) - q_{ij} \tag{9.4}$$

式中，$\max_i Q(a_i, \theta_j)$是在第j种状态下，正确决策有可能得到的最大收益，q_{ij}是收益矩阵的元素。显而易见，$r_{ij} \geqslant 0$。

在完全不确定型的决策中，虽然各种状态的概率无法估计，但是一旦作出决策并付诸实施，则必然会遇到某种状态。如果实际选择的方案正好是这种状态下的最优方案（有可能带来最大收益的方案），则后悔值为0；如果实际选择的方案不如最优方案，决策者就会感到后悔。后悔值越大表明所选的方案与最优方案差距越大。

最小的最大后悔值准则主张：应在求出后悔矩阵的基础上，先选出各种状态下每个方案的最大后悔值，然后再从中选择最小者，并以其相对应的方案作为所要选择的方

案。该准则的数学表达式为：

$$a^* = \min_i \max_j \{r_{ij}\} \tag{9.5}$$

【例9-3】假设例9-1中，有关市场状态的概率完全不知道，试求出后悔矩阵并根据最小的最大后悔值准则进行决策。

解：

(1) 在市场需求大的情况下，采用方案一可获得最大收益，故有：

$$\max_i Q(a_i, \theta_1) = 450$$

在市场需求中的情况下，采用方案二可获得最大收益，故有：

$$\max_i Q(a_i, \theta_2) = 200$$

在市场需求小的情况下，采用方案三可获得最大收益，故有：

$$\max_i Q(a_i, \theta_3) = 0$$

将其代入(9.4)式，可求得以下后悔矩阵(参见表9-3)。

(2)由表9-3可知：方案一的最大后悔值为285万元，方案二的最大后悔值为250万元，方案三的最大后悔值为450万元。根据最小的最大后悔值准则，应选择方案二。

表9-3 某项投资的后悔矩阵表 单位：万元

状态		需求大	需求中	需求小
方案	方案一	0	200	285
	方案二	250	0	85
	方案三	450	200	0

(四)折中准则

该准则认为，对未来的形势既不应该盲目乐观，也不应过分悲观。主张根据经验和判断确定一个乐观系数 $\delta(0 \leqslant \delta \leqslant 1)$，以 δ 和 $1-\delta$ 分别作为最大收益值和最小收益值的权数，计算各方案的期望收益值 $E(Q(a_i))$

$$E(Q(a_i)) = \delta \max_i \{q_{ij}\} + (1-\delta) \min_i \{q_{ij}\} \tag{9.6}$$

以期望收益值最大的方案作为所要选择的方案。该准则的数学表达式为：

$$a^* = \max_i E(Q(a_i)) \tag{9.7}$$

【例9-4】假设例9-1中，有关市场状态的概率不知，根据经验判断的乐观系数为0.6，试根据折中准则进行决策。

解：

将有关数据代入(9.6)式，可得：

$$E(Q(a_1)) = 0.6\times450 + (1-0.6)(-285) = 156$$
$$E(Q(a_2)) = 0.6\times200 + (1-0.6)(-85) = 86$$
$$E(Q(a_3)) = 0.6\times0 + (1-0.6)\times0 = 0$$

因为在可选择的方案中，方案一的期望收益值较大，所以根据折中原则，应选择方案一。

（五）等可能性准则

该准则认为：既然我们不知道未来各种状态出现的可能性有多大，那么不妨假定其发生的概率相等。在此基础上求各方案收益的期望值，并以期望收益值最大的方案作为所要选择的方案。该准则的数学表达式为：

$$a^* = max\,E(Q(a_i)) \tag{9.8}$$

$$E(Q(a_i)) = \frac{1}{n}\sum_{j=1}^{n} q_{i,j} \qquad (i=1,2,\cdots,m) \tag{9.9}$$

【例9－5】假设例9－1中，有关市场状态的概率不知，试根据等可能性准则进行决策。

解：

将有关数据代入(9.9)式，可得：

$$E(Q(a_1)) = 1/3(450 + 0 - 285) = 55$$
$$E(Q(a_2)) = 1/3(200 + 200 - 85) = 105$$
$$E(Q(a_3)) = 1/3(0 + 0 + 0) = 0$$

因为，按(9.9)式计算的方案二的期望收益值最大，所以按等可能性准则，应选择方案二。

二、各种准则的特点和适用场合

应当指出，决策方案的确定与所选用的决策准则有很大关系。表9－4列出了前面的例子中根据不同准则选取的方案。

表9－4　各种准则决策结果的比较

所依据的准则	选择的方案
最大的最大收益值准则	方案一
最大的最小收益值准则	方案三
最小的最大后悔值准则	方案二
折中准则($\alpha=0.6$)	方案一
等可能性准则	方案二

由表9－4可以看出，尽管前面的例子中所利用的基础资料都是表9－2给出的收

益矩阵，但是由于依据的决策准则不一样，选出的方案也是不同的。显而易见，这些决策结果不可能都是正确的。那么究竟应如何选择决策准则，以保证决策结果的正确性呢？由于完全不确定型决策问题相当复杂，而决策者掌握的信息又非常有限，因此，在实际决策时，决策准则的选择往往取决于决策者的偏好，也就是说对准则的选择仍带有相当程度的主观随意性。为了提高决策的科学性，减少盲目性，在选用准则时，应注意分析各种准则隐含的假定和决策时的各种客观条件。客观条件越接近于某一准则的隐含假定，则选用该准则进行的决策结果就越正确。

最大的最大收益值准则事实上是假定未来最理想状态（有可能出现最大收益值的状态）发生的可能性很大，并在这一前提下来选择收益最大的方案。这是一种比较冒进的决策准则。一般只有在客观情况确实很乐观，或者即使决策失误，也完全可以承受损失的场合才采用。这是因为，按照这一准则决策，固然有可能带来最大的收益，但是如果一旦出现不理想的状态，则可能造成较大的损失。例如，在前面的例子中，如果实际出现市场需求小的状态，则按此准则决策带来的损失最大。

最大的最小收益值准则事实上是假定未来最不理想状态（有可能出现最小收益值的状态）发生的可能性很大，并在这一前提下来选择收益最大的方案。这是一种比较保守的决策准则。采用这一准则，往往选择无所作为的方案，这样虽然可以避免出现大的损失，但是也可能损失获利的机会。例如，在前面的例子中，如果实际出现市场需求大的状态，则按此准则所选择的方案三是最差的方案。因此，这一准则适用于对未来的状态非常没有把握，或者难以承受决策失误损失的场合。

最小的最大后悔值准则是将能够获利而未获利也看成是一种机会损失，并假定发生较大的机会损失值的状态出现的可能性也较大，在这一前提下选择机会损失值最小的方案。采用该准则进行决策，可以抓住获利的机会，避免出现大的后悔。但是，如果一旦出现不理想的状态，则按此准则选择的方案也可能比其他方案带来较大损失。例如，在前面的例子中，如果实际出现市场需求小的状态，则按此准则选择的方案二将比方案三带来较大的损失。因此，这一准则适用于不愿放过较大的获利机会，同时又对可能出现的损失有一定承受力的场合。

折中准则和等可能性准则都是以各种方案的收益的期望值作为选择方案的标准。这一点与下一节将要介绍的风险决策的期望值准则是类似的。所不同的是，这两个准则只是对各种状态发生的概率作出假定，而这些假定并没有充分的理由和依据。折中准则事实上是假定未来可能发生的状态只有两种：即最理想状态和最不理想状态。前者发生的概率是 a，后者发生的概率是 $(1-a)$。当 $a=1$ 时，该准则等价于乐观准则，而当 $a=0$ 时，该准则等价于悲观准则。实际应用该准则时，应根据风险的大小、对未来状态的预计以及对决策失误的承受力，调整 a 的赋值。等可能性准则事实上是假定

各种状态出现的概率相等。该准则只适用于对未来各种状态发生的可能性完全心中无数的场合。

第三节　一般风险型决策

一、自然状态概率分布的估计

风险型决策与完全不确定型决策的不同之处在于：它是在估计出状态空间的概率分布的基础上进行决策。一般风险型决策中，所利用的概率包括客观概率与主观概率。客观概率是一般意义上的概率，可来源于频率估计，通常是由自然状态的历史资料推算或按照随机实验的结果计算出来的。例如，购买体育彩票的中奖概率就属于客观概率。主观概率是基于自身的学识、经验作出的对某一事件发生的可能性的主观判断。在很多场合，人们缺乏有关自然状态的历史资料，同时又不可能通过大量独立的随机实验去取得资料，难以用频率或一定的理论分布来估计客观概率。因此，有必要引进主观概率。虽然主观概率有相当大的主观成分，但它不是纯粹的猜测，必须具备相关的知识和经验才能给出主观概率的合理估计。

二、风险型决策的准则

利用状态的概率分布，可以对决策问题作比较细致的计算和分析，在此基础上再根据一定的准则进行选择和判断。风险型决策常用的决策准则有以下几种：

（一）期望值准则

该准则是一般风险型决策中应用最广泛的一个准则。它是以各方案收益的期望值的大小为依据，来选择合适的方案。具体决策时，可按下式计算收益的期望值：

$$E(Q(a_i))=\sum_{j=1}^{n} q_{i,j}P_j \qquad (i=1,2,\cdots,m) \tag{9.10}$$

式中，$E(Q(a_i))$是 i 方案的收益的期望值；$q_{i,j}$是 i 方案在出现 j 状态时的收益值；p_j是 j 状态出现的概率。

收益的期望值越大，表明平均来说，该方案获得的收益也越大。因此，可将各方案中收益期望值较大的方案作为最佳方案。

（二）变异系数准则

应当指出，单纯以期望值作为判断标准往往是不够充分的。这是因为收益的期望

值所反映的只是一种平均趋势，在进行决策时还应考虑其离散程度。如果某一方案虽然期望值较大，但是其方差也非常大，这就表明该方案具有较大的风险性。所以，当出现两个方案收益的期望值相差不大的情况时，应进一步观察各方案的方差，选择其中方差较小的方案。为此，可以在期望值达到一定数额的前提下，以变异系数作为选择方案的标准，以变异系数较低的方案作为所要选择的方案。

方差 $Var(a_i)$ 和变异系数 V 的计算公式如下：

$$Var(a_i) = E(Q(a_i)) = \sum_{j=1}^{n}[q_{i,j} - E(Q(a_i))]^2 p_j \qquad (i = 1,2,\cdots,m) \tag{9.11}$$

$$V_i = \frac{\sqrt{Var(a_i)}}{E(Q(a_i))} \qquad (i = 1,2,\cdots,m) \tag{9.12}$$

【例 9－6】试利用例 9－1 中给出的收益矩阵表的资料，根据期望值准则和变异系数准则选择最佳的投资方案。

解：

(1)将有关数据代入(9.10)式，可得：

$$E(Q(a_1)) = 450\times0.5 + 0\times0.3 - 285\times0.2 = 168$$

$$E(Q(a_2)) = 200\times0.5 + 200\times0.3 - 85\times0.2 = 143$$

$$E(Q(a_3)) = 0\times0.5 + 0\times0.3 + 0\times0.2 = 0$$

(2) $E(Q(a_3)) = 0$，可以从备选方案中排除。方案一和方案二的期望值虽有差别，但差别不是很大，所以再计算变异系数，帮助判断。

将有关数据代入(9.11)式和(9.12)式，可得：

$$Var(a_1) = (400-168)^2\times0.5+(0-168)^2\times0.3+(-285-168)^2\times0.2=76\,421$$

$$Var(a_2) = (200-143)^2\times0.5+(200-143)^2\times0.3+(-85-143)^2\times0.2=12\,996$$

$$V_1 = \frac{\sqrt{76\,421}}{168} = 1.645\,5$$

$$V_2 = \frac{\sqrt{12\,996}}{143} = 0.797\,2$$

所以，如果单纯根据收益期望值大小为标准，应选择方案一；如果将收益的期望值和方差结合在一起考虑，选择方案二比较合适。

(三)最大可能准则

该准则主张以最可能状态作为选择方案时考虑的前提条件。所谓最可能状态，是指在状态空间中具有最大概率的那一状态。按照最大可能准则，在最可能状态下，可实现最大收益值的方案为最佳方案。

最大可能准则是将风险条件下的决策问题，简化为确定条件下的决策问题。它不考虑其他状态下各行动方案收益值的差异对决策结果的影响。因此，只有当最可能状

态的发生概率明显大于其他状态时，应用该准则才能取得较好的效果。

【例 9－7】试利用例 9－1 中给出的收益矩阵表的资料，根据最大可能准则选择最佳的投资方案。

解：

该例的各种自然状态中，“市场需求大”的概率最大，因此，该状态为最可能状态。在市场需求大的状态下，方案一可以获得最大的收益。所以，根据最大可能准则，应选择方案一。

应当指出，在本例的场合，应用最大可能准则，不是很恰当。因为，“市场需求大”这一状态的概率虽然大于其他状态，但差距不会很大，其概率只有 0.5。也就是说，其他状态发生的可能性还有一半。所以，应用该准则决策失误的可能性有 50%。

（四）满意准则

利用这一准则进行决策，首先要给出一个满意水平。所谓满意水平，是指决策者认为比较合理、可以接受的目标值。然后，将各种方案在不同状态下的收益值与目标值相比较，并以收益值不低于目标值的累积概率为最大的方案作为所要选择的方案 a^*。该准则的数学表达式如下：

$$a^* = Max\, P\{Q(a_i,\theta_j) \geqslant A\} \quad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \tag{9.13}$$

式中，A 是给定的满意水平，$Q(a_i,\theta_j)$ 是 i 方案在 j 状态下的收益，$P\{Q(a_i,\theta_j) \geqslant A\}$ 是各方案收益值不低于目标值状态的累积概率。

应当指出：利用该准则的决策结果，与满意水平的高低有很大关系。满意水平一旦改变，所选择的方案也将随之改变。

【例 9－8】试利用例 9－1 中给出的收益矩阵表的资料，根据满意准则选择满意的投资方案，假定给出的满意水平有 200 万元和 400 万元两种。

解：

(1) $P\{Q(a_1,\theta_j) \geqslant 200\} = 0.5$

$P\{Q(a_2,\theta_j) \geqslant 200\} = 0.5+0.3 = 0.8$

$P\{Q(a_3,\theta_j) \geqslant 200\} = 0$

在备选方案中，方案二达到满意水平的累积概率最大，所以选择方案二。

(2) $P\{Q(a_1,\theta_j) \geqslant 400\} = 0.5$

$P\{Q(a_2,\theta_j) \geqslant 400\} = 0$

$P\{Q(a_3,\theta_j) \geqslant 400\} = 0$

在备选方案中，方案一达到满意水平的累积概率最大，所以选择方案一。

三、利用决策树进行风险型决策

决策树是求解风险型决策问题的重要工具，它是一种将决策问题模型化的树形图。

决策树由决策点、方案枝、机会点、概率枝和结果点组成。决策点以"□"代表,表示在该点必须对各种方案作出选择(进行决策);从决策点可引出若干条直线,表示有若干方案可以选择,故称为方案枝;与方案枝另一头相连接的"○"称为机会点;从机会点又可引出若干条直线,每一条直线表示一种自然状态,当各种自然状态发生的可能性(概率)已知时,在各条直线上应标明该状态发生的概率,故又称其为概率枝。在概率枝的末端标有不同方案在各种状态下的收益值(条件收益值),称为结果点。为了使决策树的层次更为清晰,一般对决策点和机会点还要进行编号,号数填在"□"和"○"中(决策树的具体结构可参见图9-1)。

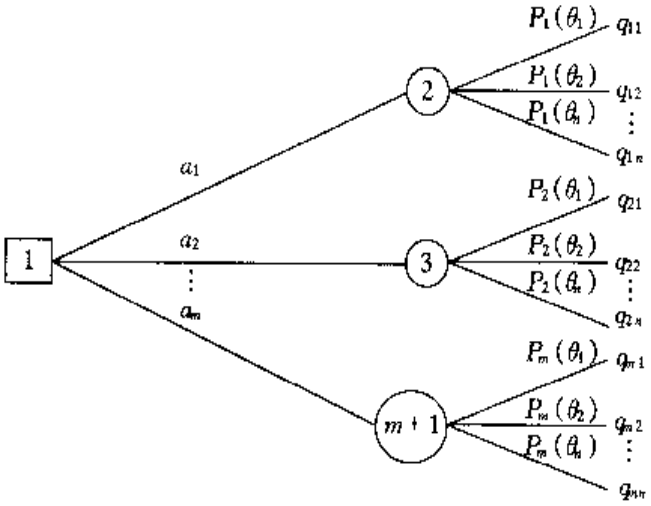


图9-1 决策树结构图

利用决策树对方案进行比较和选择,一般采用逆向分析法,即从树形结构的末端的条件结果开始,从后向前逐步分析。决策树分析通常采用期望值准则①。首先,根据条件收益值和相应状态的概率计算各方案的期望收益值,将其标在机会点"○"的上方。其次,对各方案进行比较,从中删除较差的方案,在删除的方案枝上画上"//"。这就是所谓的"剪枝"。最后,在决策树上留下的方案枝就是所要选择的最佳或满意方案。所选择的方案的期望收益值可标在相应的决策点的上方。

在一些比较复杂的决策问题中,某一方案的结果(可能得到的期望收益值)有赖于下一阶段乃至更多阶段的决策。这种场合,各种不同层次的行动空间、状态空间及其概率分布很容易混淆。利用决策树图,可以用简单直观的形式将其很好地表现出来。因此,它特别适用于求解这一类复杂的多阶段决策问题。

【例9-9】某汽车配件厂拟安排明年某零部件的生产。该厂有两种方案可供选择:

① 决策树也可采用其他准则决策,这时可按其他准则的要求进行计算、分析和比较。

方案一是继续利用现有的设备生产，零部件的单位成本是0.6万元；方案二是对现有设备进行更新改造，以提高设备的效率。更新改造需要投资100万元(假定其全部摊入明年的成本)，成功的概率是0.7。如果成功，零部件不含上述投资费用的单位成本可降至0.5万元；如果不成功，则仍用现有设备生产。另据预测，明年该厂某零部件的市场销售价格为1万元，其市场需求有两种可能：一是2 000件，二是3 000件，其概率分别为0.45和0.55。试问：①该厂应采用何种方案？②应选择何种批量组织生产？

解：

在本例中，首先要解决的问题是对生产方案的选择，但是对生产方案进行选择需要考察各种方法可能的结果。而这些结果又依赖于对生产批量的选择。因此，这是一个典型的两阶段决策问题。求解该问题，可按以下步骤进行：

1. 根据题中给出的条件，画出决策树结构图(参见图9-2)。

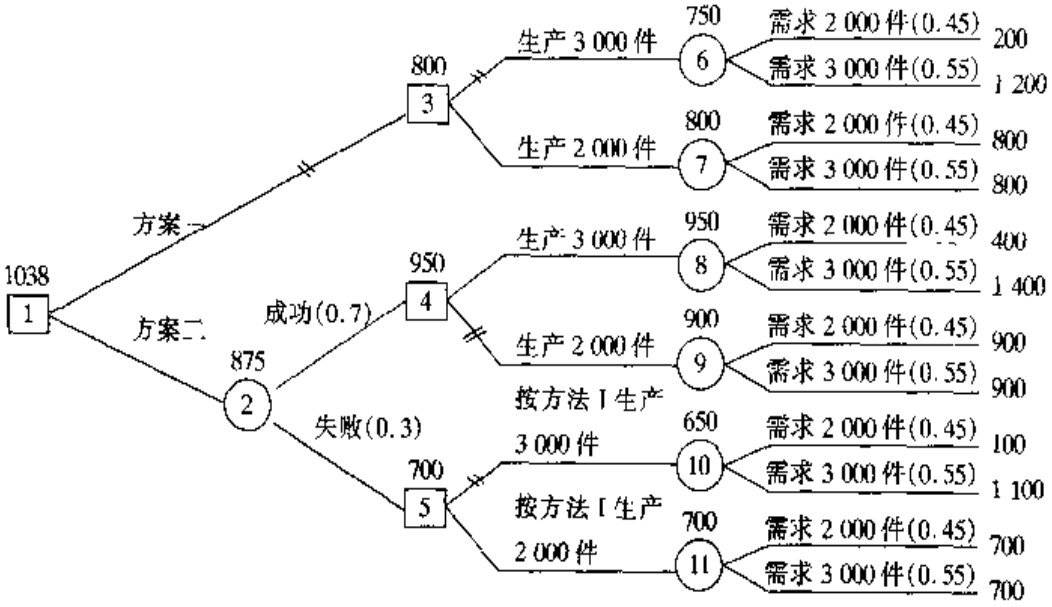


图9-2 例9-9的决策树图

在决策点[1]要对生产方案进行选择，其可供选择的方案有方案一和方案二，故可由此引出两条方案枝。如果该厂决定采用方案一，即采用原有的生产方法组织生产，则需要再对生产批量进行选择，所以，方案一分枝另一头的节点[3]是下一阶段的决策点。如果该厂决定采用方案二，即对设备进行更新改造，则可能有成功和失败两种状态，如果成功可采用新生产方法生产，如果失败则仍要采用原生产方法生产。所以，方案二分枝另一头的节点②是机会点，由该点可引出两条概率枝。无论方案二成功与否，都有一个生产批量选择的问题。因此，概率枝另一头的节点[4]和[5]也是下一阶段的决策点。从[3]、[4]和[5]又分别可引出两条方案枝，即生产3 000件和生产2 000件。于是，又

得到6个机会点。这6个机会点都面临两种可能的市场需求:2 000件或3 000件,所以可引出12条概率分枝。

2.计算决策树最末端的条件收益值。这里采用的计算式如下:

净收益=可能销售量×单价-生产量×单位成本-应摊新投资费用

当生产批量大于市场需求量时,可能销售量等于市场需求量。而当生产批量小于市场需求量时,可能销售量等于生产批量。另外,当选择按方案一组织生产时,应摊新投资费用等于0,选择按方案二组织生产时,应摊新投资费用100万元。例如:

右边第一个结果点的条件收益=2 000-3 000×0.6-0=200

3.利用各条件收益值和相应的概率分布,计算最右端各机会点的期望收益值。例如:机会点⑥的期望值=200×0.45+1 200×0.55=750

4.根据期望值准则,选出决策点[3]、[4]、[5]的最佳生产批量,并将最佳方案的期望收益值填在相应的决策点的上方。同时,剪除落选的方案枝。例如,在决策点[3]选择生产2 000件的方案,该方案的期望收益值为800万元。

5.利用[4]、[5]的结果,计算机会点②的期望收益值。将其与方案一的期望收益值比较,按照期望值准则选择最佳方案。

从图中可以看出,方案二的期望收益值为875万元,大于方案二的期望收益值(800万元)。

本例决策树分析的结论是:该汽车配件厂应按方案二对设备进行更新改造,如果能够成功,就采用新生产方法组织生产,其批量安排为3 000;如果失败,则仍采用原生产方法组织生产,其批量安排为2 000。

第四节 贝叶斯决策

一、什么是贝叶斯决策

在以上所述的一般风险性决策问题中,自然状态的概率是作为已知条件给出的。但是,在现实经济生活中,事先给出的各种状态的概率(又称为先验概率)常常是不准确的。因此,需要通过进一步的试验和调查,收集补充信息,并利用补充信息对原来估计的概率进行修订,从而求得更接近实际的新概率(利用补充信息修正的概率又称为后验概率)。所谓贝叶斯决策,就是利用补充信息,根据概率计算中的贝叶斯公式来估计后验概率,并在此基础上对备选方案进行评价和选择的一种决策方法。

利用贝叶斯决策方法，可以将先验的信息和补充的信息结合在一起进行分析与判断，从而提高了决策的可靠性。同时，利用该方法，还可以对信息的价值以及是否需要采集新的补充信息作出科学的判断。

二、贝叶斯公式与后验概率的估计

设某种状态 θ_j 的先验概率为 $P(\theta_j)$，通过调查获得的补充信息为 e_k，θ_j 给定时，e_k 的条件概率（似然度）为 $P(e_k/\theta_j)$，则在给定信息 e_k 的条件下，θ_j 的条件概率，即后验概率 $P(\theta_j/e_k)$ 可用以下贝叶斯公式计算：

$$P(\theta_j/e_k)=\frac{P(\theta_j)\cdot P(e_k/\theta_j)}{\sum_{j=1}^{n}P(\theta_j)\cdot P(e_k/\theta_j)} \tag{9.14}$$

【例 9－10】某空调机生产厂家拟向另一电子元件厂购买某种电子元器件，根据过去的经验，该电子元件厂产品发生不同次品率的概率分布如表 9－5 第 2 栏所示。但据说，该厂的产品质量最近有所提高。现从市场上该电子元件厂出售的该种元器件中，随机抽取了 10 件，结果未发现次品。试根据这一信息，对以往元器件厂次品率的概率分布进行修正。

解：

以往的概率分布可视为先验概率。在各种不同次品率给定条件下，抽查 10 件发生 0 件次品（发生 0 件为 e_0）的概率近似地服从于二项分布，其似然度可按以下方式计算：

$$P(e_0/\theta_j)=C_{10}^{0}\theta_j^{0}(1-\theta_j)^{10} \qquad (j=1,2,3,4) \tag{9.15}$$

例如，$P(e_0/\theta_1)C_{10}^{0}(0.05)^0(0.95)^{10}=0.599$

在 Excel 中，利用 BINOMDIST 函数可以方便地计算二项分布的概率。

表 9－5 的第 3 栏，给出了按照上式计算的结果。将似然度代入贝叶斯公式(9.14)式，可求得不同状态下的后验概率，结果如表 9－5 中最后一栏（第 5 栏）所示。例如，次品率为 0.05 状态的后验概率为：

$$P(\theta_1/e_0)=0.1\times0.599/(0.059\,9+0.139\,6+0.078\,8+0.010\,7)=0.207$$

从表 9－5 中的结果可以看出：出于实际抽查的次品率为 0，因此，次品率为 0.05 这种状态的后验概率大于先验概率，而次品率为 0.15 和 0.20 这两种状态的后验概率小于先验概率。也就是说，从抽查的结果看，该电子原件厂的产品质量比过去有所提高。

表 9-5 后验概率的计算

次品率 θ_j	先验概率 $P(\theta_j)$	似然度 $P(e_0/\theta_j)$	$P(\theta_j)\cdot P(e_0/\theta_j)$	后验概率 $P(\theta_j/e_0)$
0.05	0.1	0.599	0.059 9	0.207
0.10	0.4	0.349	0.139 6	0.483
0.15	0.4	0.197	0.078 8	0.273
0.20	0.1	0.107	0.010 7	0.037
	1.0		0.289	1.000

三、先验分析与后验分析

自然状态的概率分布有先验概率与后验概率之分。与之相对应,决策分析也可分为先验分析和后验分析。先验分析是利用先验概率进行决策,而后验分析则是利用后验概率作为选择与判断合适方案的依据。在不少场合,两种分析所作的结论是不一致的。由于后验分析中不仅利用了先验信息,而且还利用了补充信息,因此,一般来说,只要补充信息是准确的,则后验分析的结论更为可靠。

【例 9-11】设在例 9-10 中,对于是否向电子原件厂购买电子元器件,空调机厂有两种可供选择的方案,即:方案一购买;方案二不购买。假设其收益矩阵表如表 9-6 所示。试根据期望值准则,进行先验分析和后验分析。

表 9-6 收益矩阵表

状态:次品率		0.05	0.10	0.15	0.20
先验概率		0.1	0.4	0.4	0.1
后验概率		0.207	0.483	0.273	0.037
方案	购买 a_1	200	50	-100	-300
	不买 a_2	0	0	0	0

解:

(1)先验分析

$$E(Q(a_1))=200\times0.1+50\times0.4-100\times0.4-300\times0.1=-30$$

$$E(Q(a_2))=0$$

根据先验概率和期望值准则,应选择方案二。

(2)后验分析

$$E(Q(a_1))=200\times0.207+50\times0.483-100\times0.273-300\times0.037=27.15$$

$$E(Q(a_2))=0$$

根据后验概率和期望值准则,应选择方案一。

四、完全信息价值与补充信息价值

完全信息价值与补充信息价值是开展贝叶斯决策分析常用的两个基本指标。

所谓完全信息,是指在对某一问题进行决策时,对于所有可能出现的自然状态都可以提供完全确切的情报。决策者掌握了完全信息,便可以根据一定的目标,选择最佳的行动方案,风险型的决策问题,也就转化为确定型的决策问题。因此,完全信息的价值,可以由掌握完全信息前后所采取的不同行动方案的收益值的差额来表示。因为,不同状态下收益值的差额有所不同,所以人们用收益值差额的期望值来综合反映完全信息的价值。其计算公式如下:

$$\begin{aligned} EVPI &= E[\max_i Q(a_i,\theta_j)-Q(a_*,\theta_j)] \\ &= \sum_{j=1}^{n}P(\theta_j)[\max_i Q(a_i,\theta_j)-Q(a_*,\theta_j)] \end{aligned} \tag{9.16}$$

上式中,$EVPI$(Expected Value of Perfect Information)是完全信息价值的期望值,$\max_i Q(a_i,\theta_j)$表示各方案在状态 θ_j 下的最大收益值,$Q(a_*,\theta_j)$表示先验分析中的最佳方案在状态 θ_j下的收益值。

$EVPI$ 实质上反映了由于未掌握完全信息而造成的平均机会损失①。$EVPI$ 越大表明通过收集补充信息使决策效益提高的余地越大。同时,它也代表了为取得该项情报可付出的代价的上限。

【例 9-12】试利用例 9-10 和例 9-11 的资料和结果,计算完全信息价值的期望值。

解:

在该例中,如果能够确切知道电子元件厂产品的次品率为 0.05 或 0.10,则空调机厂应采取方案一,即购买元件厂的产品,其最大收益分别为 200 万元和 50 万元,如果能够确切知道元件厂产品的次品率为 0.15 或 0.20,则空调机厂应采取方案二,这时可避免损失,其最大收益都是 0。而在无法确切知道元件厂产品中的次品率的情况下,空调机厂只能利用以往的先验概率分布,进行决策,根据期望值准则,选择方案二。因此,当真实的次品率在 0.10 以下(含 0.10)时,先验分析的结果将带来机会损失,而当次品率大于 0.10 时,先验分析的结果也就是最大可能收益(本例中可避免负的收益)。在以上

① 参见本章第二节"最小最大后悔准则"中,关于机会损失的论述。

所述的各种状态下可能得到的最大收益值和先验分析的最佳方案 a_2 在不同状态下的收益值如表 9－7 的第 3 行与第 4 行所示，各种状态下完全信息的价值则如该表的第 5 行所示。

表 9－7 完全信息价值计算

状态：次品率	0.05	0.10	0.15	0.20
先验概率	0.1	0.4	0.4	0.1
最大收益值 Max $Q(a_i,\theta_j)$	200	50	0	0
先验分析最佳方案 a_2 的收益值 $Q(a_*,\theta_j)$	0	0	0	0
完全信息价值 = Max $Q(a_i,\theta_j)$ − $Q(a_*,\theta_j)$	200	50	0	0

利用表 9－7 中的资料，可求得完全信息价值的期望值。

$$EVPI = 200\times0.1+50\times0.4+0\times0.4+0\times0.1=40\text{ 万元}$$

本例中，完全信息价值的期望值相当高。这表明通过追加补充信息，有可能进一步提高决策的效益。

在现实经济生活中，人们通常很难获得完全信息。通过各种手段，如抽样调查、咨询等得到的补充信息量一般不会超过完全信息量，其准确性也不如完全信息。因此，它们的价值通常低于完全信息价值。补充信息 e_k 的价值 VAI 的计算公式如下：

$$VAI(e_k) = \text{先验 } EVPI - \text{后验 } EVPI(e_k) \tag{9.17}$$

式中，先验 $EVPI$ 是根据状态的先验概率计算的完全信息价值的期望值，后验 $EVPI(e_k)$ 是在了解补充信息 e_k 后，利用根据该信息修正的后验概率计算的完全信息价值的期望值。先验 $EVPI$ 反映了在先验分析中由于未能掌握关于自然状态的完全信息而造成的平均机会损失额。在后验分析中，通过补充信息的利用，使平均机会损失有所减少，但仍有部分信息未掌握，决策者仍要为这部分不确定性付出代价，它体现在后验 $EVPI$ 中。所以，补充信息的价值可以通过先验 $EVPI$ 与后验 $EVPI$ 的差额来反映。

显而易见，$VAI(e_k)$ 的取值与 e_k 有关。为了综合反映补充信息的价值，还需要计算补充信息价值的期望值 $EVAI$(Expected Value of Additional Information)。

$$\begin{aligned} EVAI &= E[VAI(e_k)] \\ &= \sum_{k=0}^{T} VAI(e_k)\cdot P(e_k) \end{aligned} \tag{9.18}$$

式中，$P(e_k)$是e_k出现的概率。

$$P(e_k)=\sum_{j=1}^{n}P(\theta_j)\cdot P(e_k/\theta_j);\qquad \sum_{k=0}^{l}P(e_k)=1 \tag{9.19}$$

EVAI 是判断收集补充信息是否有利的基本标准。一般在收集补充信息之前，应将 *EVAI* 与收集补充信息的费用加以比较，只有当收集补充信息的费用小于 *EVAI* 时，平均来看，收集补充信息才有利可图。

【例 9－13】试利用例 9－10、例 9－11 和例 9－12 的资料和结果，计算补充信息 e_0 的价值、补充信息 e_0 出现的概率和补充信息的期望值。

解：

(1)例 9－10 中已经求得根据补充信息 e_0(抽查 10 件发现 0 件次品)修正的后验概率，按照与例 9－12 类似的方法，可求得不同状态下的后验完全信息价值(参见表 9－8)。

表 9－8　后验完全信息价值计算

状态：次品率	0.05	0.10	0.15	0.20
后验概率	0.207	0.483	0.273	0.037
最大收益值 $Max\ Q(a_i,\theta_j)$	200	50	0	0
后验分析最佳方案 a_2 的收益值 $Q(a_*,\theta_j)$	200	50	－100	－300
完全信息价值＝$Max\ Q(a_i,\theta_j)-Q(a_*,\theta_j)$	0	0	100	300

利用表 9－8 的资料，可得：

$$\text{后验 } EVPI(e_0)=0\times0.207+0\times0.483+100\times0.273+300\times0.037$$
$$=38.4\text{ 万元}$$

另外，由例 9－12 的结果已知先验 *EVPI* 为 40 万元。因此，利用(9.17)式，可得：

$$VAI(e_0)=40-38.4=1.6\text{ 万元}$$

(2)利用表 9－5 的资料和(9.19)式。可得：

$$P(e_0)=\sum_{j=1}^{n}P(\theta_j)\cdot P(e_0/\theta_j)=0.289$$

(3)按照与求解 $VAI(e_0)$相类似的方法，可以求得抽查 10 件出现(1,2,3,…,10)件次品等其他几种结果发生的概率 $P(e_k)$及其补充信息价值 $VAI(e_k)$(计算过程从略)，并列表如表 9－9 所示。

表 9－9 补充信息价值期望值计算

(1) 出现次品数 e_k	(2) $P(e_k)$	(3) 先验 EVPI (万元)	(4) 后验 EVPI (万元)	(5) (3)－(4) VAI	(6) (5)×(2)
0	0.289	40	38.4	1.6	0.457
1	0.352	40	39.9	0.1	0.041
2	0.226	40	23.8	16.2	3.653
3	0.096	40	14.1	25.9	2.485
4	0.029	40	8.2	31.8	0.934
5	0.007	40	4.7	35.3	0.235
6	0.001	40	2.5	37.5	0.041
7	0.000	40	1.3	38.7	0.005
8	0.000	40	0.6	39.4	0.00
9	0.000	40	0.3	39.7	0.00
10	0.000	40	0.2	39.8	0.00
合计	1.000	—	—	—	7.852

$$EVAI = \sum_{k=0}^{T} VAI(e_k) \cdot P(e_k) = 7.852 \text{ 万元}$$

五、后验预分析

在前面讨论的例 9－11 中，我们对补充信息的一个实际结果计算后验概率分布，并据此作出相应的决策。但是，在现实经济生活中，补充信息的取得需要付出一定的费用，采集补充信息也可以有多种方案。因此，在完整的贝叶斯决策过程中，在正式进行补充信息的调查之前，还需要根据先验概率以及各种可能发生的补充信息的结果估计后验概率，计算完全信息价值的期望值和补充信息价值的期望值，比较收集补充信息所需的费用和收益，对是否值得进一步收集补充信息的问题作出判断，并选择最佳的收集补充信息的方案。这一环节被称为后验预分析。

【例 9－14】试利用例 9－12 和例 9－13 的资料和结果，进行后验预分析。

解：

根据例 9－12 计算的结果，先验完全信息价值的期望值——先验 EVPI 达 40 万

元，表明有必要进一步调查，以提高决策的效益。

根据例 9－13 的结果，补充信息价值的期望值 $EVAI$ 为 7.852 万元。这表明如果进一步调查的费用低于 7.852 万元，则平均来看，进一步调查是有利的。

由以上分析可知，空调机厂在进行是否购买该电子元件之前，应进行适当的补充调查（补充调查费用控制在 7.8 万元以下），以便更准确地把握该电子元件的概率分布。并根据经过修正的概率分布来进行决策。具体来说，抽查 10 件产品，如果未发现次品，可采纳方案一，即购买该电子元件。如果发现 1 件以上次品则应采纳方案二（其后验分析过程省略），拒绝购买该电子元件。

本章小结

1. 狭义的统计决策方法是一种研究非对抗型和非确定型决策问题的科学的定量分析方法。

2. 收益矩阵表（见表 9－10）和决策树形图是统计决策分析的两种基本工具。

表 9－10　收益矩阵表

状态		θ_1	θ_2	…	θ_n
概率		P_1	P_2	…	P_n
方案	a_1	q_{11}	q_{12}	…	q_{1n}
	a_2	q_{21}	q_{22}	…	q_{2n}
	⋮	⋮	⋮	⋮	⋮
	a_m	q_{m1}	q_{m2}	…	q_{mn}

决策树是一种将决策问题模型化的树形图，它由决策点、方案枝、机会点、概率枝和结果点组成。

3. 完全不确定性决策是在对状态的概率分布一无所知的情况下进行决策分析。其决策准则有：①最大的最大收益值准则；②最大的最小收益值准则；③最小的最大后悔值准则；④折中准则；⑤等可能性准则。

4. 一般风险型决策是在估计出状态空间的概率分布的基础上进行决策。其主要决策准则有：①期望值准则；②变异系数准则；③最大可能准则；④满意准则。

5. 上述各种决策准则具有不同的特点，适用于不同的场合。

6. 贝叶斯决策是利用补充信息，根据贝叶斯公式来估计后验概率，并在此基础上进

行决策分析。

7.后验概率 $P(\theta_j/e_k)=\dfrac{P(\theta_j)\cdot P(e_k/\theta_j)}{\sum_{j=1}^{n}P(\theta_j)\cdot P(e_k/\theta_j)}$

8.先验分析是利用先验概率进行决策,后验分析则是利用后验概率作为选择与判断合适方案的依据。

9.完全信息的价值,可以由掌握完全信息前后所采取的不同行动方案的收益值的差额来表示。它也代表了为取得有关状态确切情报可付出的代价的上限。

$$EVPI = E[\max_i Q(a_i,\theta_j) - Q(a_*,\theta_j)]$$
$$=\sum_{j=1}^{n}P(\theta_j)[\max_i Q(a_i,\theta_j) - Q(a_*,\theta_j)]$$

10.补充信息 e_k 的价值 $VAI(e_k)=$ 先验 $EVPI-$ 后验 $EVPI(e_k)$,补充信息价值的期望值 $EVAI=E[VAI(e_k)]=\sum_{k=1}^{T}VAI(e_k)\cdot P(e_k)$。它是判断收集补充信息是否有利的基本标准。

思考与练习题

一、选择题:

1.收益矩阵表包括的主要内容有:(　　)

a. 状态空间　　b.行动方案

c.状态的概率分布　　d. 结果空间

2.以下决策准则属于风险型决策准则(　　)

a. 最大可能性准则　　b. 期望值准则

c.最小最大后悔准则　　d.折衷准则

3.贝叶斯决策与一般风险型决策的主要区别在于:

a. 计算完全信息价值　　b.计算后验概率

c.后验分析　　d.决策准则不同

二、问答题:

1.某企业拟开发生产一种新产品,有三个方案可供选择。其收益矩阵表如表9-11所示。

表 9－11

状态		需求大	需求中	需求小
方案	方案一	400	100	－140
	方案二	200	200	－20
	方案三	0	0	0

要求：

根据完全不确定型决策的五种决策准则，选择合适的方案(假定乐观系数为 0.6)。

2. 销售体育彩票的收入相当部分将用作发展体育事业，只有一部分作为奖金返还。试问：

(1)如根据期望值准则进行决策，是否会购买股票？请说明理由。

(2)如有人决定购买体育彩票，其依据的决策准则是何种准则？

3. 设某贸易公司近期有 3 笔生意可做，其收益矩阵表如表 9－12 所示。

表 9－12

状态		θ_1	θ_2	θ_3
概率		0.4	0.4	0.2
方案	方案一	300	150	－150
	方案二	200	200	－100
	方案三	100	100	80

要求：

(1)画出该决策问题的决策树。

(2)根据期望值准则和变异系数准则进行决策。

(3)如果该企业希望取得 200 万收益，试问该企业宜采用何种决策准则？应选择何种方案？

(4)如果该企业必须确保 80 万收益用于偿还到期的债务，试问该企业宜采用何种决策准则？应选择何种方案？

4. 某水利工程公司拟对大江截流的施工工期作出决策。可供选择的方案有两种：一是在 9 月份施工；二是在 10 月份施工。假定其他条件都具备，影响截流的惟一因素是天气与水文状况。10 月份的天气与水文状况肯定可以保证截流成功。而 9 月份的天气水文状况有两种可能。如果天气好，上游没有洪水，9 月底前截流成功，可使整个工程的工期提前，从而能比 10 月施工增加利润 1 000 万元；如果天气坏，上游出现洪

水,截流失败,则比10月施工增加500万元的损失。根据以往经验,9月份天气好的可能性是0.6,天气坏的可能性是0.4。为了帮助决策,公司拟向某气象站购买气象预报的资料。过去的资料表明,该气象站预报好天气的准确率是0.9,预报坏天气的准确率是0.7。

要求:

(1)列出收益矩阵表,根据期望值准则进行先验分析;

(2)计算先验完全信息的价值;

(3)估计后验概率;

(4)进行后验分析;

(5)计算后验完全信息价值;

(6)计算补充信息价值的期望值;

(7)为该水利工程公司选择合适的策略。

第十章

统计指数

★ 本章学习要点与要求 ★

本章学习统计指数的理论、方法和应用。通过本章学习，要求：①正确理解统计指数涵义和作用，掌握综合指数和平均指数的编制方法和编制特点；②了解和掌握综合指数与总平均指数的因素分析方法；③了解和掌握统计指数在社会经济问题中的应用。

第一节　统计指数概述

一、统计指数的概念

统计指数也称经济指数(简称指数)，是社会经济统计中常用的一种重要方法。它产生于18世纪，最早是用于测定物价的变动。此后的200多年，统计指数的理论与方法不断发展，其应用逐步扩大到产量、成本、进出口贸易、工资、生活费用、劳动生产率、股票证券等各个方面。

指数的涵义有广义和狭义两种。广义指数是指一切说明社会经济现象数量变动的相对数。狭义的指数是一种特殊的相对数，即用来说明不能直接相加的复杂社会经济现象综合变动程度的相对数。本章所要研究的统计指数主要是指狭义的指数。

统计指数具有以下几个基本性质和特点：

第一,统计指数通常以相对数的形式表示。

第二,反映复杂现象的统计指数具有综合的性质,它综合地反映了复杂现象总体的数量变化关系。复杂现象总体的数量变化常常受到许多因素的影响,例如,受多种因素的影响,各种商品价格变动的方向和幅度经常是不一致的,有些商品价格上涨,有些商品价格下跌,而且上涨与下跌的幅度也不一样。商品价格总指数是各种商品价格综合影响变动的结果。

第三,反映复杂现象的统计指数具有平均的性质,它反映复杂现象总体中各个单位变动的平均水平。就商品综合价格指数而言,它所表明的是各种商品价格变动的平均水平。

二、统计指数的种类

由于分析对象的不同,统计指数可以划分为不同的种类。

(一)按所考察范围的不同,分为个体指数、组指数和总指数

个体指数是反映单个现象或单个事物变动的相对数,如某种品牌型号的电冰箱的产量指数与价格指数都是个体指数。组指数也称类指数,是综合反映总体内某一类现象变动的相对数,如食品类、衣着类、服务类价格指数等,组指数通常也是编制总指数的中间环节。总指数是综合反映整个复杂经济现象总体变化情况的相对数,如商品价格总指数、工业产品产量总指数等。

(二)按所反映的现象特征不同,分为数量指标指数和质量指标指数

数量指标指数反映所研究现象的数量规模变动,如产品产量指数、商品销售量指数等。质量指标指数反映所研究现象的质量水平变动,如商品价格指数、产品成本指数、劳动生产率指数等。

(三)按所反映的时间状态不同,分为动态指数和静态指数

统计指数的最初的涵义是动态指数,由两个不同时间上的经济量对比形成,反映社会经济现象在不同时间上的发展变化。动态指数按所对比的基期的不同,分为定基指数与环比指数两种。静态指数是后来在实际应用中才提出的,包括空间指数和计划完成情况指数两种。空间指数指不同空间的同类现象水平在同一时间对比的结果,反映现象在不同区域的差异程度。计划完成情况指数则是将某种现象的实际水平与计划水平对比的结果,反映计划的完成程度。

三、统计指数的作用

统计指数的作用主要可以表现为以下几个方面:

第一,运用统计指数可以分析复杂经济现象总体的变动方向和程度。例如,股票价

格的变动是一种复杂现象,通过编制股票价格指数可以反映股票市场价格总变动的情况。再如,商品销售额、商品销售量、商品价格水平的变动也都是复杂的经济现象,通过编制统计指数可以反映其总体的变动方向和程度。

第二,运用统计指数可以分析复杂经济现象总体变动中各个因素的变动,以及它们的变动对总体变动的影响程度。复杂现象总体的变动是由各种因素综合影响的结果,而各种因素自身变动的幅度和变动方向常常是不一致的,对总体变动的影响也不同。例如,某地区2000年商品销售额对比1999年为110.9%,说明2000年该地区商品销售额的增长幅度为10.9%。这个变动是销售量与价格两个因素共同作用的结果,借助于统计指数法可以深入分析和测定这两个因素的变动及其对销售额变动所带来的影响。

第三,运用统计指数可以分析复杂现象平均水平的变动中各个因素的变动,以及它们的变动对总平均水平变动的影响程度。例如,城镇就业人口平均工资水平的变动,既受各行业职工工资水平变动的影响,也受各行业职工构成变动的影响。借助于统计指数方法,就能对全体就业人口的工资水平变动进行分析,同时分析各行业职工平均工资变动及其对全体就业人口平均工资变动的影响,分析各行业职工所占比重的变动及其对全体就业人口平均工资的影响。

第四,运用统计指数可以分析复杂经济现象总体的长期变化趋势。借助连续编制的动态指数形成的指数数列,可以反映现象在长时间的变化趋势。如果把两个相互联系的指数数列(如居民收入指数和价格指数)加以比较,还可以进一步认识复杂现象总体之间数量上的变动关系。

第五,运用统计指数可以对多指标复杂社会经济现象进行综合测评。如对综合国力、社会发展水平的综合评价研究。

第二节 综合指数

一、综合指数编制的基本原理

统计研究的对象主要是总体现象。因此,从研究对象的范围来看,编制指数主要是指总指数。综合指数就是总指数的基本形式之一。

编制综合指数首先必须明确两个概念:一是"指数化指标",二是"同度量因素"。所谓指数化指标就是编制综合指数所要测定的因素,如商品价格综合指数所要测定的因

素是价格，所以价格就是指数化指标。所谓同度量因素是指媒介因素，借助媒介因素，把不能直接加总的因素过渡到可以加总，所以称其为同度量因素。编制综合指数的目的是测定指数化指标的变动，因此，在对比的过程中对同度量因素应加以固定。综合指数的基本公式如下：

$$I_p=\frac{\sum p_1q_i}{\sum p_0q_i} \tag{10.1}$$

$$I_q=\frac{\sum q_1p_i}{\sum q_0p_i} \tag{10.2}$$

公式中，p 表示质量因素，q 表示数量因素，下标 1 表示所分析的时期称为报告期，0 表示比较时期称为基期，i 为同度量因素固定的时期。公式(10.1)是质量指标综合指数形式。价格、成本、劳动生产率指数等属于质量指标指数。公式(10.2)是数量指标综合指数形式。产量、劳动量指数、经济发展速度等属于数量指标指数。

编制综合指数的基本方法是"先综合，后对比"。例如，编制价格综合指数反映市场商品价格总变动，其步骤是：首先把市场各种商品价格乘以同度量因素加以综合，然后再进行对比。

编制综合指数必须解决以下两个问题。

首先，必须根据指数化指标的性质确定同度量因素的性质。一般而言，质量指标指数的指数化指标是 p，其同度量因素是数量指标 q，两者的乘积 pq 是一价值总量；数量指标指数的指数化指标则是 q，其同度量因素是 p，两者的乘积 qp 也是一价值总量，价值量指标加总后，其经济意义具体明确。

其次，必须确定同度量因素所固定的时期。根据不同的研究目的和所研究的现象的特点，同度量因素可以固定在基期，也可以固定在报告期。但分子分母中的同度量因素，必须固定在同一时期。只有这样，才能把握所要测定的指标的变动。

需要指出的是，同度量因素具有权数的性质。例如，在价格综合指数中，同度量因素 q 不仅可以使各种不同商品的销售价格转化为可以相加的价值量，而且由于各种商品的销售量不同，其各自价格变动对综合价格指数影响的大小也有所差别。由于基期与报告期各种商品销售结构可能不一样，因此，将 q 固定在不同的时期(基期或报告期)，所得到的综合指数的计算结果是不相同的。所以，在编制综合指数时，如何选择合适的权数，是一个需要着重研究的问题。

二、综合指数的类型

(一)拉氏指数

拉氏指数(简记为 L)，是德国经济统计学家拉斯佩雷斯(E. Laspeyres)在 1864 年

提出的。拉氏指数公式的特点是将同度量因素固定在基期水平上，因此也称基期综合指数，其公式如下：

$$L_q=\frac{\sum q_1p_0}{\sum q_0p_0} \tag{10.3}$$

$$L_p=\frac{\sum p_1q_0}{\sum p_0q_0} \tag{10.4}$$

【例 10－1】某企业三种产品报告期和基期销售资料如表 10－1 所示。

表 10－1

产品类别	计量单位	销售量		销售价格(元)		销售额(百元)			
		q_0	q_1	p_0	p_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
1	万件	450	500	700	770	3 150	3 850	3 500	3 465
2	万件	500	520	350	350	1 750	1 820	1 820	1 750
3	台	900	1 080	100	110	900	1 188	1 080	990
合计	—	—	—	—	—	5 800	6 858	6 400	6 205

解：

以表 10－1 的资料，计算拉氏价格指数和销售量指数：

$$L_q=\frac{\sum q_1p_0}{\sum q_0p_0}=\frac{6\ 400}{5\ 800}=110.34\%$$

$$L_p=\frac{\sum p_1q_0}{\sum p_0q_0}=\frac{6\ 205}{5\ 800}=106.98\%$$

（二）帕氏指数

帕氏指数(简记为 P)，是德国的另一位经济统计学家帕舍(H. Pasche)在 1874 年提出的。与拉氏指数的不同之处是，帕氏指数将同度量因素固定在报告期水平上，因此也称报告期综合指数。其公式如下：

$$P_q=\frac{\sum q_1p_1}{\sum q_0p_1} \tag{10.5}$$

$$P_p=\frac{\sum p_1q_1}{\sum p_0q_1} \tag{10.6}$$

【例 10－2】以表 10－1 的资料，计算帕氏价格指数和销售量指数：

解：

$$P_q=\frac{\sum q_1p_1}{\sum q_0p_1}=\frac{6\ 858}{6\ 205}=110.52\%$$

$$P_p=\frac{\sum p_1q_1}{\sum p_0q_1}=\frac{6\ 858}{6\ 400}=107.16\%$$

(三)马埃公式和理想公式

由上面的计算结果不难看出,拉氏指数和帕氏指数两者之间的差异,它们各自的分析意义也有所不同。在某种意义上讲,马埃公式和理想公式则是对拉氏指数和帕氏指数的偏差加以折中的综合指数形式。

1.马埃公式。马埃公式(简记为 E)由英国著名经济学家马歇尔(A. Marshall)和埃奇沃斯(F. Y. Edgeworth)等人于1887~1890年间提出。该指数对拉氏指数和帕氏指数的同度量因素进行简单平均。公式具体形式如下:

$$E_p=\frac{\sum p_1\left(\frac{q_0+q_1}{2}\right)}{\sum p_0\left(\frac{q_0+q_1}{2}\right)}=\frac{\sum p_1(q_0+q_1)}{\sum p_0(q_0+q_1)}=\frac{\sum p_1q_0+\sum p_1q_1}{\sum p_0q_0+\sum p_0q_1} \tag{10.7}$$

$$E_q=\frac{\sum q_1\left(\frac{p_1+p_0}{2}\right)}{\sum q_0\left(\frac{p_1+p_0}{2}\right)}=\frac{\sum q_1(p_1+p_0)}{\sum q_0(p_1+p_0)}=\frac{\sum q_1p_1+\sum q_1p_0}{\sum q_0p_1+\sum q_0p_0} \tag{10.8}$$

2.理想指数。理想公式(简记为 F)由美国经济学家沃尔什(G. M. Walsh)和庇古(P. C. Pigou)等人于1901~1902年先后提出,后来美国统计学家费雪(Irving Fisher)比较验证了其所具有的优良性后,将它命名为理想公式。也有人称其为费雪指数。理想指数是对拉氏指数和帕氏指数的几何平均。具体公式如下:

$$F_p=\sqrt{\frac{\sum p_1q_1}{\sum p_0q_1}\cdot\frac{\sum p_1q_0}{\sum p_0q_0}} \tag{10.9}$$

$$F_q=\sqrt{\frac{\sum q_1p_1}{\sum q_0p_1}\cdot\frac{\sum q_1p_0}{\sum q_0p_0}} \tag{10.10}$$

(四)固定权数综合指数

固定权数综合指数由英国经济学家杨格(A. Young)提出,因此也称杨格指数。在固定加权综合指数中,同度量因素所属时期既不固定在报告期也不固定在基期,而是固定在一个特定的水平上。公式具体形式如下:

$$I_p=\frac{\sum p_1q_n}{\sum p_0q_n} \tag{10.11}$$

$$I_q=\frac{\sum q_1p_n}{\sum q_0p_n} \tag{10.12}$$

式中,q_n 和 p_n 分别表示特定的物量和价格水平。

由于固定权数综合指数的同度量因素不因比较时期(报告期或基期)的改变而改变,因此采用固定权数综合指数,它不但方便指数的编制,而且便于观察现象长期发展变化的趋势。

第三节　平均指数

一、平均指数编制的基本原理

与综合指数相同，平均指数也是总指数的基本形式之一，用来反映复杂现象的总变动。平均指数之所以被称为平均指数，是因为它利用了加权算术平均数 $\bar{x}=\sum xf/\sum f$ 和加权调和平均数 $\bar{x}=\sum m/\sum(m/x)$ 的计算形式。在下面的讨论中，我们可以看到，平均指数的编制原理与综合指数编制的基本原理是相互贯通的。

平均指数编制的基本方法是"先对比，后平均"。所谓"先对比"是指先通过对比计算个体指数：$k_q=q_1/q_0$ 或 $k_p=p_1/p_0$；所谓"后平均"则是指将个体指数赋予适当的权数 p_0q_0 或 p_1q_1，加以平均得到总指数。

二、平均指数的类型

（一）算术平均指数

算术平均指数一般用 p_0q_0 加权，基本公式如下：

$$I_p=\frac{\sum\frac{p_1}{p_0}p_0q_0}{\sum p_0q_0} \tag{10.13}$$

$$I_q=\frac{\sum\frac{q_1}{q_0}p_0q_0}{\sum p_0q_0} \tag{10.14}$$

【例 10－3】以表 10－1 的资料，分别计算价格指数和销售量的算术平均指数。

解：

$$I_p=\frac{\sum\frac{p_1}{p_0}p_0q_0}{\sum p_0q_0}=\frac{\frac{770}{700}\times 3\ 150+\frac{350}{350}\times 1\ 750+\frac{110}{100}\times 900}{3\ 150+1\ 750+900}=\frac{6\ 205}{5\ 800}=106.98\%$$

$$I_q=\frac{\sum\frac{q_1}{q_0}p_0q_0}{\sum p_0q_0}=\frac{\frac{500}{450}\times 3\ 150+\frac{520}{500}\times 1\ 750+\frac{1\ 080}{900}\times 900}{3\ 150+1\ 750+900}=\frac{6\ 400}{5\ 800}=110.34\%$$

这两式的计算结果和前面综合指数中的拉氏指数的计算结果完全相同。不难发现，这是因为当个体指数与总值权数之间存在严格的一一对应关系时，采用基期总值加权的平均指数，实际上是拉氏综合指数的变形；

$$I_p=\frac{\sum\frac{p_1}{p_0}p_0q_0}{\sum p_0q_0}=\frac{\sum p_1q_0}{\sum p_0q_0} \tag{10.15}$$

$$I_q=\frac{\sum\frac{q_1}{q_0}p_0q_0}{\sum p_0q_0}=\frac{\sum q_1p_0}{\sum q_0p_0} \tag{10.16}$$

需要指出的是,算术平均指数不仅仅是综合指数的变形,在许多场合下它还是一种相对独立的总指数编制方法,具有比综合指数更广泛的适用性。以价格指数为例,其计算公式可变形为:

$$I_p=\frac{\sum\frac{p_1}{p_0}p_0q_0}{\sum p_0q_0}=\sum\frac{p_1}{p_0}\times\frac{p_0q_0}{\sum p_0q_0}=\sum\frac{p_1}{p_0}w \tag{10.17}$$

上式表明算术平均指数不仅可以用绝对数权数加权,也可以用相对数权数 w 加权。而相对数权数可以根据全面资料确定也可以根据非全面资料确定。在一定情况下,还可以将相对数权数加以固定,大大便利指数编制工作。这正是综合指数所不具备的。

(二)调和平均指数

调和平均指数一般用 p_1q_1 加权,基本计算形式如下:

$$I_p=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{p_1}{p_0}}p_1q_1} \tag{10.18}$$

$$I_q=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{q_1}{q_0}}p_1q_1} \tag{10.19}$$

【例 10-4】以表 10-1 的资料,分别计算价格指数和销售量的调和平均指数。

解:

$$I_p=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{p_1}{p_0}}p_1q_1}=\frac{3\,850+1\,820+1\,188}{\frac{1}{\frac{770}{700}}\times3\,850+\frac{1}{\frac{350}{350}}\times1\,820+\frac{1}{\frac{110}{100}}\times1\,188}=\frac{6\,858}{6\,400}=107.16\%$$

$$I_q=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{q_1}{q_0}}p_1q_1}=\frac{3\,850+1\,820+1\,188}{\frac{1}{\frac{500}{450}}\times3\,850+\frac{1}{\frac{520}{500}}\times1\,820+\frac{1}{\frac{1\,089}{900}}\times1\,188}=\frac{6\,858}{6\,205}=110.52\%$$

这两式的计算结果和前面综合指数中帕氏指数的计算结果完全相同。不难发现,这也是因为当个体指数与总值权数之间存在严格的一一对应关系时,采用报告期总值

加权的平均指数实际上是帕氏综合指数的变形：

$$I_p=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{p_1}{p_0}}p_1q_1}=\frac{\sum p_1q_1}{\sum p_0q_1} \tag{10.20}$$

$$I_q=\frac{\sum p_1q_1}{\sum\frac{1}{\frac{q_1}{q_0}}p_1q_1}=\frac{\sum p_1q_1}{\sum p_1q_0} \tag{10.21}$$

在指数的实际编制中，与编制加权算术平均指数一样，公式中的权数可以采用相对数权数。同理，在许多场合，编制调和平均指数比编制综合指数更为切实可行。

第四节　指数体系与因素分析

如前所述，指数不仅可以反映社会经济现象数量的变动程度，而且还可以用于分析影响总量变动的各个因素的作用。因素分析是借助于指数体系来进行的。

一、指数体系的概念

这里讨论的指数体系是指指数之间存在的相互联系。一般地说，三个或三个以上在性质上相互联系、在数量上存在一定关系的指数便构成指数体系。利用指数体系可以分析社会经济现象各种因素的变动，以及它们对总体发生作用的影响程度。

社会经济现象存在的客观联系，可以通过相应的指标体系表现出来。例如：

总产值＝产品产量×价格

总成本＝产品产量×单位成本

销售额＝销售量×价格

从上面的三个关系式我们可以看到，现象的总体可以分解为一个数量因素和一个质量因素。而现象总体的变化就可以归结为数量因素和质量因素共同作用的结果。上述指标体系按指数形式表现时，乘积关系仍然成立。即：

总产值指数＝产品产量指数×价格指数

总成本指数＝产品产量指数×单位成本指数

销售额指数＝销售量指数×价格指数

这些指数关系可以归纳为：

现象总体变动指数＝数量指标指数×质量指标指数

利用指数体系可以从数量方面研究分析社会经济现象总体变动中各个因素变动的影响程度和绝对额,即进行因素分析;也可以利用指数之间的联系进行必要的推算。

利用指数体系进行因素分析,主要分析如下两方面的问题:

1.分析现象总体总量指标的变动受各种因素变动的影响程度。这是利用综合指数体系,从数量指标指数和质量指标指数的相互联系中,分析各个因素的变动影响关系。例如,编制多种产品的销售量指数和价格指数,分析销售量和价格的变动对销售总额变动的影响。

2.分析社会经济现象总体平均指标变动受各种因素变动的影响程度。这种分析,是通过平均指标指数体系来进行的。

二、连锁替代法

(一)连锁替代法的概念

利用指标体系进行因素分析,一般采用连锁替代法。所谓连锁替代法,就是在被分析指标的因素结合式中,根据各因素的性质和相互联系的数量关系,将各个因素的基期数字顺次以报告期的数字替代,有多少因素就替代多少次;每次替代后的结果与替代前的结果进行对比,从相对数和绝对数两方面分析各因素对现象总体的影响。

用连锁替代法进行因素分析时应注意如下几个问题:

1.各因素的排列顺序。运用连锁替代法进行因素分析,一般是按各因素排列的先后顺序展开的。因此,什么因素排在前面,什么因素排在后面,需要事先加以考虑。一般原则是先数量因素后质量因素,先内涵因素后外延因素。

2.注意相邻因素之间的经济含义。运用连锁替代法进行因素分析,各因素排列顺序还要考虑它们相乘后的经济含义。例如,在研究企业利润额变动时,影响利润额变动的各因素排列的顺序为销售量、销售价格、利润率。销售量乘以销售价格等于销售额,销售额乘以利润率等于利润额,都有明确的经济意义。假如销售价格和利润率排列的位置互换,销售量乘以利润率的经济含义就难以确定,不符合指标分解逻辑。

(二)连锁替代法的步骤

以销售额变动分析为例,用 p,q 分别表示销售价格、销售量,下标 0 和 1 分别表示基期和报告期,即以 p_0,q_0 分别表示基期的数值,p_1、q_1 表示报告期的数值。连锁替代的过程如下:

$$\left.\begin{array}{l}\text{分析的起点}\rightarrow q_0p_0\\ \text{第一次替代}\rightarrow q_1p_0\end{array}\right\}q\text{ 因素变动的影响}$$

$$\left.\begin{array}{l}\text{第一次替代}\rightarrow q_1p_0\\ \text{第二次替代}\rightarrow p_1q_1\end{array}\right\}p\text{ 因素变动的影响}$$

具体的计算公式和步骤：

1. 计算被分析指标的总变动：

$$变动的程度=\frac{\sum q_1p_1}{\sum q_0p_0} \tag{10.22}$$

$$变动的绝对额=\sum q_1p_1-\sum q_0p_0 \tag{10.23}$$

2. 计算各因素变动影响的程度和绝对额：

$$q\ 因素变动影响程度=\frac{\sum q_1p_0}{\sum q_0p_0} \tag{10.24}$$

$$q\ 因素影响绝对额=\sum q_1p_0-\sum q_0p_0 \tag{10.25}$$

$$p\ 因素变动影响程度=\frac{\sum q_1p_1}{\sum q_1p_0} \tag{10.26}$$

$$p\ 因素影响绝对额=\sum q_1p_1-\sum q_1p_0 \tag{10.27}$$

3. 影响因素的综合分析。总变动程度等于各因素变动影响程度的连乘积：

$$\frac{\sum q_1p_1}{\sum q_0p_0}=\frac{\sum q_1p_0}{\sum q_0p_0}\times\frac{\sum q_1p_1}{\sum q_1p_0} \tag{10.28}$$

总因素变动绝对额等于各因素变动影响绝对额的总和：

$$\sum q_1p_1-\sum q_0p_0=(\sum q_1p_0-\sum q_0p_0)+(\sum q_1p_1-\sum q_1p_0) \tag{10.29}$$

三、两因素分析

下面结合例子说明用连锁替代法进行两因素分析的计算过程和实际分析意义。

【例 10－5】以表 10－1 资料，进行销售额变动的因素分析。

1. 销售额变动分析：

$$销售额变动指数=\frac{\sum q_1p_1}{\sum q_0p_0}=\frac{6\ 858}{5\ 800}=118.24\%$$

$$销售额变动的绝对额=\sum q_1p_1-\sum q_0p_0=6\ 858-5\ 800=10.58(万元)$$

2. 销售量变动影响分析：

$$销售量变动影响程度=\frac{\sum q_1p_0}{\sum q_0p_0}=\frac{6\ 400}{5\ 800}=110.34\%$$

$$销售量变动影响绝对额=\sum q_1p_0-\sum q_0p_0=6\ 400-5\ 800=6.00(万元)$$

3. 销售价格变动影响分析：

$$销售价格变动影响程度=\frac{\sum q_1p_1}{\sum q_1p_0}=\frac{6\ 858}{6\ 400}=107.16\%$$

$$销售价格变动影响绝对额=\sum q_1p_1-\sum q_1p_0=6\ 858-6\ 400=4.58(万元)$$

4. 影响因素综合分析：

$$\frac{\sum q_1p_1}{\sum q_0p_0}=\frac{\sum q_1p_0}{\sum q_0p_0}\times\frac{\sum q_1p_1}{\sum q_1p_0}$$

$$118.24\%=110.34\%\times107.16\%$$

$$\sum q_1p_1-\sum q_0p_0=(\sum q_1p_0-\sum q_0p_0)+(\sum q_1p_1-\sum q_1p_0)$$

$$10.58=6.00+4.58$$

分析结果表明:从相对数方面看,该企业的销售额报告期比基期增长 18.24%,是由于销售量增长 10.34% 和销售价格上涨 7.16%,这两个因素共同作用的结果;从绝对数方面看,该企业销售额报告期比基期增加 10.58 万元,是由于销售量上升使销售额增加 6 万元,以及销售价格上涨使销售额增加 4.58 万元共同作用的结果。

四、多因素分析

指数体系因素分析法,可以推广到三个、四个甚至更多因素分析。假如要对企业的利润额进行因素分析,可将利润额按下述方式分解:

利润额=销售量(q)×销售价格(p)×利润率(r)

按照连锁替代法,可以得到多因素分析指数体系:

$$\frac{\sum q_1p_1r_1}{\sum q_0p_0r_0}=\frac{\sum q_1p_0r_0}{\sum q_0p_0r_0}\times\frac{\sum q_1p_1r_0}{\sum q_1p_0r_0}\times\frac{\sum q_1p_1r_1}{\sum q_1p_1r_0} \tag{10.30}$$

$$\sum q_1p_1r_1-\sum q_0p_0r_0=(\sum q_1p_0r_0-\sum q_0p_0r_0)+(\sum q_1p_1r_0-\sum q_1p_0r_0)+(\sum q_1p_1r_1-\sum q_1p_1r_0) \tag{10.31}$$

【例 10-6】某企业利润额资料如表 10-2 所示,试分析销售量、销售价格、利润率对利润总额的影响。

表 10-2

产品类别	计量单位	销售量		销售价格(元)		利润率(%)		利润额(百元)			
		q_0	q_1	p_0	p_1	r_0	r_1	$q_1p_1r_1$	$q_1p_1r_0$	$q_1p_0r_0$	$q_0p_0r_0$
1	万件	450	500	700	770	8	9	346.5	308.0	280.0	252.0
2	万件	500	520	350	350	7	11	200.2	127.4	127.4	122.5
3	台	900	1 080	100	110	12	10	118.8	142.6	129.6	108.0
合计	—	—	—	—	—	—	—	665.5	578.0	537.0	482.5

解:

(1)利润额总指数$=\frac{\sum q_1p_1r_1}{\sum q_0p_0r_0}=\frac{665.5}{482.5}=137.93\%$

利润额报告期比基期增加 37.93%,绝对额为:

$$\sum q_1p_1r_1-\sum q_0p_0r_0=665.5-482.5=183(百元)$$

(2)销售量指数$=\dfrac{\sum q_1p_0r_0}{\sum q_0p_0r_0}=\dfrac{537.0}{482.5}=111.30\%$

销售量报告期比基期增长了 11.30%，由此而增加的利润额为：

$$\sum q_1p_0r_0-\sum q_0p_0r_0=537.0-482.5=54.5(百元)$$

(3)销售价格指数$=\dfrac{\sum q_1p_1r_0}{\sum q_1p_0r_0}=\dfrac{578}{537}=107.64\%$

销售价格报告期比基期提高 7.64%，由此而增加的利润额为：

$$\sum q_1p_1r_0-\sum q_1p_0r_0=578-537=41(百元)$$

(4)利润额指数$=\dfrac{\sum q_1p_1r_1}{\sum q_1p_1r_0}=\dfrac{665.5}{578}=115.14\%$

利润率报告期比基期提高 115.14%，由此而增加的利润额为：

$$\sum q_1p_1r_1-\sum q_1p_1r_0=665.5-578=87.5(百元)$$

上列计算结果的综合分析：

$$137.93\%=111.30\%\times107.64\%\times115.14\%$$

$$183=54.5+41+87.5$$

分析结果说明，报告期由于销售量增加 11.3%，销售价格提高 7.64% 和利润率提高 15.14%，三方面因素综合作用的结果使利润额提高 37.93%；利润总额增加 1.83 万元，是由于销售量增加而增加利润 0.545 万元，由于销售价提高而增加利润 0.41 万元和由于利润率提高而增加利润 0.875 万元的共同结果。

五、总平均数指数与平均指标变动的因素分析

总平均数指数也称平均指标指数，它是对总体平均指标变动的测定，如平均工资指数、劳动生产率指数都属于平均指标指数。平均指标指数有两个特点：

首先，它是利用分组资料计算的指数。它所测定的总平均数是对组平均数的加权平均，其权数是各组的单位数。总平均数即加权算术平均数，其表现形式为：$\bar{x}=\sum x_if_i/\sum f_i$，式中被平均标志 x_i 为各组平均水平，权数 f_i 为各组单位数。它所综合的变量并非不可同度量，而是不同地区、不同单位或不同部门的同一指标，无需采用同度量因素。

其次，从作用看，总平均数指数除了测定总体平均指标的变动程度之外，还测定了总体内部各组平均数变动和各组权数(结构变动)对总平均指标变动的影响。

分析平均指标的变动，需要计算以下 3 种指数：

(1)可变构成指数。分析总平均数的变动。计算形式为：

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \tag{10.32}$$

(2)固定构成指数。分析总体内部各组平均水平变动对总平均数的影响。计算形式为：

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} \tag{10.33}$$

(3)结构变动影响指数。分析总体内部各组权数(结构变动)对总平均指标变动的影响。计算形式为：

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \tag{10.34}$$

可变构成指数、固定构成指数和结构变动指数都是反映复杂现象变动的指数，因此都是总指数。这三个指数存在内在的联系：

(1)可变构成指数＝结构变动指数×固定结构指数

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \left(\frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \right) \times \left(\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} \right) \tag{10.35}$$

(2)总平均数变动绝对额＝结构变动影响额＋各组平均数变动影响额

$$\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} = \left(\frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} \right) + \left(\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_1}{\sum f_1} \right) \tag{10.36}$$

下面结合例子说明总平均数指数的计算与应用。

【例 10－7】某企业职工有关工资资料见表 10－3。

表 10－3 某企业职工工资情况表

职工类别	月工资(元)		职工人数(人)		工资额(百元)		
	基期 x_0	报告期 x_1	基期 f_0	报告期 f_1	基期 x_0f_0	报告期 x_1f_1	按基期工资水平计算的报告期工资额 x_0f_1
车间工人	800	850	300	270	2 400	2 295	2 160
管理人员	1 000	1 050	50	55	500	577.5	550
合计	—	—	350	325	2 900	2 872.5	2 710

从表 10－3 中的数据可以看出，该企业两类职工的工资水平和人数在基期和报告期都不同，并且变动的幅度也不同，这必然导致该企业报告期和基期的总平均工资水平的变动。

解：

(1)计算总平均工资变动指数：

$$可变构成指数=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}=\frac{883.8}{828.6}=106.67\%$$

$$变动绝对额=\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}=883.8-828.6=55.2(元)$$

(2)计算各组工资变动影响指数：

$$固定构成指数=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_1}{\sum f_1}=\frac{883.8}{833.8}=106.00\%$$

$$各组变量水平变动绝对额=\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_1}{\sum f_1}=883.8-833.8=50(元)$$

(3)计算结构变动影响指数：

$$结构影响指数=\frac{\sum x_0f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}=\frac{833.8}{828.6}=100.63\%$$

$$结构影响变动绝对额=\frac{\sum x_0f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}=833.8-828.6=5.2(元)$$

总平均工资指数＝结构变动影响指数×各组平均工资变动影响指数

$$106.67\%=100.63\%\times106.00\%$$

总平均工资变动绝对额＝结构变动影响额＋各组工资变动影响额

$$\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}=\left(\frac{\sum x_0f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}\right)+\left(\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_1}{\sum f_1}\right)$$

$$883.8-828.6=(833.8-828.6)+(883.8-833.8)$$

$$55.2=5.2+50$$

分析结果表明：从相对数方面看，该企业总平均工资报告期比基期上涨了6.67%，这是因为各组职工工资水平变动使总平均工资上涨6.00%，职工结构变动影响使总平均工资上涨0.63%，二者共同作用的结果；从绝对数方面看，该企业总平均工资报告期比基期增加55.2元，这是因为各组职工工资水平变动使总平均工资增加50元，职工结构变动影响使总平均工资增加5.2元的共同结果。

根据上述计算结果，还可以进一步分析由于平均工资水平的变动对企业工资总额支出的影响。

总平均工资的变动使企业工资总额支出增加额：

$$\left(\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}\right)\sum f_1=(883.8-828.6)\times325=55.2\times325=17\ 940(元)$$

其中：

(1)由于职工结构变动使平均工资提高，企业工资总额支出增加额：

$$\left(\frac{\sum x_0f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}\right)\sum f_1=(833.8-828.6)\times325=5.2\times325=1\ 690(元)$$

(2)由于各组职工工资变动使平均工资提高，企业工资总额支出增加额：

$$\left(\frac{\sum x_1 f_1}{\sum f_1}-\frac{\sum x_0 f_1}{\sum f_1}\right)\sum f_1=(883.8-833.8)\times 325=50\times 325=16\ 250(元)$$

第五节 统计指数的应用

统计指数在社会经济统计中应用很广泛，下面介绍几种主要的应用。

一、生产指数

生产指数概括反映一国或一地区各种产品产量的综合变动，它是衡量经济增长水平的指标。生产指数的编制方法有多种，下面介绍常见的三种方法。

（一）不变价格法

不变价格又称固定价格。不变价格法生产指数，即采用某一时期或时点的产品平均出厂价格作为同度量因素的固定权数综合指数。这种方法是我国建国以来长期沿用的方法。公式具体形式如下：

$$I_q=\frac{\sum q_1 p_n}{\sum q_0 p_n} \tag{10.37}$$

公式中，p_n 代表不变价格。我国先后采用过 1952 年、1957 年、1970 年、1980 年和 1990 年的不变价格标准。

（二）价格指数紧缩法

随着经济体制的转变，不变价格的编定变得相当困难，我国目前正在进行生产指数的改革。从 2002 年 4 月开始试行用价格紧缩法计算工业发展速度。

价格指数紧缩法是根据指数体系原理，利用价格指数消除产品价值量动态指标中的价格因素的影响，反映工业发展速度。计算方法有单缩法和双缩法两种。

1. 单缩法。单缩法是在工业产品的出厂价格指数与工业原材料等的投入价格指数包含的内容相差不大的情况下，用工业产品的出厂价格指数对现价工业发展速度进行缩减。计算公式如下：

$$I_q=\frac{\sum q_1 p_1}{\sum q_0 p_0}\div\frac{\sum r_1 z_1}{\sum r_1 z_0} \tag{10.38}$$

式中，右边第一项是按现价计算的工业发展速度，第二项是工业产品出厂价格指数。

2. 双缩法。双缩法就是利用产品的出厂价格指数和原材料、燃料、动力等购进价格指数，同时消除总产品和中间消耗中的价格变动因素，计算出可比价格工业发展速度。计算公式如下：

$$I_q=\frac{\sum m_1 p_1}{\sum m_0 p_0}\div\frac{\sum q_1 z_1-\sum k_1 s_1}{\sum q_0 z_1-\sum k_0 s_1} \tag{10.39}$$

式中，右边第一项是按现价计算的工业增加值发展速度，第二项是双缩法计算的工业增加值价格指数。q，z 为总产品的数量和价格，k，s 为中间消耗的数量和价格(有关总产品、中间消耗和增加值的概念，请参阅第 12 章)。

(三)工业生产指数法

工业生产指数法为世界上大多数市场经济国家所采用。它首先根据代表产品报告期与基期的产量数据，分别计算出各产品的个体指数，然后用其各自的权数对个体指数进行加权平均，计算出类指数。例如，对小类所有代表产品的产量个体指数进行加权平均，即为小类指数；对中类各代表产品产量指数加权平均即为中类指数。类似地，可计算出大类及总指数，综合反映工业发展速度。编制工业生产指数，一般采用基期固定权数，计算公式如下：

$$I_q=\frac{\sum i_q p_0 q_0}{\sum p_0 q_0} \tag{10.40}$$

式中，i_q 为各种工业品的个体产量指数，$p_0 q_0$ 为相应产品的基期增加值。在实践中，为了简化指数的编制工作，常常以各种工业品的增加值比重作为权数，并且将这些比重权数相对固定起来，运用“固定加权算术平均指数”连续地编制各个时期的工业生产指数。计算公式如下：

$$I_q=\frac{\sum i_q w}{\sum w} \tag{10.41}$$

二、产品成本指数

产品成本指数概括反映一个部门或企业各种产品成本的综合变动，它是衡量综合成本水平的指标。产品成本指数有以下几种形式：

1. 帕氏形式的以基期成本作为比较基准的成本综合指数：

$$P_p=\frac{\sum p_1 q_1}{\sum p_0 q_1} \tag{10.42}$$

该指数表示由于报告期单位成本比基期成本 p_0 节约或提高，报告期总成本降低或提高的幅度，分子分母之差表示报告期总成本的降低或增加额。

2. 帕氏形式的以计划成本作为比较基准的成本综合指数：

$$P_p=\frac{\sum p_1 q_1}{\sum p_n q_1} \tag{10.43}$$

该指数表示由于报告期单位成本比计划成本 p_n 节约或提高，报告期总成本降低或提高的幅度，分子分母之差表示报告期总成本的降低或增加额。

3.拉氏形式的以计划成本为比较基准的成本综合指数：

$$L_p=\frac{\sum p_1q_n}{\sum p_nq_n} \tag{10.44}$$

该指数与公式(10.42)和(10.43)不同，同度量因素不是报告期产量而是计划产量q_n，所以该指数是按照计划规定的产量结构，表示报告期总成本降低或提高的幅度。当管理部门同时制定成本和产量计划时，用这个指数来检查计划执行情况，可以考察部门或企业是否通过改变产量构成来达到完成成本计划的目的。

三、空间价格指数

空间价格指数概括反映同一时间、不同国家或不同地区各种商品价格水平的差异，也称区域价格指数。不同地区之间价格水平的比较，是经济领域里最敏感的现象。所以空间价格指数也是进行国际对比或地区对比的重要指标。与动态指数不同，空间价格指数是一种静态指数，而它的编制和分析也有一些特殊的要求。

假定对A,B两个地区进行价格比较，如果以B地区为比较基准，采用拉氏形式编制，则指数形式为：

$$L_p^{\frac{A}{B}}=\frac{\sum p_Aq_B}{\sum p_Bq_B} \tag{10.45}$$

如果反过来以A地区为对比基准，同样采用拉氏形式编制，则指数形式为：

$$L_p^{\frac{B}{A}}=\frac{\sum p_Bq_A}{\sum p_Aq_A} \tag{10.46}$$

下面，我们通过一个例子来看看用公式(10.45)和(10.46)比较A,B两个地区的价格水平，会得到什么结果。

【例10－8】要比较A,B两地的水果价格，以两地李子和桃子为代表商品，收集到如表10－4的数据资料，试计算两地水果价格比较指数。

表10－4

商品	计量单位	销售量		销售价格(元/千克)	
		q_A	q_B	p_A	p_B
李子	万千克	20	5	1.4	2.0
桃子	万千克	10	15	2.0	1.2

解：

$$L_p^{\frac{A}{B}}=\frac{\sum p_Aq_B}{\sum p_Bq_B}=\frac{1.4\times50\ 000+2.0\times150\ 000}{2.0\times50\ 000+1.2\times150\ 000}=132.14\%$$

$$L_p^{\overset{B}{A}}=\frac{\sum p_B q_A}{\sum p_A q_A}=\frac{2.0\times 200\ 000+1.2\times 100\ 000}{1.4\times 200\ 000+2.0\times 100\ 000}=108.33\%$$

不难发现,上面两式的计算结果互相矛盾。

这表明计算空间价格指数用拉氏形式和帕氏形式计算都不大合适,因为它们在互换基准后的指数计算结果不能保持一致。空间价格指数编制和分析的特殊要求是,互换基准后指数的结论应保持一致。用公式表示,即:

$$I_p^{\overset{A}{B}}=\frac{1}{I_p^{\overset{B}{A}}} \tag{10.47}$$

或

$$I_p^{\overset{A}{B}}\times I_p^{\overset{B}{A}}=1 \tag{10.48}$$

要达到这个要求,编制空间价格指数应采用马埃公式或理想公式。因为:

$$E_P^{\overset{A}{B}}\times E_P^{\overset{B}{A}}=\frac{\sum p_A(q_A+q_B)}{\sum P_B(q_A+q_B)}\times\frac{\sum p_B(q_A+q_B)}{\sum P_A(q_A+q_B)}=1 \tag{10.49}$$

$$F_P^{\overset{A}{B}}\times F_P^{\overset{B}{A}}=\sqrt{\frac{\sum p_A q_B}{\sum P_B q_B}\times\frac{\sum p_A q_A}{\sum p_B q_A}}\times\sqrt{\frac{\sum p_B q_A}{\sum p_A q_A}\times\frac{\sum p_B q_B}{\sum p_A q_B}}=1 \tag{10.50}$$

以B地区为比较基准,用马埃公式计算,A,B两地空间价格指数为:

$$E_P^{\overset{A}{B}}=\frac{\sum p_A(q_A+q_B)}{\sum P_B(q_A+q_B)}=106.25\%$$

计算结果表明A地区水果价格比B地区高。

四、居民消费价格指数

居民消费价格指数在国外也被称为消费者价格指数(Consumer Price Index,简记为CPI),是度量一组代表性消费品及服务项目价格水平随时间而变动的指数,反映居民家庭所购买的生活消费品和服务价格水平变动的情况。通常被用来作为反映通货膨胀或通货紧缩程度的指标,观察和分析价格水平变动对居民货币工资的影响,作为研究居民生活和宏观经济分析和决策、价格总水平监测和调控的依据。我国现行居民消费价格指数的编制要点如下:

(一)消费品分类和代表规格品的选择

居民消费价格指数包括居民用于日常生活的全部消费品和服务项目。现行国家统计制度规定,将居民消费的商品分为八大类:①食品;②衣着;③家庭设备及用品;④交通和通讯工具;⑤娱乐教育文化用品;⑥医疗保健用品;⑦居住;⑧服务项目。每个大类包括若干个中类,中类之下又有基本分类,基本分类中包括若干代表规格品。如衣着大类分为服装、衣着材料、鞋帽袜及其他衣着三个中类;在鞋帽袜及其他衣着中类下又分为鞋类、袜子、帽子、其他衣着四个基本分类。

由于社会商品的种类极其繁多，每种商品的牌号、型号、规格、等级、花色、式样等千差万别。要编制包括所有商品规格的价格指数，在客观上是不可能的。因此，必须从全部商品中选择一些购销量较大的商品作为代表规格品。用这些代表规格品的价格升降情况来综合反映全部商品价格变动的趋势和程度。

（二）基本分类商品价格指数的计算

第一，计算某基本分类所属各规格品的环比价格指数。公式如下：

$$G_{t_i}=\frac{p_{t_i}}{p_{(t-1)_i}} \qquad (i=1,2,\cdots,n) \tag{10.51}$$

式中，G_{t_i}代表t期第i种代表规格品的环比价格指数，如果所属代表规格品有n种，就可以分别计算n个环比价格指数。

第二，计算各规格品环比价格指数的几何平均数。公式如下：

$$K_t=\sqrt[n]{G_{t_1}\times G_{t_2}\times\cdots\times G_{t_n}}\times 100\% \tag{10.52}$$

式中，K_t表示基本分类指数；$G_{t_1},G_{t_2},\cdots,G_{t_n}$分别为第1个至第$n$个代表规格品的环比价格指数。

（三）计算中类指数

基本分类指数乘以相应的权数，便得到中类指数。计算公式如下：

$$I_{中类}=\frac{\sum W_{t-1}K_t}{\sum W_{t-1}} \tag{10.53}$$

式中，W_{t-1}表示$t-1$期居民各类消费额比重。

【例10-9】从调查资料得到某月粮食中类各代表规格品的环比指数，如表10-5中所示，试计算粮食中类价格指数。

表10-5 粮食中类价格指数计算表

单位：%

	规格等级	计量单位	权数	指数	指数×权数
粮食中类			100		101.99
大米基本分类			60	102.5	61.50
	早　米（散装）	千克		100.0	
	东北米（真空包装）	千克		102.5	
	月牙米（袋装）	千克		105.1	
面粉基本分类			10	103.1	10.31
	富强粉			106.3	
	精致粉			100.0	
粮食制品基本分类			20	101.8	20.36
其他			10	98.2	9.82

大米基本分类指数为：

$$I_{大米}=\sqrt[3]{100.0\times 102.5\times 105.1}=102.5\%$$

面粉基本分类指数为：

$$I_{面粉}=\sqrt[2]{106.3\times100.0}=103.1\%$$

粮食中类指数为：

$$I_{粮食}=\frac{\sum W_{t-1}K_t}{\sum W_{t-1}}=\frac{102.5\times60+103.1\times10+101.8\times20+98.2\times10}{100}=101.99\%$$

（四）计算大类指数

各中类指数乘以相应的权数，便得到大类指数。计算公式的形式同中类指数。

（五）计算总指数

大类指数乘以相应的权数，便得到总指数。计算公式如下：

$$I_{总}=\frac{\sum W_{t-1}I_{t类}}{\sum W_{t-1}} \tag{10.54}$$

【例 10－10】根据表 10－6 的资料计算居民各类消费品价格指数。

表 10－6　居民消费价格指数计算表

类　　别	类指数 $I_{类}(\%)$	权数(W_{t-1})	总指数 $I_{总}=\frac{\sum W_{t-1}K_t}{\sum W_{t-1}}$
居民消费价格指数	100.87	100	100.87
食品	96.9	48.2	46.71
衣着	98.7	8.4	8.29
家庭设备及用品	97.8	5.30	5.18
交通和通讯工具	118.6	6.80	8.06
娱乐教育文化用品	100.8	2.40	2.42
医疗保健用品	99.6	7.10	7.07
居住	100.1	9.10	9.11
服务项目	110.5	12.70	14.03

$$居民消费价格指数\ I_{总}=\frac{\sum W_{t-1}I_{t类}}{\sum W_{t-1}}$$

$$=\frac{96.9\%\times48.2+118.6\%\times6.8+98.7\%\times8.4+97.8\%\times5.3+100.8\%\times2.4+99.6\%\times7.1+100.1\%\times9.1+110.5\%\times12.7}{100}=100.87\%$$

计算结果表明，该市居民消费品价格上涨了 0.87%。

五、股票价格指数

股票价格指数综合反映股票市场价格的变动程度，它是影响投资人决策行为的重要因素，而且股票价格的波动和走向也是反映经济景气状况的敏感指标。股票价格指

数的编制方法有多种，综合指数公式是其中的一种重要方法。我国的上证指数、美国标准普尔指数、香港恒生股票指数等，都是采用综合指数公式编制。其计算公式为：

$$I_p=\frac{\sum p_t q_0}{\sum p_0 q_0} \tag{10.55}$$

(10.55)式是以基期的股票发行量(或流通量)为同度量因素的拉氏综合指数。式中，q_0 代表基期股票发行量(或流通量)。

不同股价指数的样本范围和基期日期的选定都不同。例如，美国标准普尔指数，样本范围包括 500 种股票(其中工业股票 400 种、公用事业股票 40 种、金融业股票 40 种、运输业股票 20 种)，选择 1941～1943 年为基期。香港恒生指数选择了 33 种具有代表性的股票(成份股)为指数计算对象(其中金融业 4 种、公用事业 6 种、地产业 9 种、其他行业 14 种)，选择 1964 年 7 月 31 日为基期。而我国的上海证券交易所股票价格指数包括全部上市股票，基期为 1990 年 12 月 19 日，股票的基期指数定为 100，股票价格的变动幅度，以"点"数来表示，每上升或下降一个单位称为"1 点"。例如，当股票价格指数为 2 100 点时，就表明股票价格报告期比基期上升了 2 000 点。

六、指数时间序列

统计指数作为一种分析经济现象动态变化的统计指标，需要连续地进行编制，由此得到按时间顺序排列的一系列指数，构成指数时间序列。根据比较基期的不同，有环比指数时间序列和定基指数时间序列；根据采用的指数形式的不同，有拉氏和帕氏两种形式；根据指数所反映的内容的不同，又可以区分为质量指标指数时间序列和数量指标指数时间序列。下面是数量指标指数时间序列的基本形式：

$$\text{定基指数}\begin{cases}\text{拉氏：}\dfrac{\sum q_1 p_0}{\sum q_0 p_0},\dfrac{\sum q_2 p_0}{\sum q_0 p_0},\dfrac{\sum q_3 p_0}{\sum q_0 p_0},\cdots,\dfrac{\sum q_n p_0}{\sum q_0 p_0},\cdots\\ \text{帕氏：}\dfrac{\sum q_1 p_1}{\sum q_0 p_1},\dfrac{\sum q_2 p_2}{\sum q_0 p_2},\dfrac{\sum q_3 p_3}{\sum q_0 p_3},\cdots,\dfrac{\sum q_n p_n}{\sum q_0 p_n},\cdots\end{cases}$$

$$\text{环比指数}\begin{cases}\text{拉氏：}\dfrac{\sum q_1 p_0}{\sum q_0 p_0},\dfrac{\sum q_2 p_1}{\sum q_1 p_1},\dfrac{\sum q_3 p_2}{\sum q_2 p_2},\cdots,\dfrac{\sum q_n p_{n-1}}{\sum q_{n-1} p_{n-1}},\cdots\\ \text{帕氏：}\dfrac{\sum q_1 p_1}{\sum q_0 p_1},\dfrac{\sum q_2 p_2}{\sum q_1 p_2},\dfrac{\sum q_3 p_3}{\sum q_2 p_3},\cdots,\dfrac{\sum q_n p_n}{\sum q_{n-1} p_n},\cdots\end{cases}$$

如果采用不变价格 p_n 编制各期生产指数，可以得到观察长期经济发展趋势的生产指数时间序列：

$$\text{定基指数：}\frac{\sum q_1 p_n}{\sum q_0 p_n},\frac{\sum q_2 p_n}{\sum q_0 p_n},\frac{\sum q_3 p_n}{\sum q_0 p_n},\cdots,\frac{\sum q_n p_n}{\sum q_0 p_n} \tag{10.56}$$

$$\text{环比指数：}\frac{\sum q_1 p_n}{\sum q_0 p_n},\frac{\sum q_2 p_n}{\sum q_1 p_n},\frac{\sum q_3 p_n}{\sum q_2 p_n},\cdots,\frac{\sum q_n p_n}{\sum q_{n-1} p_n} \tag{10.57}$$

不难看出，不变价格产量定基指数时间序列和环比指数时间序列存在两个数量关系：①环比指数的连乘积等于相应时期的定基指数；②相邻时期的定基指数之商等于相应时期的环比指数。利用这两个数量关系，可以在定基指数和环比指数之间相互推算。

前面已经提到，我国的不变价格标准先后采用过1952年、1957年、1970年、1980年、1990年的不变价格标准。因此，在分析较长时期产量的动态时，遇到不变价格更换时，还必须消除不变价格本身变动的影响。消除不变价格变动的一般步骤是：①计算不变价格换算系数；②将以往各年按旧的不变价格计算的产值乘以换算系数，求得按新的不变价格计算的产值。不变价格换算系数计算公式为：

$$\text{不变价格换算系数} = \frac{\sum q_m p'_n}{\sum q_m p_n} \tag{10.58}$$

式中，q_m为交替年份的产量，p'_n为新的不变价格，p_n为旧的不变价格。

【例10·11】某省2000年工业总产值按1990年不变价格计算为582亿元，1990年工业总产值按1980年不变价格计算为242亿元。又知该省1991年工业总产值按1980年不变价格计算为260亿元，按1990年不变价格计算为270亿元。试计算该省2000年对比1990年的工业生产指数。

$$\text{不变价格换算系数} = \frac{\sum q_{91} p_{90}}{\sum q_{91} p_{80}} = \frac{270}{260} = 1.04$$

1990年不变价格计算的工业总产值：

$$242 \times 1.04 = 251.68(\text{亿元})$$

2000年对比1990年的工业生产指数：

$$\frac{582}{251.68} = 231.25\%$$

计算结果表明该省2000年工业生产比1990年增长了231.25%。

本章小结

1. 统计指数是用来分析社会经济现象数量变动的对比性指标。广义指数是指一切说明社会经济现象数量变动的相对数。狭义的指数是一种特殊的相对数，即用来说明不能直接相加的复杂社会经济现象综合变动程度的相对数。

2. 综合指数的编制特点是先综合后对比。编制综合指数必须明确指数化指标和同度量因素，指数化指标是被测定的因素，同度量因素也即权数，作为同度量因素的指标固定在哪个时期上，不是固定不变的。拉氏指数将同度量因素固定在基期水平上；帕氏指数将同度量因素固定在报告期水平上。通常情况下，数量指标指数按拉氏公式计算，

质量指标指数按帕氏公式计算。

3.平均指数的编制特点是从个体指数出发,先对比后平均。平均指数有算术平均指数和调和平均指数两种形式。算术平均指数一般用基期总值来加权;调和平均指数一般用报告期总值来加权。

4.指数的因素分析是对现象总变动中各个因素变动的影响程度,用连锁替代法从相对数和绝对数两个方面进行分析。利用综合指数体系,可以分析现象总变动中数量因素和质量因素的影响。相对数分析是从各个指数计算结果来分析;绝对数分析是从各个指数分子与分母指标之差来分析。

5.总平均数指数是对总体平均指标的测定,分析总平均数指数的变动,需要计算3种指数:①可变构成指数;②固定构成指数;③结构变动影响指数。利用上述3种指数的相互联系,可以分析现象总体平均指标变动受各组平均水平和各组结构变动的影响程度。

6.指数时间数列是将一系列总指数按照时间先后顺序排列形成的数列,用以分析现象的动态变化趋势。指数时间数列分为环比指数数列和定基指数数列。在指数时间数列中,各个时期指数的同度量因素所属时期是变动的,称可变权数;各个时期的同度量因素所属时期是固定的,称不变权数。

7.空间价格指数用拉氏指数与帕氏指数度量不大合适。空间价格指数编制和分析的特殊要求是:互换基准后指数的结论应保持一致。

8.我国现行居民消费价格指数的编制步骤:①消费品分类和代表规格品的选择;②基本分类商品价格指数的计算;③计算中类指数;④计算大类指数;⑤计算总指数。

思考与练习题

一、选择题:

1.编制综合指数时,应固定的因素是(　　)。

a.指数化指标　　b.个体指数　　c.同度量因素　　d.被测定的因素

2.指出下列哪一个数量加权算术平均数指数恒等于综合指数形式拉氏数量指标指数(　　)。

a. $\frac{\sum k_q q_0 p_1}{\sum q_0 p_1}$　　b. $\frac{\sum k_q q_1 p_1}{\sum q_1 p_1}$　　c. $\frac{\sum k_q q_0 p_0}{\sum q_0 p_0}$　　d. $\frac{\sum k_q q_1 p_0}{\sum q_1 p_0}$

3.同度量因素,也可以叫权数,是因为(　　)。

a.它使得不同度量单位的现象总体转化为数量上可以加总

b.客观上体现它在实际经济现象或过程中的份额

c. 是我们所要测定的那个因素

d. 它必须固定在相同的时期

4. 编制数量指标综合指数所采用的同度量因素是(　　)。

a. 质量指标　　b. 数量指标　　c. 综合指标　　d. 相对指标

5. 空间价格指数一般可以采用(　　)指数形式来编制。

a. 拉氏指数　　b. 帕氏指数　　c. 马埃公式　　d. 平均指数

二、问答题:

1. 某企业生产 A,B 两种产品,报告期和基期产量、出厂价格资料如表 10－7 所示。

表 10－7

产品	产量(件)		出厂价(元)	
	基期	报告期	基期	报告期
A	2 000	2 200	12.0	12.5
B	5 000	6 000	6.2	6.0

要求:

(1)用拉氏公式编制产品产量和出厂价格指数。

(2)用帕氏公式编制产品产量和出厂价格指数。

(3)比较两种公式编制的产量和销售量指数的差异。

2. 某地区 1995 年和 2000 年两类商品收购价格类指数和收购额指数资料如表 10－8所示。

表 10－8

商品种类	收购总额(万元)		收购价格类指数(%)
	1995 年	2000 年	
甲	140	138.6	105
乙	60	78.4	98

要求:

编制这两类商品收购价格总指数。

3. 据调查,某地甲、乙、丙、丁四种商品的个体价格指数分别为 110%,95%,100% 和 105%,各类代表商品的固定权数分别为 10%,30%,40% 和 20%。

要求:

计算这四种商品价格总指数。

4. 某企业生产三种产品,有关资料如表 10－9 所示。

表 10－9

产品种类	基期		产量个体指数（%）
	价格(元)	产量(万件)	
甲	10	20	109.2
乙	9	16	121.5
丙	8	15	98.6

要求：

计算该企业三种产品产量总指数。

5.某地区 1999 年农副产品收购总额为 400 亿元，2000 年比 1999 年的收购总额增长了 10%，农副产品收购价格总指数为 102%。试问 2000 年与 1999 年对比：

(1)农民因交售农副产品共增加多少收入？

(2)农副产品收购量增加了百分之几？农民因此增加了多少收入？

(3)由于农副产品收购价格提高了 2%，农民又增加了多少收入？

(4)验证以上三方面的分析结论能否协调一致。

6.甲、乙两城市水果销售资料如表 10－10 所示。

表 10－10

种类	甲城市		乙城市	
	价格(元)	销售量(千克)	价格(元)	销售量(千克)
1	14	200 000	20	50 000
2	20	100 000	12	150 000

要求：

以乙城市为比较基准分别用马埃公式计算甲、乙两城市水果价格比较指数，并加以简要文字说明。

7.某工厂生产两种不同种类的产品，有关资料如表 10－11 所示。

表 10－11

产品种类	计量单位	产量		价格(元)	
		基期	报告期	基期	报告期
甲	件	20 000	24 600	40	45
乙	台	108	120	500	450

要求：

(1)计算该厂工业总产值指数及总产值增长额。

(2)从相对数和绝对数两方面对总产值变动进行因素分析。

(3)用文字说明分析结果。

8.某公司所属两企业有关资料如表 10－12 所示。

表 10－12

企业	平均工人数(人)		工作月长度(日)		工作日长度(时)		单位劳动生产率(件)	
甲	200	220	26	25	7.5	7.5	100	120
乙	200	200	26	25	8.0	7.0	120	125
合计	400	420	—	—	—	—	—	—

要求：

根据上列资料分析说明该公司工业总产量变动情况及劳动生产率、工作日长度、工作月长度和工人数各因素影响程度和绝对额。

9.某市 2001 年与 2000 年对比的各类零售商品价格指数及权数资料如表 10－13 所示。

表 10－13　某市居民消费价格指数计算表

类　　别	平均价格(元)		指数(%)	权数
	基期	报告期		
居民消费价格指数				100
食品类			96.9	48
衣着类			98.3	5
家庭设备及用品类			97.8	5
医疗保健			100.8	3
交通和通讯工具				7
交通工具			99.6	(40)
通讯工具				(60)
固定电话机	200	180		
移动电话机	1 900	1 980		
娱乐教育文化用品类			100.1	10
居住			110.5	13
服务项目			118.6	9

要求：

计算居民消费价格指数。

10.某企业1988～1995年按不变价格计算的总产值资料如表10－14所示。

表10－14　　单位:万元

	1988	1989	1990	1991	1992	1993	1994	1995
按1980年不变价格计算	300	316	360	400				
按1990年不变价格计算				416	440	462	500	540

要求:

(1)计算不变价格换算系数。

(2)将1992年以后各年产值换算为按1980年不变价格计算的总产值。

(3)将1990年以前各年产值换算为按1990年不变价格计算的总产值。

(4)以1988年为基期编制各年的产量定基指数。

11.表10－15是某企业1995年到2000年各种产品产量及不变价格资料。

表10－15

产品	单位	不变价格(元/件)	产量					
			1995	1996	1997	1998	1999	2000
甲	万件	15	220	240	250	300	320	320
乙	万台	8	480	500	500	550	480	520
丙	吨	5	800	800	900	1 000	980	1 200

要求列表计算:

(1)各年按不变价格计算的总产值。

(2)以1995年为基期的各年产值的定基指数。

(3)各年环比指数。

(4)验证定基指数和环比指数的关系。

12.生产甲、乙、丙三种产品的某企业,1996年和2000年的销售数据如表10－16所示。

表10－16

产品名称	销售额(万元)	
	1996	2000
甲	700	800
乙	600	700
丙	650	850

以上年为基期的产品价格指数如表 10－17 所示。

表 10－17

产品名称＼年份	1997（%）	1998（%）	1999（%）	2000（%）
甲	108	112	105	95
乙	126	122	130	118
丙	103	105	106	105

要求：

(1)计算丙产品 1996～2000 年价格平均增长率。

(2)计算三种产品以 1996 年为基期的 2000 年价格总指数。

13. 某企业工人数及工资资料如表 10－18 所示。

表 10－18

	工人数		工资水平	
	基期	报告期	基期	报告期
技术工	400	630	1 600	1 700
普通工	600	870	800	900

要求：

(1)计算各组工人工资个体指数。

(2)计算总平均工资指数。

(3)对总平均工资变动进行因素分析。

(4)比较各组工人工资个体指数与总平均工资指数的差异，并分析出现差异的原因。

(5)计算由于平均工资水平的变动对企业工资总额支出的影响额。

第十一章

统计综合评价

★ 本章学习要点与要求 ★

本章介绍统计综合评价的基本概念及其常规评价的方法。通过学习要求：①能够以定量的方法——聚类法确定评价指标；②掌握对评价指标进行同度量处理的三种方法；③相对化处理法、功效系数法和标准化处理法。掌握评价体系中权重的计算方法，能够利用不同的方法(包括总分评定法、加权平均综合法和最优值距离法)和对单项评价结果进行综合。

第一节　综合评价概述

一、统计综合评价的基本步骤

统计评价是对所研究的对象及其组成部分的数量规模大小、水平高低、速度快慢、质变程度及内部协调状况作出定量的判断。评价包括单项评价与综合评价。单项评价是利用一个统计指标对事物的某一个方面作出判断。综合评价，则是根据研究的目的建立一个统计指标体系，对事物的各个方面进行定量分析，得出概括性的结论，从而揭示事物的本质及其发展规律。

统计综合评价的基本步骤如下：

(一)明确评价的目标

进行统计综合评价,必须先明确其评价的目标。即搞清楚为什么要进行综合评价,确定所要评价的对象、评价的精确度以及评价所要说明的问题等等。

(二)建立评价指标体系

根据研究的目的,选择合适的统计指标,建立一个能够从不同角度,不同侧面反映评价对象的评价指标体系。评价指标体系可以是单一层次的,也可以是多层次的。例如,我国评价独立核算工业企业的经济效益,是利用工业增加值率、总资产贡献率、资产负债率、流动资产周转次数、工业成本费用利润率、全员劳动生产率、产品销售率等项指标进行评价,该评价项目体系是单一层次的。对于因素更多的复杂现象,则可用二层次或多层次综合评判法进行评价。

(三)评价指标的量化处理

综合评价需要运用由多个指标组成的指标体系。而这些指标性质不同,计量单位往往也不一样,因而必须先确定各单项指标的评价标准,进行同度量处理,使之具有可比性,在此基础上才能综合汇总。

(四)确定各评价指标的权数

综合评价涉及许多因素,而每一因素又有各种不同的评价值。例如,一种产品是否受欢迎受到多种因素的影响,这些因素包括了产品的质量、性能、价格、式样、包装及售后服务等等,对上述这六个因素作出综合评价,而这六个因素对我们评价的目标——产品受欢迎程度所起的作用强度是不同的。因此,在综合评价中,权衡不同项目重要性的数值被称为权重,或称权数,在评价中,须根据评价的目的和各个项目的内在涵义对各个目标值赋予相应的权数。

如果评价项目体系是单一层次的,要求所有指标的权数总和等于100%。如果评价项目体系是多层次的,则要求每一层次中,同一层次各项目的权数总和都等于100%。

(五)进行综合评价

根据前面建立的评价指标体系和相应的权数体系,在单项评价的基础上,采用一定的方法,汇总计算与分析所要研究的事物,得出综合的判断。

二、统计综合评价的特点

统计综合评价的特征是定量与定性分析相结合,它是在定性分析的前提下,通过现象的数量表现,对研究对象进行更深刻、更全面的认识。就统计活动过程而言,统计综合评价是在统计调查、统计整理之后的一项重要工作,是充分发挥统计功能的重要环

节。

应该指出，由于目前统计综合评价的理论与方法还不是十分成熟，它也存在一定的局限性。主要表现在以下两点：

（一）综合评价结果具有相对性

综合分析与评价尽管采用了一定的数学模式，其结果用数值表示，但它只有相对的意义，一般情况下，它仅适用于在性质相同的对象之间进行比较和排序。此外，采用不同的评价方法，有可能得出不同的结果。

（二）综合评价结果常带有主观性

在综合评价中，各指标的评价标准以及权数的确定，常常需要依靠有关专家来确定，不同的专家给出的标准和权重会有所差异，因此，综合评价的结果，往往带有一定的主观性。

我们在开展统计综合评价时，必须注意各种方法的特点，采用多种方法进行比较与分析，尽可能地减少主观因素的干扰，提高评价结果的客观性和稳定性。

第二节 评价指标的选择与数据处理的方法

一、选择评价指标的原则

如前所述，对客观现象进行统计综合评价，关键是要科学地选择评价项目，建立一个合适的评价指标体系。选择评价指标应遵循如下原则：①目的性指标的选择应与研究目的相吻合；②客观性，评价指标体系能够准确地把握所要研究问题的本质和内涵，能够客观地反映事物的特征；③全面性，各评价指标能从不同的角度综合反映分析对象的全貌；④敏感性，所选择的指标能比较敏感地反映分析对象的变化；⑤相互独立性，尽可能选择相关程度低的指标，这是因为如果指标之间相关程度过高，用一项指标就能说明问题，采用多项指标，事实上加大了这类指标的权重；⑥可比性，评价指标要涵义明确，计算口径一致，达到动态可比、横向可比；⑦可操作性，评价指标体系要考虑资料收集的可能性，尽可能地利用现有的统计资料。

二、评价指标的选择方法

在实践中，根据上述原则选择具体评价指标的方法有两种：定性方法和定量方法。

（一）定性方法

常用的定性方法有:综合法和分析法。

综合法一般是通过研讨会或征询意见的方式,集中专家们的意见,以确定评价指标。该方法是借着专家的智力优势和经验以选择指标的。

分析法是将评价的对象划分为若干个组成部分或不同的侧面,明确各个侧面所要评价问题的内涵与外延,在此基础上,对每一侧面分别选用一个或若干个指标以反映评价对象的特征。

(二)定量方法

常用的定量方法有:试算法和系统聚类法。

试算法是通过历史资料的试算来判断指标的有效性。例如,要评价各地区2002年可持续发展战略的实施绩效,可以用2000和2001年的资料进行试算,通过试算判断所选择的指标是否合适,然后反复修正指标体系直至满意为止。

系统聚类法是通过判断指标之间的相似程度来筛选指标的一种方法。假设有 N 个指标,将每个指标看作一类,根据指标间的相似程度,通过比较类间距离进行并类。把距离最小的两类加以合并,此时,余下 $N-1$ 类;再选择类间距离最小的加以合并。这样,每合并一次,就减少一类,反复继续这一过程,最终形成由小到大的分类系统。整个分类过程可以绘制成一张聚类图,用以反映所有指标的亲疏关系,我们可依据这种关系确定指标体系中所包含的指标个(类)数。系统聚类法的具体操作步骤如下:

第一步,度量指标(类)间的相似程度。度量指标(类)间的相似程度常用的方法是相关系数或判别系数法。根据 N 个指标的历史资料,分别计算两两指标的相关系数(或判别系数)并形成相关系数矩阵 R(或判别系数矩阵 R^2),以相关系数矩阵 R(或判别系数矩阵 R^2)表示指标间的相关关系。

第二步,度量指标(类)间的距离。利用相关系数矩阵 R(或判别系数矩阵 R^2)表示指标(类)间相似程度时,必须将其变换为距离 d,d 值越小表明两指标关系越密切。$d=1-|r_{ij}|$或者 $d^2=1-r_{ij}^2$。

第三步,根据聚类图确定指标(类)的个数,从每类中选择出最具有代表性的指标。

【例11-1】现有6个指标,根据历史资料计算每两个指标的相关系数,建立相关系数矩阵 R(见表11-1),并将相关系数矩阵 R 转变为距离矩阵(见表11-2)。

在距离矩阵中(见表11-2)找到距离最小的两个指标,由 $d_{34}=0.11$ 得知,指标3和指标4关系最为密切,可聚为一类。又在距离矩阵中找到距离第二小的两个指标,由 $d_{56}=0.20$ 得知,指标5和指标6的关系密切程度仅次于指标3与指标4,它们可聚为一类。由 $d_{12}=0.25$ 得知,指标1和指标2关系较为密切,可聚为一类。如此继续这一过程,逐步选择较小距离的指标,直至指标聚为一类为止,并据此绘制聚类图,具体图形如图11-1所示。

表 11－1 相关系数矩阵

指标	1	2	3	4	5	6
1	1	0.75	0.58	0.54	0.44	0.60
2		1	0.55	0.56	0.56	0.62
3			1	0.89	0.73	0.60
4				1	0.74	0.72
5					1	0.80
6						1

表 11－2 距离矩阵

指标	1	2	3	4	5	6
1	0	0.25	0.42	0.46	0.56	0.40
2		0	0.45	0.44	0.44	0.38
3			0	0.11	0.27	0.40
4				0	0.26	0.28
5					0	0.20
6						0

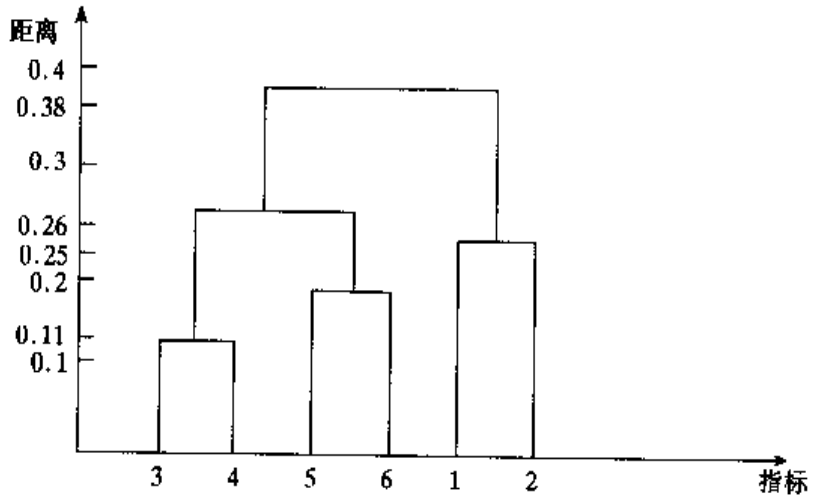


图 11－1 聚类图

通过定量和定性判断，根据评价的目的、实际操作的可能性和各个指标之间关系的密切程度确定指标体系的容量。假定评价某一现象要选择四个评价指标，可在指标3和指标4中选择一个指标，将它确定为第一个评价指标，又在指标5和指标6中选择一个指标，将它确定为第二个评价指标，确定指标1为第三个评价指标，确定指标2为第四个评价指标。又假定评价另一现象要选择两个评价指标，可在指标3、指标4、指标5和指标6中选择一个指标，将它确定为第一个评价指标，又在指标1和指标2中选择一个指标，将它确定为第二个评价指标。

三、评价项目的量化处理方法

在综合评价中，有些评价项目本来属于定性评价，对这些项目应进行量化处理。例如，评价工程项目质量是以优、良、合格、不合格作为评价标准的；顾客对某项服务的评价是以满意、一般、不满意来反映的。量化处理常用的方法是：先对评价的各等级赋予不同的分值，再聘请足够数量的评议者，请评议者对评价对象进行评议，在所列的评语等级中选择一个他认为最恰当的等级，而后计算各等级得票数，再用得票数对比总票数，计算出得票频率。

现以一实例说明对评价项目进行量化处理的步骤。

【例11－2】对教师教学质量进行综合评价，建立教学质量评估体系如表11－3所示。假定对教学质量的评价等级分为：好、较好、一般和较差四个等级，各评价项目的权数分别为：0.40，0.30，0.10，0.20，对教师教学质量评价项目进行量化处理。

表11－3　教学质量评估体系

评价项目	评分标准				权重
	10	8	6	4	
教学内容	好：符合大纲、内容丰富、联系实际	较好	一般	较差	0.40
教学方法	好：讲授生动、深入浅出、有启发性	较好	一般	较差	0.30
教学态度	好：认真对待教学、虚心听取学生意见	较好	一般	较差	0.10
教学效果	好：课堂气氛活跃、学生知识扎实	较好	一般	较差	0.20

解：

对教师教学质量评价的四个等级“好”、“较好”、“一般”和“较差”量化处理的具体步骤：

第一步，对这四个等级分别赋予不同的分值：10分、8分、6分和4分，由此将评语等级“好”、“较好”、“一般”和“较差”转换为相应的数值：10，8，6，4。

第二步,对各评价项目评判。假定请100名学生对F教师分别就上述4个评估项目评分,根据得票数计算各评语等级的得票率,如表11-4所示。

表11-4 100名学生对F教师教学质量评分结果分组表

评估项目	好(10)		较好(8)		一般(6)		较差(4)	
	得票数	得票率	得票数	得票率	得票数	得票率	得票数	得票率
教学内容	40	0.4	40	0.4	20	0.20	0	0
教学方法	24	0.24	60	0.6	16	0.16	0	0
教学态度	50	0.50	36	0.36	10	0.10	4	0.04
教学效果	30	0.30	20	0.20	30	0.30	20	0.20

第三步,计算单个评价项目的分值如下:

F教师的教学内容得分$=10\times0.4+8\times0.4+6\times0.2+4\times0=8.4$

F教师的教学方法、教学态度和教学效果的得分分别为:8.16,8.64和7.2。

第四步,综合评价结果。F教师的教学质量总分为:$8.4\times0.4+8.16\times0.3+8.64\times0.1+7.2\times0.2=8.112$。教学质量总分值是10,假定得分9以上者为"好",7.5~8.9者为"较好",可以得出结论,F教师的教学质量为"较好"。

四、同度量处理方法

在指标体系建立之后,由于各指标的计量单位不同(具有不同的量纲)而无法进行直接汇总,为此,一般在完成资料搜集工作后,还需要进行消除量纲处理,即同度量处理。较常用的同度量处理方法有如下三种:

(一)相对化处理方法

进行相对化处理,必须先对每个评价指标确定一个标准值,而后计算实际值与标准值之比。

指标有"正"、"逆"之分,所谓正指标,就是指指标数值越大就越好的指标,例如,产值、利润、劳动生产率等指标。所谓逆指标,就是指指标数值越小就越好的指标,例如,单位产品成本、万元产值综合能耗、物耗率等指标。对正、逆指标的相对化处理公式如下:

$$\text{正指标}: x_i' = x_i / x_m \qquad \text{逆指标}: x_i' = x_m / x_i$$

式中,x_i为实际值,x_m为标准值。通常,可用参加评价单位某一时期的平均值或最优值作为标准值。根据研究目的不同,也可以选择国际先进水平、历史最高水平或计划规定水平等作为标准值。

【例 11－3】2000 年我国 4 个直辖市的国有及规模以上非国有工业企业主要经济效益指标资料如表 11－5 所示，试对这些资料进行相对化处理。

解：

表 11－5 所列示的指标均为正指标，以实际值对比标准值。经过相对化处理后，将原来不同度量的指标转化为无量纲的、同度量的相对指标。具体的处理结果见表11－5 和表 11－6 所示。

表 11－5　我国 4 个直辖市国有及规模以上非国有工业企业主要经济效益指标

指标 / 直辖市	工业增加值率（%）	总资产贡献率（%）	流动资产周转次数（次/年）	工业成本费用利润率（%）	全员劳动生产率（元/人年）	产品销售率（%）
北　京	28.17	6.94	1.43	5.02	63 878	95.95
天　津	24.18	8.50	1.79	6.48	52 425	98.39
上　海	27.19	8.87	1.56	6.43	82 327	99.02
重　庆	29.48	6.09	1.18	1.67	31 252	99.09

资料来源：国家统计局，中国统计年鉴 2001 年．北京：中国统计出版社，2002

表 11－6　相对化处理结果资料表

指标 / 直辖市	工业增加值率	总资产贡献率	流动资产周转次数	工业成本费用利润率	全员劳动生产率	产品销售率
标准值*	30.00	9.00	1.80	6.00	60 000	99.00
北　京	0.94	0.77	0.79	0.84	1.06	0.97
天　津	0.81	0.94	0.99	1.08	0.87	0.99
上　海	0.91	0.99	0.87	1.07	1.37	1.00
重　庆	0.98	0.68	0.66	0.28	0.52	1.00

*表中标准值为假定。

（二）功效系数法处理

功效系数法是对多目标规划原理中的功效系数加以改进，经计算而得到综合评判的分数。它借助功效系数，把确定要评价的各项指标值转化为可以度量的评判分数。

利用功效系数法进行消除量纲影响的处理，必须对评价的指标确定一对阈值，包括

一个下限值(不容许值)和一个上限值(满意值),并通过功效系数公式计算出每项指标的评价分,其计算公式如下:

$$d_i^* = \frac{x_i - x_i^{(s)}}{x_i^{(h)} - x_i^{(s)}} \times 40 + 60 \tag{11.1}$$

式中,x_i 为第 i 个指标实际值;$x_i^{(s)}$ 为第 i 个指标的不容许值;$x_i^{(h)}$ 为第 i 个指标的满意值,d_i^* 为第 i 项指标单项评分。一般地,对于不容许值可令第 i 项指标历年来最差值为不容许值,或第 i 项指标数列中较差的 10% 的数值平均数为不容许值;对于满意值,可令第 i 项指标中历年来最优值为满意值,或第 i 项指标数列中较好的 10% 的数值平均数为满意值。

在上述公式中,把指标值处于不容许状态看做及格状态,当某项指标值等于不容许值时,d_i^* 为 60 分。一般情况下,实际指标值在不容许值和满意值之间,若 x_i 比 $x_i^{(h)}$ 值更高,$d_i^* > 100$ 分;若 x_i 比 $x_i^{(s)}$ 值更低,则 $d_i^* < 60$ 分。根据表 11-5 所提供的资料,利用公式(11.1)计算:

北京市工业增加值率功效系数分值:$d^* = \frac{28.17-24}{50-24} \times 40 + 60 = 66.4$,

北京市总资产贡献率功效系数分值:$d^* = \frac{6.94-6}{10-6} \times 40 + 60 = 69.4$

同理,我们可以得到各个直辖市有关工业经济效益指标的功效系数分值。详细的计算结果列在表 11-7 中。

表 11-7 功效系数法处理结果

指标 / 直辖市	工业增加值率	总资产贡献率	流动资产周转次数	工业成本费用利润率	全员劳动生产率	产品销售率
不容许值*	24.00	6.00	1.2	2.00	30 000	95.00
满意值*	50.00	10.00	1.8	10.00	90 000	100.00
北 京	66.4	69.4	75.3	75.1	82.6	67.6
天 津	60.3	85.0	99.3	82.4	75.0	87.1
上 海	64.9	88.7	84.0	82.2	94.9	92.2
重 庆	68.4	60.9	58.7	58.4	60.8	92.7

* 表中不容许值和满意值为假定。

(三)标准化处理

利用标准化处理的基本前提是：须进行标准化处理的变量服从正态分布。在标准化处理中，将变量值转化为数学期望为0，方差为1的标准化数值。具体操作步骤如下：

(1)求出各变量(指标)的算术平均值(数学期望值)$\bar{x}_i$和标准差σ_i。

(2)利用公式(11.2)进行标准化处理：

$$x'_{ij}=(x_{ij}-\bar{x}_i)/\sigma_i \tag{11.2}$$

式中，x'_{ij}为标准化变量；x_{ij}为实际变量(指标)值；$\bar{x}_i$为各变量(指标)的算术平均值(数学期望值)；σ_i为标准差。

(3)进行标准化处理的指标中若有逆指标，改变处理后的指标的符号。

经标准化处理后的数值围绕着0上下波动，数值大于0的说明高于平均水平，数值小于0的说明低于平均水平。利用表11－5所提供的资料进行标准化处理，先计算各指标的平均数与标准差，而后分别对各直辖市的指标值进行标准化：

$$\text{北京市工业增加值率标准值}=\frac{28.17-27.26}{1.95}=0.47$$

$$\text{北京市总资产贡献率标准值}=\frac{6.94-7.6}{1.13}=-0.58$$

同理，我们可以得到4个直辖市各个指标的标准值。具体计算处理结果见表11－8。

表11－8　标准化处理结果

指标 / 直辖市	工业增加值率	总资产贡献率	流动资产周转次数	工业成本费用利润率	全员劳动生产率	产品销售率
平均数$\bar{x}$*	27.26	7.6	1.49	4.9	57 470.5	98.11
标准差*	1.95	1.13	0.22	1.95	18 518.7	1.28
北　京	0.47	－0.58	－0.27	0.06	0.35	－1.69
天　津	－1.58	0.80	1.36	0.81	－0.27	0.22
上　海	－0.04	1.12	0.32	0.78	1.34	0.71
重　庆	1.14	－1.34	－1.41	－1.66	－1.42	0.77

*表11－8中的平均值与标准差是根据表11－5的资料计算的。

第三节 权重的确定与评价结果的综合

一、权重的分类

在综合评价中，不同类别的权重往往代表着不同的经济含义和不同的数学特点。在综合评价中的统计权数主要有如下几种分类：

按权数的表现形式划分，可分为绝对数形式权数和相对数形式权数（或称比重权数），相对数权数能较直观地显示权数在评价中的作用。

按权数的形成方式划分，可分为自然权数和人工权数。由于变换统计资料的表现形式与统计指标合成方式而得到的权数是自然权数，这种权数也被称为客观权数。人工权数是指根据研究目的和评价指标的内涵，人为地构造出反映各个评价指标重要程度的权数，这种权数也被称为主观权数。

按权数形成的数量特点划分，可分为定性赋权和定量赋权。在综合分析评价过程中，定性赋权和定量赋权的方法结合使用，往往能获得较好的效果。

按权数与待加权的单个指标之间的相关程度划分，可分为独立权数与相关权数。独立权数是指评价指标的权重与该指标值的大小无关，在综合评价中较多的运用这种权数，基于这种权数建立的综合评价模型被称为“定权综合”模型。相关权数是指评价指标的权重与该指标值呈函数关系，例如，当某一评价指标值达到一定水平时，该指标的重要性相应减弱；或者当某一评价指标值达到另一定水平时，该指标的重要性相应增大。这种权数产生于评价指标的重要性随着指标取值的不同而发生变化的场合，基于这种权数建立的综合评价模型被称为“变权综合”模型。“变权综合”模型较多的运用于对环境质量的评价。

二、确定权重的方法

权数的确定方法有多种，本书仅介绍几种较常用的方法。

（一）统计平均法

统计平均法是根据专家们对各评价指标所赋予的相对重要性系数分别求其算术平均数，所得的平均数作为各指标的权重。

（二）最大组中值法

评价指标体系 U 中含有 $u_1, u_2, \cdots, u_n$ 等指标，请有关专家（$m \geq 30$ 人组成）根据

权数分配表，对 U 中的各项指标，提出自己认为最合适的权数。再根据专家估测的资料，进行单因素的权数统计，具体步骤如下：

1.对于指标 $u_i(i=1,2,\cdots,n)$，在专家所赋予的权数 a_{ij} 中找出最大值 M_i 和最小值 m_i。$M_i=\max\{a_{ij}\}$，$m_i=\min\{a_{ij}\}(i=1,2,\cdots,n;j-1,2,\cdots,m)$

2.适当选取正整数 P，利用公式 $(M_i-m_i)/P$ 计算出权数分组的组距，并将权数从小到大分为 P 组。

3.计算各组的权数的频率与频数。

4.根据频数与频率分布情况，取最大频率所在组的组中值为指标 u_i 的权数 $a_i(i=1,2,\cdots,n)$，从而得出权重向量。

【例 11－4】某公司对管理人员的工作质量进行评估，建立评价指标体系 U，U 中含有四个指标，u_1：管理水平；u_2：业务知识；u_3：廉洁奉公；u_4：工作态度。现请 100 名职员对这四个指标写出自己认为最合理的权重分配，据此确定四个评价指标的权数。

解：

根据 100 名职员所列的权重分配表，对于指标 u_1（管理水平），最大值 $M_1=0.5$，最小值 $m_1=0.05$。选 $P=5$ 作为组数，则组距 $=(0.5-0.05)/5=0.09$，具体分组情况如表 11－9 所示。

表 11－9　管理水平指标权重分配表

分组	0.05～0.14	0.14～0.23	0.23～0.32	0.32～0.41	0.41～0.50	Σ
频数	8	33	40	10	9	100
频率	0.08	0.33	0.40(max)	0.10	0.09	1.00

对于指标 u_2（业务知识），最大值 $M_2=0.75$，最小值 $m_1=0.25$，组距 $=(0.75-0.25)/5=0.1$，分组情况如表 11－10 所示。

表 11－10　业务知识指标权重分配表

分组	0.25～0.35	0.35～0.45	0.45～0.55	0.55～0.65	0.65～0.75	Σ
频数	7	25	43	20	5	100
频率	0.07	0.25	0.43(max)	0.20	0.05	1.00

对于指标 u_3（廉洁奉公），最大值 $M_3=0.50$，最小值 $m_3=0.05$，组距 $=\frac{0.50-0.05}{5}=0.09$，分组情况如表 11－11 所示。

表 11－11 廉洁奉公指标权重分配表

分组	0.05～0.14	0.14～0.23	0.23～0.32	0.32～0.41	0.41～0.5	Σ
频数	12	55	18	9	6	100
频率	0.12	0.55(max)	0.18	0.09	0.06	1.00

对于指标 u_4(工作态度),最大值 $M_4=0.25$,最小值 $m_4=0$,组距$(0.25-0)/5=0.05$,分组情况如表 11－12 所示。

表 11－12 工作态度指标权重分配表

分组	0～0.05	0.05～0.1	0.1～0.15	0.15～0.2	0.2～0.25	Σ
频数	8	33	49	6	4	100
频率	0.08	0.33	0.49(max)	0.06	0.04	1.00

表 11－9 至表 11－12 反映了 100 名职员对四个评价指标给出的权重系数的分布情况,取频率最大组的组中值作为相应指标 u_i 的权。如果所得的权数之和不为 1,可做相应的调整。如本题,各评价指标中,频率最大的组中值分别为 0.275,0.5,0.185,0.125,它们的和不为 1,须调整。

$$\text{调整系数}=\frac{1}{0.275+0.5+0.185+0.125}=0.921\ 66$$

$0.275\times0.921\ 66=0.253$,$0.5\times0.921\ 66=0.461$,$0.185\times0.921\ 66=0.171$,$0.125\times0.921\ 66=0.115$,由此,我们可以得到评价指标的权重向量 $A=(0.253,0.461,0.171,0.115)$。

(三)AHP(Analytic Hierarchy Process)构权法

AHP 构权法是利用运筹学的层次分析法的原理,确定统计综合评价中各评价指标(或项目)的相对重要性系数。运用层次分析法进行构权,基本思路是:将复杂的评价对象表示为一个有序的递阶层次结构的整体,通过人们在各个评价项目间进行两两的比较、判断,进而计算各个评价项目的相对重要性系数,即权数。AHP 构权法有单准则构权法和多准则构权法之分,本节只介绍单准则构权法。用单准则 AHP 构权法确定权数的具体步骤如下:

1.确定指标的量化标准。AHP 构权法的关键是建立一个构造合理且一致的判断矩阵,判断矩阵的合理性受到标度的合理性的影响。标度是指人们对各个评价指标(项目)重要性的等级差异进行量化的概念。确定指标重要性的量化标准常用的方法有:比例标度法和指数标度法。以比例标度法为例,比例标度法是以人们对事物质的区别的

评判标准为基础，一般地，以5种判别就可以很好地表示事物质的等级差别。当评判需要更高的精度时，可以在相邻判别之间做出比较，形成9种判别，对判别等级量化，用数值表示就是9个标度。常见的标度值体系如表11－13所示。

表11－13　比例标度值体系(重要性分数 x_{ij})

取值含义	1～9标度	5/5～9/1标度	9/9－9/1标度
i与j同等重要	1	(5/5＝)1	(9/9＝)1
i比j稍微重要	3	(6/4＝)1.5	(9/7＝)1.286
i比j明显重要	5	(7/3＝)2.33	(9/5＝)1.8
i比j强烈重要	7	(8/2＝)4	(9/3＝)3
i比j极端重要	9	(9/1　)9	(9/1＝)9
介于上述相邻两级之间重要程度的比较	2,4,6,8	(5.5/4.5＝)1.222 (6.5/3.5＝)1.857 (7.5/2.5＝)3 (8.5/1.5＝)5.67	(9/8＝)1.125 (9/6＝)1.5 (9/4＝)2.25 (9/2＝)4.5
j与i的比较	上述各数的倒数	上述各数的倒数	上述各数的倒数

2.初始权数的形成。在综合评价中，确定初始权数采用定性分析与定量分析相结合的方法，往往能取得较好的效果。一般地是先通过组织专家评判，再综合专家的意见，最终形成初始权数。具体操作的步骤如下：

(1)将统计分析目的、已建立的指标体系和第一步确定的指标重要性量化标准分发给各个专家，让专家们根据上述的比例标度值体系所提供的等级重要性系数，请他们独立地给出各指标的重要性权数(权重)。

(2)根据专家们所给出的各个指标的权数，分别计算各指标权重的均值和标准差。

(3)将所得的均值和方差的资料反馈给专家，要求专家再次提出修改意见或更改指标的建议，并在此基础上重新确定权数。

(4)重复以上1～3的步骤，直至获得较为满意的结果(或各专家对各个评价项目所确定的权数趋于一致)为止，便得到初始权数。

3.对初始权数进行处理。步骤如下：

(1)建立判断矩阵 $\boldsymbol{A}$。通过专家对 N 个评价指标的评判，进行两两比较，其初始权数形成判断矩阵 $\boldsymbol{A}$。判断矩阵 $\boldsymbol{A}$ 中第 i 行第 j 列的元素 x_{ij} 表示指标 x_i 与 x_j 比较后所得的标度。

(2)计算判断矩阵 $\boldsymbol{A}$ 每一行各标度连乘积的 N 次方根 W_i。

(3)进行归一化处理。归一化处理就是利用公式 $W_i = W_i / \sum_{i=1}^{n} W_i$ 计算,确定各个评价指标的权数。

4.对判断矩阵的一致性进行检验。判断矩阵的一致性检验是指当需要赋权的指标较多时,矩阵内的初始权数可能出现相互矛盾的现象,对于阶数较高的判断矩阵,难以直观地对其一致性进行判断,这就必须进行一致性检验。本书省略了一致性检验的步骤。

【例 11-5】有三个评价指标 X_1,指标 X_2,指标 X_3,其判断矩阵 A 如表 11-14 所示,试确定这三个指标的权数。

表 11-14 判断矩阵 A

指标	X_1	X_2	X_3
X_1	1	4	1/8
X_2	1/4	1	1/5
X_3	8	5	1

解:

根据已知条件计算 W_i

$W_1 = \sqrt[3]{1 \times 4 \times 1/8} = 0.793\ 7$　　$W_2 = \sqrt[3]{1/4 \times 1 \times 1/5} = 0.368\ 4$

$W_3 = \sqrt[3]{8 \times 5 \times 1} = 3.42$

进行归一化处理:$\sum_{i=1}^{3} W_i = 0.793\ 7 + 0.368\ 4 + 3.42 = 4.582\ 1$

$W'_1 = W_1 / \sum_{i=1}^{3} W_i = 0.793\ 7/4.582\ 1 = 0.173\ 2$

同理,$W'_2 = 0.080\ 4$,$W'_3 = 0.746\ 4$

通过以上计算,初步确定三个指标的权数为:0.173 2,0.080 4,0.746 4。

(四)变异系数法

与前面介绍的几种方法不同,变异系数法不需要依靠专家先对各指标的权重作出评定,而是直接利用各指标所包含的信息通过计算得出指标的权重。它是一种客观赋权方法。这一方法的基本思想是:在评价指标体系中,指标取值差异越大的指标,也就是越难实现的指标。差异越大的指标越重要,因为它更能反映出参加评价的各单位的差距。例如,选取人均 GNP 作为评价现代化的重要指标,不仅在于它能够反映各国经济水平的高低,而且还在于现代化程度不同的国家其人均 GNP 差别较大。试想一下,如果各国人均 GNP 没有差别,都能达到现代化标准,则没有必要再将其作为一项衡量

的指标。

评价指标体系中各指标的量纲不同，不宜直接比较其差异程度。为了消除各指标量纲不同的影响，用各指标的变异系数来衡量各项指标取值的差异程度。

$$\text{各指标的变异系数 } V_i = \frac{\sigma_i}{\bar{x}_i} \quad (i = 1, 2, \cdots, 10) \tag{11.3}$$

式中，σ_i 是第 i 项指标的标准差，$\bar{x}_i$ 是第 i 项指标的均值。

$$\text{各指标的权重 } W_i = \frac{V_i}{\sum_{i=1}^{10} V_i} \tag{11.4}$$

【例 11－6】试利用变异系数法，来确定英国社会学家英克尔斯提出的衡量一个国家现代化程度的指标体系的权重。

解：

选取包括中国在内的近 40 个中等收入水平以上国家 1998 年的数据为样本，根据 11.3 式计算各项指标在这近 40 个国家中的变异系数，以反映各国在各项指标方面的差距，从而作为衡量各项指标权重的依据，再根据 11.4 式计算出各指标的权重，计算结果如表 11－15 所示。

表 11－15　现代化评价指标的权重

	人均 GNP（国际美元）	农业占 GDP 比重(%)	第三产业占 GDP 比重(%)	非农业劳动力比重(%)	城市人口比重(%)	人口自然增长率	平均预期寿命(岁)	成人识字率(%)	大学生占适龄人口比重(%)	千人拥有医生	总和
标准差	7 966.27	7.316	12.94	0.170	19.339	0.831 9	5.375	9.050	20.477	1.314	
平均数	11 938.43	9.352	54.86	0.826	69.792	0.721 4	72.632	93.34	36.556	2.446	
变异系数	0.667	0.782	0.236	0.206	0.277	1.153	0.074	0.097	0.560	0.537	4.59
权数	0.145	0.170	0.051	0.045	0.060	0.251	0.016	0.021	0.122	0.117	1.00

资料来源：根据中国统计出版社《2000 年国际统计年鉴》中所列的近 40 个中等收入以上国家 1998 年数据计算，分别是中国、中国香港、印度尼西亚、伊朗、以色列、日本、哈萨克斯坦、韩国、马来西亚、菲律宾、新加坡、斯里兰卡、泰国、土耳其、埃及、南非、加拿大、墨西哥、美国、阿根廷、巴西、委内瑞拉、保加利亚、白俄罗斯、捷克共和国、法国、德国、意大利、荷兰、波兰、罗马尼亚、俄罗斯、西班牙、英国、乌克兰、澳大利亚、新西兰。其中，大学生占适龄人口比重的数据来源于世界银行《1999 年世界发展指标》。

三、评价结果的综合

在确定评价指标的权数之后，可将经同度量处理后的变量值进行综合汇总。综合汇总有多种方法，本书侧重介绍总分评定法、加权平均综合法和距离综合法。

（一）总分评定法

总分评定法是通过将变量值转化为分值，而后将各个评价对象的分值加总的方法。将变量值转化为分值，可以有多种的记分方法。由于变量的计量尺度不同，对于不同计量层次的变量采用不同的记分方法。

1. 对于以定距尺度或定比尺度计量的评价指标，其记分方式可根据各评价指标的优劣排次，名次在前计高分，名次在后计低分。例如，采用的指标是动态对比指标，以报告期值对比基期值，采用三档记分：报告期对比基期增加者，记满分；报告期与基期持平者，减半记分；报告期比基期下降者计零分。

2. 对于以定序尺度计量的评价项目，一般把该项目划分为若干个评分等级，并对不同等级赋予适当的分值，然后请足够多的评判者按预定的评分等级对该项目评分。

3. 最后综合评判结果，即把各指标（或各项目）得分加总，得到该评价对象的总分。

【例 11－7】考核工业企业经济效益，选择工业增加值率、资本金利税率、成本费用利润率和流动资产周转次数四个指标为评价指标，对上述四个指标确定的标准分分别为 20，40，30，10，四个地区的资料如表 11－16 所示。

表 11－16 四个地区独立核算工业企业四项经济效益指标

指标 / 地区	工业增加值率（%）			资本金利税率（%）			成本费用利润率（%）			流动周转次数（次/年）		
	报告期	基期	报告期/基期	报告期	基期	报告期/基期	报告期	基期	报告期/基期	报告期	基期	报告期/基期
甲	36.32	35.07	103.6	12.92	12.39	104.3	8.86	9.90	89.49	1.51	1.59	95.00
乙	28.38	30.36	93.47	12.91	14.34	90.03	8.88	8.37	106.1	1.44	1.75	82.3
丙	29.14	26.51	109.9	12.53	15.26	82.11	5.91	8.16	72.4	1.84	2.11	87.2
丁	30.29	36.54	82.90	7.45	10.21	72.97	2.66	4.26	62.4	1.35	1.72	78.5

评分标准：动态指标大于 1 者记满分；动态指标等于 1 者减半记分；动态指标小于 1 者记 0 分。

解：

四个地区的得分情况如下：

甲:20+40+0+0=60

乙:0+0+30+0=30

丙:20+0+0+0=20

丁:0+0+0+0=0

上面计算结果表明,甲地区企业经济效益最佳,乙、丙次之,丁地区最差。

(二)加权平均综合法

加权平均综合法有两种形式,一种是加权算术平均形式,另一种是加权几何平均形式。

1.采用加权算术平均形式计算,其基本公式如下:

$$加权算术平均综合值=\frac{\sum_{i=1}^{n}x_iW_i}{\sum_{i=1}^{n}W_i} \tag{11.5}$$

上式中,给定评价指标体系由 n 个指标构成,x_i 为已经经同度量处理过的第 i 个评价指标的相对值,$i=1,2,\cdots,n$,W_i 为各项指标的权重,$W_1+W_2+\cdots+W_n=100\%$(或=1)。

【例11-8】假定企业效益评价指标体系中的评价指标:工业增加值率、总资产贡献率、流动资产周转次数、工业成本费用利润率、全员劳动生产率和产品销售率的权重向量为(0.15,0.2,0.1,0.15,0.2,0.2),对表11-7功效系数法处理结果的资料进行加权算术平均,计算各直辖市的加权算术平均综合值,并按独立核算企业经济效益的高低对各直辖市进行排序。

解:

北京市企业经济效益功效系数加权算术平均综合值:

$$\frac{66.4\times0.15+69.4\times0.2+75.3\times0.1+75.1\times0.15+82.6\times0.2+67.6\times0.2}{0.15+0.2+0.1+0.15+0.2+0.2}=72.675$$

同理,可得天津市、上海市和重庆市的功效系数综合值,它们分别为:80.755,85.625,67.77。

按企业经济效益功效系数综合值的高低排序:上海市第一,天津市次之,北京市第三,重庆市最差。

2.采用加权几何平均形式计算,其基本公式如下:

$$加权几何平均值=\sqrt[\sum_{i=1}^{n}W_i]{\prod x_i^{W_i}} \tag{11.6}$$

上式中,评价指标体系的容量为 n,x_i 为已经经同度量处理过的第 i 个评价指标的相对值,$i=1,2,\cdots,n$,W_i 为各项指标的权重。

【例11-9】以表11-7功效系数法处理结果的资料和例11-8给出的权数,采用

加权几何平均法，计算各直辖市的功效系数加权几何平均综合值。

北京市企业效益功效系数加权几何平均综合值：

$$\sqrt[0.15+0.2+0.1+0.15+0.2+0.2]{66.4^{0.15}\times 69.4^{0.2}\times 75.3^{0.1}\times 75.1^{0.15}\times 82.6^{0.2}\times 67.6^{0.2}}=72.34$$

同理，我们可得到天津、上海和重庆等直辖市的企业经济效益功效系数加权几何平均综合值，它们分别是：79.68，85.28，66.04。按几何平均法的计算结果排序：上海市第一，天津市次之，北京市第三，重庆市最差。

（三）最优值距离法

最优值距离法是以各项评价指标的实际值与最优值（或标准值）之间距离的大小来衡量评价对象效益的高低。对评价指标体系中的各评价指标进行相对化处理后，运用距离法进行综合，其综合值反映了各评价指标与相应的最优值（或标准值之间的距离）。第 j 个参评单位的距离综合值的计算公式如下：

$$S_j=\sqrt{(1-X'_{1j})^2W_1+(1-X'_{2j})^2W_2+\cdots+(1-X'_{nj})^2W_n} \tag{11.7}$$

上式中，X'_{ij} 是经相对化处理后的变量值，$X'_{ij}=X_{ij}/X_{i0}$，i 表示第 i 个评价指标，$i=1,2,3,\cdots,n$；j 表示第 j 个参评单位，$j=1,2,3,\cdots,m$。X_{ij} 为第 j 个参评单位的第 i 个评价指标的实际值，X_{i0} 为第 i 项指标的最优值或标准值，它们可以是全国的先进水平或历史的最高水平，也可以根据经济发展规划而确定。权数 W_i 是根据各项评价指标在社会再生产过程中的重要程度分别确定相应的权数。

S_j 值越小，与最优值越接近，表明效益越佳；S_j 值越大，越远离最优值，表明效益越低。

【例 11－10】根据表 11－5 的资料，各项指标：工业增加值率、总资产贡献率、流动资产周转次数、工业成本费用利润率、全员劳动生产率和产品销售率的最优值分别为：50%，10%，1.8 次/年，10%，90 000 元/人年，100%，各项指标相应的权重向量为（0.15，0.2，0.1，0.15，0.2，0.2），试计算各直辖市的距离综合值。

解：

先对各个指标实际值进行相对化处理，相对化处理是以各评价指标的实际值对比最优值，具体计算如下：

北京市企业增加值率相对值：28.17/50＝0.56

北京市企业总资产贡献率相对值：6.94/10＝0.69

同理，我们可以计算其他指标的相对值。经相对化处理后的数据如表 11－17 所示。

北京市最优距离综合值：

$$S_1=\sqrt{(1-0.56)^2\times 0.15+(1-0.69)^2\times 0.2+(1-0.79)^2\times 0.1+\cdots+(1-0.96)^2\times 0.2}=0.327\,6$$

表 11－17　相对化处理结果表

指标 / 直辖市	工业增加值率	总资产贡献率	流动资产周转次数	工业成本费用利润率	全员劳动生产率	产品销售率
最优值	50.00	10.00	1.80	10.00	90 000	100.00
北　京	0.56	0.69	0.79	0.50	0.71	0.96
天　津	0.48	0.85	0.99	0.65	0.58	0.98
上　海	0.54	0.89	0.87	0.64	0.91	0.99
重　庆	0.59	0.61	0.66	0.17	0.35	0.99

同理，我们可以得到天津、上海和重庆三个直辖市的最优距离综合值，它们分别是：0.3143，0.2434，0.5050。根据各个城市的企业经济效益的最优距离综合值排序，上海居首位，天津次之，北京第三，重庆第四。

本章小结

1. 综合评价就是根据研究的目的，以统计资料为依据，借助一定的方法，对不能直接加总的、性质不同的项目进行综合，得出概括性的结论，从而揭示事物的本质及其发展规律。

2. 系统聚类法是通过判断指标之间的相似程度来筛选指标的一种方法。系统聚类法的具体操作步骤是：第一步，度量指标(类)间的相似程度；第二步，度量指标(类)间的距离；第三步，根据聚类图确定指标(类)的个数，从每类中选择出最具有代表性的指标。

3. 同度量处理主要有二种方法：相对化处理法要先对每个评价指标确定一个标准值。指标有“正”、“逆”之分，对于正指标，用公式 $x_i' = x_i / x_m$ 进行相对化处理；对于逆指标，用公式 $x_i' = x_m / x_i$ 进行相对化处理。

功效系数法是借助功效系数，把须评价的各项指标值转化为可以度量的评判分数。功效系数的计算公式是：$d_i^* = (x_i - x_i^{(s)}) / (x_i^{(h)} - x_i^{(s)}) \times 40 + 60$。

标准化处理法是将变量值转化为数学期望为 0、方差为 1 的标准化数值。具体操作步骤如下：第一步，求出各变量(指标)的算术平均值(数学期望值) $\bar{x}_i$ 和标准差 σ_i；第二步，利用公式 $x_{ij}' = (x_{ij} - x_i) / \sigma_i$ 进行标准化处理。

4. 权数的确定方法主要有四种：统计平均法、最大组中值法、AHP 构权法和变异系

数法，前三种属于主观赋权法，后一种属于客观赋权法。

5.对单项评价指标评价结果的合成，一是采用总分评定法，它先将变量值(指标实际值)转化为分值，而后将各个评价项目的分值加总；二是采用加权平均综合法，加权平均综合法有两种形式，一种是加权算术平均形式，另一种是加权几何平均形式；三是采用最优值距离法，它是以各项评价指标的实际值与最优值(或标准值)之间距离的大小来衡量评价对象效益的高低。

思考与练习题

一、判断题：

1.综合评价是对客观对象从不同的侧面、不同的角度进行全面的分析与评价，故它完全消除了主观因素的影响。(　　)

2.资产负债率、万元产值消耗能源比率和物耗率皆为逆指标。(　　)

3.评价指标值若服从正态分布，可利用标准化方法进行消除量纲影响的处理。(　　)

4.进行归一化处理后，权数之和一定为1。(　　)

5.经过标准化处理后的数据 x'_{ij} 取值范围 $0\leqslant x'_{ij}\leqslant 1$。(　　)

二、问答题：

1.现有7个指标，根据历史资料计算每两个指标的相关系数，建立如下相关系数矩阵，见表11－18。

表11－18　指标间相关系数

指标	1	2	3	4	5	6	7
1	1	0.85	0.58	0.54	0.44	0.74	0.60
2		1	0.55	0.56	0.56	0.34	0.62
3			1	0.89	0.73	0.72	0.60
4				1	0.74	0.41	0.72
5					1	0.54	0.80
6						1	0.66
7							1

要求：

根据上述资料绘制聚类图。假定评价某一项目的评价指标体系指标容量为4，请根据聚类图从中选择4个评价指标。

2.某电器公司为了解本公司×型号电冰箱的受欢迎程度,请客户对该型号冰箱从其制冷量、耗电量、耐用度及售后服务四个项目进行评价。对每一评价项目,按满意10分、较满意8分、一般6分、较差4分四个等级计分。现对收回的500个客户的调查卷的资料整理结果如表11-19所示。

表11-19 四个评价项目的得分情况表

评估项目	得票数				
	满意(10分)	较满意(8分)	一般(6分)	较差(4分)	合计
制冷量	280	160	60	0	500
耗电量	120	200	100	80	500
耐用度	160	160	100	80	500
售后服务	50	50	100	300	500

又假定各评价项目的权重分别为:0.3,0.2,0.3,0.2。

要求:

按总分评定法计算×型号电冰箱的得分。

3.污水处理厂的经营管理状况综合评判是采用如下指标反映的:①每日污水处理量(千吨/日);②BOD_5去除率(%);③悬浮物去除率(%);④处理1吨污水消耗空气量(米^3/吨);⑤去除1公斤BOD_5耗电量(度/公斤)。以上各项指标的权数为:0.10,0.30,0.30,0.15,0.15。又已知某污水处理厂按月记录的历史数据划分为三个等级,各等级的平均值和该厂2002年6月的实际指标值如表11-20所示。

表11-20 污水厂主要技术管理指标

指标	单位	好	中	差	实际值
每天污水处理量	千吨/日	18.5	16.5	14.5	15.0
BOD_5去除率	%	95.0	87.0	79.0	95.0
悬浮物去除率	%	95.0	87.0	79.0	94.0
处理1吨污水消耗空气量	米^3/吨	6.5	8.5	10.5	8.0
去除1公斤BOD_5耗电量	度/公斤	0.85	1.05	1.25	1.05

要求:

(1)以评语"好"为满意值,以评语"差"为不容许值,用功效系数法对上述资料进行同度量处理。

(2)以评语“好”为最优值,用相对化处理法对上述资料进行同度量处理。

(3)对(1)计算结果的数据,分别采用加权算术平均法和加权几何平均法进行综合。

(4)对(2)计算结果的数据,采用最优值距离法进行综合。

(5)对三种综合方法的计算结果进行比较。

第十二章

国民经济统计概述

★ 本章学习要点与要求 ★

国民经济核算是经济统计的核心内容，本章扼要介绍国民经济核算体系基本框架、国内生产总值等最主要的国民经济总量指标的计算，以及基本的宏观经济统计分析方法。通过本章学习，要求：①了解国民经济核算体系的基本框架；②了解国民经济核算的主要分类；③理解确定国民经济核算主体范围的基本原则；④掌握国内生产总值及相关总量指标的概念、计算方法及其数量关系；⑤掌握有关统计分析指标的含义，能够对有关宏观经济问题进行分析。

第一节　国民经济核算的基本问题

一、国民经济核算体系的发展

国民经济核算是国民收入统计演化的结果。最早的国民收入统计可以追溯到1665年英国经济学威廉·配第在他的《政治算术》一书中对当时英国国民收入的估算。1929～1933年资本主义世界爆发了严重的经济大危机，主张“国家干预”的凯恩斯经济理论应运而生，作为政府管理经济基本依据的国民经济核算体系便顺应这一历史时期的要求而诞生了。

英国、美国等国家的政府和联合国等国际组织对国民经济核算体系的创建与发展

作出了重要贡献。1984年诺贝尔经济学奖得主英国经济学家、剑桥大学教授斯通领导了联合国国民经济核算体系研究和统计制度的设计工作。1971年诺贝尔经济学奖得主美国经济学家西蒙·库兹涅茨以美国国民经济研究局(NBER)的工作为基础创立和发展了美国国民经济核算理论、方法和实际统计工作。联合国1947年公布了《国民收入的计量和社会核算表的编制》,1953年公布了《国民核算表及补充表体系》(简称旧SNA,A System of National Accounting),标志着国民经济核算体系正式诞生。在世界各国实践和理论研究的基础上,1968年联合国又颁布了《国民经济核算体系》(简称新SNA),意味着国民经济核算体系的发展成熟。1993年联合国第27届统计委员会议又通过了对这一体系的修改方案,其内容有了较大的扩充,形成的核算体系被称为1993SNA。联合国的国民经济核算理论与方法,对世界各国国民经济核算体系的建立与发展起了根本性的指导作用。目前已有170多个国家和地区采用了这一体系。

我国在改革开放之前,采用原苏联的国民经济核算体系,即物质产品平衡表体系(Material Product Balance System 简称MPS)。MPS只对物质生产部门的再生产成果进行核算,而且核算内容也不完整。随着改革开放的深入发展,为使我国国民经济核算制度适应国民经济管理和宏观经济调控的需要,1984年国务院成立专门的国民经济统一核算领导小组,参照SNA对我国的新国民经济核算体系进行设计,经过若干年的努力,制定和颁布了我国的《新国民经济核算体系(试行方案)》。从1992年起我国开始向新的国民核算体系转轨,到1995年,基本完成了向新国民经济核算体系的全面过渡。

二、国民经济核算的对象和内容

国民经济核算是以国民经济为对象的整体核算。国民经济核算体系通过采用一整套的统计表、统计指标、账户体系、平衡表、数学模型对国民经济运行过程作系统描述和反映。为了实现对国民经济运行过程的系统描述,现代国民经济核算体系已经发展成由国内生产总值(Gross Domestic Product, 简称GDP)的生产及使用核算、投入产出核算、资金流量核算、国际收支核算和资产负债核算五个有机联系的子系统(五大核算)组成的科学体系。五大核算的具体内容如下:

(一)GDP生产及使用核算

GDP生产及使用核算是以GDP为核心指标,对国民经济活动中的生产、消费、投资和进出口等基本经济总量进行全面、系统的核算,综合反映国民经济发展的规模、速度和结构。

(二)投入产出核算

投入产出核算是对GDP核算的延伸,它运用投入产出分析法,将国民经济各部门生产中的各种投入品来源与产品去向,反映在纵横交错的棋盘式平衡表中。投入产出

核算详细地反映国民经济各产品之间、部门之间互相依存的技术经济联系。

(三)资金流量核算

资金流量核算是以全社会资金运动为对象,核算各部门资金的来源和运用,是国民经济核算在分配领域的展开,反映全社会各种资金在各部门间的流量、流向,包括各部门收入的形成,初次分配、再分配和各部门资金余缺的调节和融通。

(四)国际收支核算

国际收支核算是以对外经济交易为对象,对一定时期一国与其他国家或地区之间的贸易、非贸易往来和资本往来进行系统的核算,综合反映本国的国际收支平衡状况、收支结构和外汇储备状况。

(五)资产负债核算

资产负债核算是反映一个国家在一定时点上的国民经济资产和负债总量、结构,为研究国民财产状况和经济实力以制定产业发展政策和调整产业结构提供依据。

为了适应可持续发展的需要,1993 年的 SNA 又增加了附属的环境核算子系统。

国民经济核算体系内容庞大复杂。目前,我国的新国民经济核算体系有两大组成部分:第一部分为社会再生产核算表,其主体部分是国民经济五大核算基本表,此外,还有八张补充表;第二部分为经济循环账户,包括一套国民经济账户体系,基本框架如图 12-1 所示。

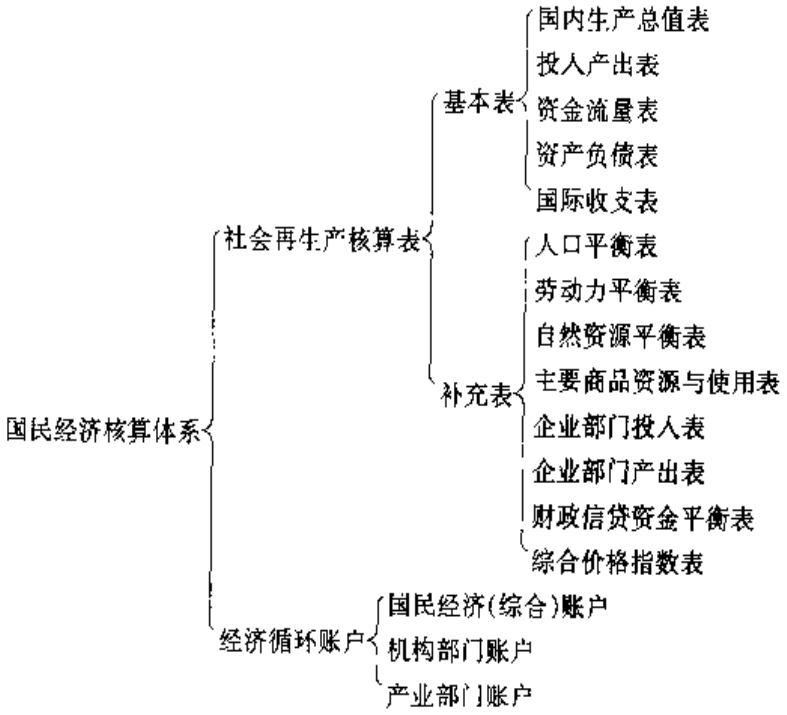


图 12-1 我国新国民经济核算体系的基本框架

三、国民经济核算的行业与部门分类

国民经济是一个复杂的有机整体，包括极其众多的组成部分，需要分门别类地加以研究。搞好国民经济分类，是国民经济核算的一项十分重要的基础性工作，也是开展国民经济深入系统的研究和分析的基本条件。

国民经济分类具有多层次、相对性和不断演化的特点。一方面根据考察对象和研究目的的不同，可以选择不同标志，从不同角度对国民经济进行分类。另一方面，国民经济本身的构成也是随着社会生产力和社会分工的发展而发展，随着社会生产关系和社会经济体制的变化而变化的。例如，由于经济技术的迅速发展，不断促成一次又一次社会分工，出现新的活动领域，产生新的国民经济部门。历史上第一次社会大分工，促成农业（种植业）与原始畜牧业（采集渔猎）的分离，第二次社会大分工又使手工业从农业中分离出来，第三次社会大分工则产生了专门从事商品流通等活动的领域，从而在国民经济中逐渐形成了第一次产业、第二次产业和第三次产业。主要的国民经济核算行业与部门分类有以下几种：

（一）机构部门分类和活动部门分类

机构部门和活动部门分类是从不同角度对国民经济的双重交叉分类，这两种分类有不同的功能，可以满足国民经济核算不同方面的要求和对国民经济活动不同的分析需要。

1. 机构部门分类。机构部门分类是对国民经济各常住单位，按其取得收入和支配收入、筹集资金和运用资金的财务决策权同一性的标准进行分类的一种方法。分类的基本单位是能够拥有资产，承担负债，从事经济活动并与其他单位进行经济交易的实体。国民经济中的机构部门和机构子部门由常住机构单位构成。一个单位如果只有生产决策权而没有财务决策权，就只能作为经济活动部门分类的基本单位，不能作为机构部门分类的基本单位。

在SNA的分类中，有五个机构部门：

(1)非金融公司部门，包括主要从事市场货物和非金融服务生产的机构单位，即一般企业部门。

(2)金融公司部门，包括主要从事金融中介活动或辅助金融活动的机构单位。

(3)一般政府，包括除负有政治责任和经济调节作用外，主要提供用于个人和公共消费的非市场服务，并对收入和财富进行再分配的机构单位。

(4)为住户服务的非营利机构部门，包括主要为住户从事非市场服务的法律实体。

(5)住户，包括经济体中的所有自然人。

我国国民经济核算体系的常住机构单位与联合国的分类方法有一些不同，分为四个机构部门：

(1)非金融企业部门。指主要从事各种非金融生产经营活动的独立核算企业所组成的部门。包括除金融服务企业之外的国有、集体、三资工商企业、建筑企业、运输邮电企业以及其他各种服务企业。

(2)金融机构部门。指主要从事金融中介活动的部门。包括从事资金借贷、保险业务、证券投资活动的银行、信用社、信托投资公司、保险公司、证券公司、租赁公司等各类金融机构。

(3)政府部门。指主要从事国家管理活动的中央政府行政机关、地方政府行政机关、军队、警察等。此外还包括由国家财政差额预算拨款的非营利事业单位以及群众团体，如医院、学校、广播电视、科研机构和一些社团组织等。

(4)住户部门。包括以消费者身份出现的城乡居民户和从事各种生产经营活动的个体经营户。

2.活动部门分类。活动部门分类是从生产的角度，按各基层单位活动性质的同类性对常住单位的分类。一个企业，特别是一个大企业，可能同时从事不同的生产活动，如果按主要活动对企业进行分类，就会产生某些企业生产的货物和服务活动不同质的情况，不利于对生产过程的技术经济联系进行分析。所以，在对生产技术起重要作用的生产过程进行分析时，必须排除不同的生产活动，而将性质相同的生产活动归并在一起。按活动部门分类，也就是对掌握生产经营决策权的基层单位进行分类。同一类性质活动的基层单位归并在一起，就形成产业部门或行业部门。但这种分类存在层次差别，根据分析的需要，可粗可细。

(二)三次产业分类

三次产业分类是英国经济学家阿·费希尔(A. Fisher)于20世纪30年代提出的。如上所述，三次产业是就产业形成的时序加以划分，同时也反映了劳动对象的特点及其满足人类需求的层次。人类最早形成的第一产业是以自然资源为对象的产业，包括农业和采掘业；以后相继出现了以农产品和采掘品为对象进行加工和再加工的第二产业，包括加工工业(制造业)和建筑业；第三产业提供各种服务产品，包括除第一、第二产业以外的所有行业。

三次产业分类，在国际上已经通用。但各国的分类标准并不完全相同。比如，美国把采矿业列入第一产业，日本等国则把采矿业列入第二产业。有的国家把煤气、电力和自来水划入第二产业，也有的国家将其划入第三产业。第三产业的内容非常丰富，并且正在随着人类经济技术的发展而不断发展。改革开放以前，我国没有采用三次产业分类。随着改革开放的不断深化，社会生产力的发展和提高，我国第三产业得到了迅速发

展，国家统计局于1985年提出并报经国务院批准《关于建立第三产业统计的报告》，对三次产业作了如下划分：

第一产业：农业（包括林业、牧业、渔业等）；

第二产业：工业（包括采掘业、制造业、自来水、电力、蒸汽热水、煤气）和建筑业；

第三产业：除上述第一、第二产业以外的其他各业。

由于第三产业包括的行业多，范围广，根据我国的实际情况，第三产业又分为两大部门：一是流通部门，二是服务部门。具体又分为四个层次：

第一层次：流通部门，包括交通运输业、邮电通讯业、商业、饮食业、物资供销和仓储业。

第二层次：为生产和生活服务的部门，包括金融业、保险业、地质普查业、房地产业、公用事业、居民服务业、旅游业、咨询服务业和各类技术服务业等。

第三层次：为提高科学文化水平和居民素质服务的部门。包括教育、文化、广播电视业、科学研究事业、卫生、体育和社会福利事业等。

第四层次：为社会公共需要服务的部门。包括国家机关、政党机关、社会团体及军队和警察等。

（三）国民经济行业分类

行业分类是一个相对完备的国民经济分类体系，是最为基本、也是最为重要的国民经济分类。与三次产业分类相比，国民经济行业分类更为系统、完整，并构成三次产业和其他一些重要的国民经济分类的基础。世界各国和联合国等有关国际组织都制定了专门的国民经济行业分类标准。

行业分类具有多层次的结构，便于灵活地进行相应的分解和归并处理，满足国民经济管理和核算的各种需要。国民经济行业的分类对象是从事各种生产经营活动的所有经济主体的总和。其分类标志则是它们所从事的生产经营活动的经济性质。

联合国1989年颁布的《国际标准产业分类》将国民经济分为17个主要部门，60个大类，173个中类，301个小类。17个主要部门是：①农业、狩猎业、林业；②渔业；③矿业和采矿业；④制造业；⑤电力、煤气、自来水供应；⑥建筑业；⑦批发和零售商业，汽车、摩托车修理及个人家庭财产修理业；⑧饭店旅馆业；⑨运输、仓储、邮电通讯业；⑩金融、保险业；⑪房地产业和租赁业；⑫公共行政、国防、社会治安部门；⑬教育；⑭卫生；⑮社会团体、其他社会和个人服务；⑯受雇佣私人家庭服务；⑰跨地区的团体和实体。

我国的国民经济行业分类曾经作过多次修订。在1978年修订时，将国民经济划分

为10大部门、47个大类、56个中类、149个小类。1985年修改颁发的《国民经济行业分类和代码》实行四级分类，将国民经济分为13个门类、25个大类、310个中类、667个小类。1993年统计年报新的国民经济行业分类标准将国民经济分为16个门类，即：①农、林、牧、渔业；②采掘业；③制造业；④电力、煤气及自来水的生产和供应业；⑤建筑业；⑥地质勘查业、水利管理业；⑦交通运输、仓储及邮电通讯业；⑧批发和零售贸易、餐饮业；⑨金融、保险业；⑩房地产业；⑪社会服务业；⑫卫生、体育和社会福利业；⑬教育、文化艺术及广播电视业；⑭科学研究和综合技术服务业；⑮国家机关、政党机关和社会团体；⑯其他行业。

四、国民经济核算的主体范围

国民经济核算必须首先界定空间界限范围，目前国际上通行的界定原则是"常住经济单位原则"。所谓常住经济单位(简称常住单位)，是指在一国的经济领土上具有经济利益中心的经济单位。理解这个定义需要搞清楚"经济领土"和"经济利益中心"这两个概念。"经济领土"与人们通常所讲的主权领土不完全一致，它除了包括本国大陆的领地、领海和领空外还包括以下两部分：一是位于国际水域，但本国具有捕捞和海底开采管辖权的大陆架；二是本国在其他国家领土范围内的所谓领土飞地，即位于别国境内但实际为本国政府所拥有或租借的土地，如由于外交、科研和国际援助等目的，通过正式协议获得，并具有明确边界线的区域范围。相应的，在本国领土范围内而被其他国家和经济组织租借和实际使用的领土飞地，就不能记入本国的经济领土范围内。"经济利益中心"是指一个经济活动单位在本国的经济领土范围内具有一定的生产经营活动场所，如住房、厂房或其他建筑物，从事一定规模的经济活动并在时间上超过一年。与常住经济单位相对应的概念是非常住经济单位。所谓非常住经济单位是指那些不在本国经济领土上，或在本国经济领土上没有经济利益中心的活动单位。

就实际情况而言，本国境内的外商投资企业是本国的常住经济单位，而本国在境外投资的企业则不是本国的常住经济单位；本国的驻外使领馆和新闻机构是本国的常住经济单位，而外国驻本国的使领馆和新闻机构则是非常住经济单位。常住经济单位在主权领土以外的生产活动仍然属于本国民经济总体的生产活动，属于本国国民经济核算的主体范围，而非常住经济单位在本国主权领土之内的生产活动则不属于本国民经济总体的生产活动，也不属于本国国民经济核算的主体范围。

第二节 国内生产总值核算

一、国内生产总值的概念

GDP是对一个国家或一个地区在一定时期内常住单位活动的总成果的一种计量。GDP作为对全社会所有经济部门生产活动的产品总量的计量，从总体上反映了一个国家或地区经济活动的总规模、综合实力。

1.从实物运动角度看，在国民经济中，生产活动的成果可以分为两类：一类叫做中间产品，另一类叫做最终产品。中间产品就是本期生产出来，并在本期再次投入生产活动，作为原材料、燃料、动力、辅助材料的产品，它是构成企业本期生产活动的成本的重要内容。最终产品则是指本期生产，本期不再加工，可供社会最终消费、使用或者库存的产品。例如，纺纱厂用某农场本期生产的棉花作为原料生产出棉纱，棉花就是中间产品。而纺织厂以棉纱为原料生产坯布，棉纱也是中间产品。染织厂以坯布为原料生产色布，坯布也是中间产品。最后，服装厂以色布为原料生产服装，色布也成了中间产品，而只有服装最终卖给消费者使用，才是最终产品。国民经济生产活动的目的是为社会和居民提供可供最终消费和使用的产品，中间产品只是在国民经济再生产过程中循环，为最终产品的生产而出现的过渡性产品，所以只有社会最终产品才是国民经济生产活动的最终有效成果。GDP反映的就是一国的经济常住单位为社会提供的全部最终产品的价值总量。

就其实物内容来看，最终产品的去向有三个方面：①居民和政府消费的消费品和服务产品；②用作投资的产品；③净出口产品。

2.从价值的角度看，在国民经济生产活动中的价值也可以分为两种类型：一种是中间消耗价值，一种是增加值。中间消耗(C_2)就是在生产过程中所消耗的产品的价值，它是上一个生产环节的活动成果；增加值则是在生产过程中把中间产品加工成最终产品所追加的价值。GDP反映的就是一国的各常住单位所创造的增加值的总和。

就其价值构成来看，增加值包括三部分：①固定资产折旧(C_1)，即固定资产的磨损价值；②劳动者报酬(V)；③社会纯收入(M)，包括国家生产税净额收入和企业营业盈余。

从全社会的角度看，中间消耗是中间产品的价值表现；增加值是最终产品的价值表现。全社会的增加值总和等于最终产品的价值。这个关系可通过表12－1说明。表

12－1 中的最终产品是成衣，它的价值为 A＋B＋C＋D＋E，成衣所经过的农场、纺纱厂、织布厂、印染厂、服装厂的增加值分别为 A，B，C，D，E，它们的总和为 A＋B＋C＋D＋E，与最终产品成衣的价值完全相等。

表 12－1　中间消耗与最终产品的关系

生产单位	产品	中间消耗 C_2	增加值 C_1+V+M	总产值 C_2+C_1+V+M
农　场	棉花	0	A	A
纺纱厂	棉纱	A	B	A＋B
织布厂	坯布	A＋B	C	A＋B＋C
印染厂	色布	A＋B＋C	D	A＋B＋C＋D
服装厂	成衣	A＋B＋C＋D	E	A＋B＋C＋D＋E
合计			A＋B＋C＋D＋E	

二、国内生产总值核算的原理和方法

从理论上看，GDP 的生产量、分配量和使用量是完全相等的。这在国民经济核算中被称为 GDP 统计的“三方等价原理”。国民经济生产的目的就是为了社会的最终使用，生产多少就使用多少，而生产的成果在使用前首先要通过货币形式对各生产要素按其对生产的贡献进行分配。就总量而言，生产多少就分配多少，生产的总量等于收入分配的总量，也等于支出使用的总量。从表 12－1 也可以很清楚地看到这一平衡关系。生产的最终产品成衣的价值量为 A＋B＋C＋D＋E；供社会最终使用的成衣的价值是 A＋B＋C＋D＋E；而社会分配的价值，各生产环节固定资产折旧、劳动者报酬和社会纯收入的增加值（C_1+V+M）总和也是 A＋B＋C＋D＋E。

根据三方等价原理，可以分别从生产、分配和使用三种不同的角度对 GDP 进行核算，即生产法、收入法和支出法。

（一）生产法

生产法是从生产的角度以生产的总成果，即货物和服务的总产品价值为出发点，从中扣除生产过程中投入的中间产品的价值，得到增加值。各企事业、各行业的增加值之和便是 GDP。用生产法核算 GDP 的公式是：

$$增加值=总产值-中间消耗 \tag{12.1}$$

这里的总产值（也称总产出）是指一个国家（或地区）的常住单位在一定时期内生产或提供的所有货物和服务产品的总价值。它反映了常住单位的生产规模。中间消耗则是指常住单位在生产货物或提供服务产品的过程中，因生产所需消耗和使用的所有非

固定资产形式的货物和服务的价值。从总产值中扣除中间消耗,剩下的便是常住单位在本期生产过程中创造的新增价值和固定资产的转移价值,即增加值。

例如,某工业企业报告期生产了价值1 000万元的产品,是该企业的总产值。企业在生产过程中消耗了700万元的原材料、燃料、动力、辅助材料等中间产品,扣除这700万元,剩下的300万元是这个企业创造的增加值。

GDP等于各企事业、行政单位创造的增加值之和。

(二)收入法

收入法是从生产领域进行初次分配的角度,也是从生产要素在生产领域得到初次分配收入的角度来进行计算的,也称分配法。

收入法的计算方法是:从生产过程创造收入的角度,对常住单位增加值构成项目进行核算。常住单位增加值的计算公式为:

$$增加值=劳动者报酬+生产税净额+固定资产折旧+营业盈余 \tag{12.2}$$

等式右边的四个项目,分别反映了参与GDP初次分配的劳动力、政府、固定资产和企业经营者四方各自在GDP初次分配中的所得份额。这四个项目的具体内容是:

1.劳动者报酬。指劳动者因从事生产活动所获得的全部报酬,包括各种形式的工资、奖金和津贴。既包括货币形式的,也包括实物形式的,还包括劳动者所得到的上班交通补贴和由单位支付的社会保险费等。

2.生产税净额。指各种生产单位向政府缴纳的生产税减去政府对企业的生产补贴以后的净值。生产税是政府对生产单位从事生产、销售和经营活动以及因为从事生产经营活动而使用某些生产要素,例如固定资产、土地、劳动力所征收的各种税、附加费和规费。生产税具体包括销售税及附加、增值税、应缴纳的养路费、排污费和水费附加、烟酒专卖上缴政府的专项收入等。生产补贴与生产税相反,是政府对生产单位政策性的单方面收入转移,故而作为生产税的对立面进行扣减处理。生产补贴具体包括政策性亏损补贴、外贸企业出口退税等。

3.固定资产折旧。指在一定时期内,为弥补固定资产损耗,按照核定的固定资产折旧率计提的固定资产折旧。其中,政府单位、非企业化管理的事业单位和居民自有住房等实际中是不计提固定资产折旧的,而是按国民经济核算中统一规定的折旧率虚拟计算固定资产折旧。固定资产折旧反映了全社会固定资产在本期生产中磨损的价值量。

4.营业盈余。指常住单位创造的增加值扣除劳动者报酬、生产税净额和固定资产折旧后的余额。它相当于企业的经营利润加上生产补贴,但要扣除从利润中开支的工资和福利。

各常住单位的增加值的总和就是GDP。

（三）支出法

支出法也称使用法。是从社会产品最终使用去向，也就是购买支出的角度来计算GDP。

支出法的计算公式为：

$$GDP = 最终消费支出 + 资本形成总额 + 货物和服务的净出口额 \quad (12.3)$$

等式右边的三个项目反映了GDP使用的三个去向，具体包括的内容是：

1.最终消费支出。最终消费支出分居民消费支出和政府消费支出两部分。指本国的常住单位为满足物质、文化和精神生活的需要，所购买的货物和服务的支出。

2.资本形成总额。指常住单位在一定时期内对固定资产和存货的投资支出合计，包括固定资本形成总额和存货增加两部分。

3.货物和服务的净出口额。指货物和服务出口额减进口额的差额

【例12－1】 某地区有关国民经济核算资料如下：

1.总产值与中间消耗如表12－2所示。

表12－2

部　门	总产值(万元)	中间消耗率(%)
农业	5 450	36
工业	22 240	72
建筑业	3 058	70
运输业	650	22
商业	1 615	48
服务企业	1 606	27
政府服务行政事业	1 608	28

2.分配再分配如表12－3所示。

表12－3　　单位：万元

项　目	数　额
劳动报酬	8 520
生产税净额	2 370
所得税	360
营业盈余	2 297
固定资产折旧	980

3.支出使用情况如表12－4所示。

表 12－4 单位：百元

项 目	数 额
居民个人消费	7 080
政府消费	1 820
社会总投资	5 107
固定资产投资	4 700
库存增加	407
净流出(流出－流入)①	160

试用生产法、收入法和支出法计算国内生产总值。

解：

(1)按生产法计算：

国内生产总值＝总产值－中间消耗

＝5 450(1－36%)＋22 240(1－72%)＋3 058(1－70%)＋650(1－22%)＋1 615(1－48%)＋1 606(1－27%)＋1 608(1－28%)

＝14 309.54(万元)

(2)按收入法计算：

国内生产总值＝劳动者报酬＋生产税净额＋固定资产折旧＋营业盈余

＝8 520＋2 370＋2 297＋980

＝14 167(万元)

(3)按支出法计算：

国内生产总值＝最终消费支出＋资本形成总额＋货物和服务的净流出额

＝(7 080＋1 820)＋5 107＋160

＝14 167(万元)

从理论上说按上述三种方法核算出来的 GDP，应该是完全相等的。但在国民经济核算的实践中，由于受资料来源、统计口径等因素的限制，三种方法的计算结果往往不相等。特别是支出法所得出的 GDP 数值与生产法、收入法的核算结果之间经常存在一定的出入。在我国国民经济核算实践中，生产法和收入法所计算的 GDP 数值相等。这是因为，生产法和收入法都是对各个产业部门增加值的计算。支出法核算的 GDP 则与

① 对于地区来讲流入和流出不仅包括对外国的进出口，还包括从国内其他地区流入或流出到其他地区的货物与服务。

之存在一定的统计误差。根据《2000年中国统计年鉴》公布的数据，我国1999年按生产法和收入法计算的GDP是81 910.9亿元，而按支出法计算的GDP则是82 429.7亿元，两者的统计误差为518.8亿元。

三、国内生产总值核算的价格问题

GDP反映了经济发展的规模和水平，其实质内容是国民经济活动为社会提供的货物和服务产品，而不是抽象的价值符号。人类社会从事经济活动的目的就是希望获得更多更好的物质产品和服务来满足消费需求。但是，由于社会产品服务种类繁多、千差万别，对其总量的核算不可能采用实物量指标，只能采用价值量指标。由于以价值量表现的GDP的数值大小由两方面的因素决定，一是所生产的实物产品与服务产品数量；二是产品和服务的价格水平。这样，如果国民经济价格水平发生变动，不同时期按当年价格计算的GDP数量变化中，就包含了价格变动因素，因而不能准确反映实际生产水平的变化。所以用GDP来反映经济发展水平时，必须注意其价格问题。

如果各个时期的GDP都是采用当年价格计算，称为"名义的GDP"。如果消除了价格变动因素，各个时期的GDP都采用固定价格、可比价格或某个特定时期的价格来计算，则GDP的数值变化就不包括价格因素的变化，称为"实际的GDP"。

根据统计指数理论，通过名义的GDP除以实际的GDP，可以得到"GDP紧缩价格指数"（又称"GDP平减指数"或"综合价格换算系数"）。如果事先编制了GDP紧缩价格指数，然后对名义的GDP数值进行缩减计算，就可以求得实际的GDP。用公式表示如下：

$$\text{实际 GDP} = \frac{\text{名义 GDP}}{\text{GDP 紧缩价格指数}} \tag{12.4}$$

我国的统计部门目前还没有编制和公布GDP紧缩价格指数，但公布了按照可比价格（不变价格）计算的GDP指数，这实际上就是不包含价格变动的实际GDP的发展速度，反映了国民经济的实际发展速度。据此可以间接地推算GDP平减指数，近似地说明问题。

表12－5中列出了我国若干年份的GDP指数，以及名义GDP、推算的实际GDP和GDP平减指数。表12－5的数据以1978年为基期，表中第4栏GDP平减指数反映了各年国民经济总的价格水平相对于1978年上升的幅度。例如，1999年GDP平减指数为330.6%，表明1999年国民经济总的价格水平相对于1978年上升了230.6%（=330.6%－100%）。

表 12－5 我国 GDP 指数、名义 GDP、推算的实际 GDP 与 GDP 平减指数

年份	定基 GDP 指数(%)	名义 GDP (亿元)	实际 GDP (亿元)	GDP 平减指数(%)
	①	②	③＝3624×①	④－②÷③
1978	100.0	3 624	3 624.1	100.0
1980	116.0	4 518	4 204.0	107.5
1985	192.9	8 964	6 991.0	128.2
1990	281.7	18 548	10 209.1	181.7
1995	496.5	58 478	17 993.7	325.0
1996	544.1	67 884	19 718.7	344.3
1997	592.2	74 463	21 461.3	347.0
1998	638.4	78 345	23 135.6	338.6
1999	683.7	81 911	24 777.3	330.6

资料来源：国家统计局.中国统计年鉴 2000 年.北京：中国统计出版社，2001

四、与 GDP 有关的国民经济总量指标

作为反映国民经济和社会经济发展的统计指标，GDP 不是独自存在的。在国民经济核算体系中存在多个有关国民经济活动最终产出总量的统计指标，构成国民经济总量核算的指标体系。

(一)国民生产总值

国民生产总值(Gross National Product，简称 GNP)，与 GDP 是两种不同内涵的总量指标。

在前面我们已经知道，国民经济核算以常住单位界定核算的空间范围。按照这一原则，凡属本国常住单位生产的最终产品都是 GDP 统计的对象，而非常住单位的生产成果一律排除在外。但是当前各国经济都是开放的经济体系，与世界其他国家发生广泛的经济联系。常住单位在生产中要使用非常住单位的生产要素，非常住单位也要使用常住单位的生产要素。例如，从国外引进技术、资本、雇佣劳动力，或本国的技术、资本、劳动力流向国外。所以本国的常住单位要对非常住单位的生产要素支付利息、工资等要素报酬，这是本期生产的 GDP 在生产领域初次分配；反之，本国的常住单位也会从非常住单位获得劳动者报酬、利息收入等初次分配收入。GNP 就是本国的 GDP 减去支付给非常住单位的原始收入，再加上从非常住单位获得的原始收入。用公式表示就是：

国民生产总值＝国内生产总值－付给国外的要素收入＋来自国外的要素收入

＝国内生产总值(GDP)＋来自国外的要素收入净额　　(12.5)

从这里我们可以看到，GNP 实际上是一个收入概念。早期的国民经济核算中，以 GNP 作为核心指标。在世界经济发展和国民经济核算体系改革的过程中，GDP 逐步取代了 GNP，《1993SNA》将 GNP 改称为 GNI(Gross National Income)，即国民总收入。

一般来讲，经济发达的国家，资本和技术输出较多，从国外取得的要素收入要比付给国外的要素收入多，即来自国外的要素收入净额为正值，从而 GNP 大于 GDP。发展中国家，资本和技术输入比输出多，从国外取得的要素收入要比付给国外的要素收入少，即来自国外的要素收入净额为负值，从而 GNP 小于 GDP。我国在 1980 年以前的 GNP 与 GDP 相等，改革开放以来的 20 多年中，GDP 与 GNP 数值的变化，表现为从 GNP 大于 GDP 到 GDP 大于 GNP 的转变过程，且差额不断扩大。表 12－6 列出了我国若干年来 GNP 和 GDP 的数据及其差额。

表 12－6　1952～1999 年我国的 GNP 与 GDP　　单位：亿元

年份	GDP	GNP	GDP－GNP	年份	GDP	GNP	GDP－GNP
1952	679.0	679.0	0	1995	58 478.1	57 494.9	983.2
1970	2 252.7	2 252.7	0	1996	67 884.6	66 850.5	1 034.1
1980	4 517.8	4 517.8	0	1997	74 762.6	73 142.7	1 619.9
1985	8 964.4	8 989.1	－24.7	1998	78 345.2	76 967.1	1 378.1
1990	18 547.9	18 598.4	－50.5	1999	81 910.9	80 422.8	1 488.1

资料来源：国家统计局. 中国统计年鉴 2000 年. 北京：中国统计出版社，2001

(二)国民生产净值和国内生产净值

国民生产净值(Net National Product，简称 NNP)是国民生产总值减去折旧后的价值，而国内生产净值(Net Domestic Product，简称 NDP)则是国内生产总值减去折旧后的价值。用公式表示就是：

国民生产净值＝国民生产总值－固定资产折旧　　(12.6)

国内生产净值＝国内生产总值－固定资产折旧　　(12.7)

因为固定资产折旧是转移价值，与中间消耗一样都是过去劳动的成果，因此在 GNP 和 GDP 中减去固定资产折旧后的国民生产净值和国内生产净值就是完全排除了转移价值以后的生产新创价值。国民生产净值是常住单位的原始净收入，而国内生产净值则是常住单位生产的最终净成果。国民生产净值与国内生产净值之间的关系用下面的公式表示：

国民生产净值＝国内生产净值－付给国外的要素收入＋来自国外的要素收入

＝国内生产净值＋来自国外的要素收入净额　　(12.8)

和国民生产总值一样，国民生产净值也是一个反映收入的指标，把国民生产总值称为国民总收入，国民生产净值就相应地可称为国民净收入。

（三）国民可支配总收入和国民可支配净收入

国民可支配总收入是国民总收入加上来自国外的政府援助、非官方赠予等非借贷、无须偿还的各种现期净转移收入之和，它是一国可以用作消费和储蓄的全部收入。用公式表示为：

$$\begin{aligned}\text{国民可支配收入} &= \text{国民总收入} + \text{来自国外的现期转移收入} - \\ &\quad \text{付给国外的现期转移支出} \\ &= \text{总消费} + \text{总储蓄}\end{aligned} \tag{12.9}$$

$$\begin{aligned}\text{国民可支配净收入} &= \text{国民可支配总收入} - \text{固定资产折旧价值} \\ &= \text{总消费} + \text{净储蓄}\end{aligned} \tag{12.10}$$

（四）个人收入和个人可支配收入

个人收入是指个人从各种来源获得的初次分配和再分配收入的总和，包括个人得到的劳动报酬、财产收入和各种转移收入。在国民净收入中，有一部分收入不能构成个人收入。如生产税净额归国家所有，企业未分配利润留归企业所有。个人收入用公式表示为：

$$\begin{aligned}\text{个人收入} = {} & \text{国民净收入} - \text{生产税净额} - \text{企业未分配利润} + \\ & \text{政府、企业对个人的债券利息支付}\end{aligned} \tag{12.11}$$

个人可支配收入指实际可用于个人消费支出和储蓄支出的收入，它是缴纳个人税和各种非税支付后所余下的收入，等于个人收入减去所得税、财产税以及其他个人缴纳的余额。

$$\text{个人可支配收入} = \text{个人收入} - \text{个人税及非税支付} \tag{12.12}$$

【例 12－2】 接例 12－1 数据资料，并知某地区得自国外与其他地区的要素收入 400 万元，分配给国外和其他地区的要素收入 1 060 万元；从区外得到的现期转移收入 200 万元，付给区外的现期转移支出 100 万元；企业营业盈余中未分配利润占 50%，所得税中个人税占 45%。试计算该地区国内生产净值、国民总收入、国民净收入、国民可支配总收入、国民可支配净收入、个人收入和个人可支配收入。

解：

$$\begin{aligned}\text{国内生产净值} &= \text{国内生产总值} - \text{固定资产折旧} \\ &= 14\ 167 - 980 \\ &= 13\ 187(\text{万元})\end{aligned}$$

$$\begin{aligned}\text{国民总收入} &= \text{国内生产总值} + \text{区外净要素收入} \\ &= 14\ 167 + (400 - 1\ 060) \\ &= 14\ 167 + (-660) \\ &= 13\ 507(\text{万元})\end{aligned}$$

国民净收入＝国民总收入－固定资产折旧
＝13 507－980
＝12 527(万元)
国民可支配总收入＝国民总收入＋区外现期转移净收入
＝13 507＋(200－100)
＝13 507＋100
＝13 607(万元)
国民可支配净收入＝国民可支配总收入－固定资产折旧
＝13 607－980
＝12 627(万元)
个人收入＝国民净收入－生产税净额－企业未分配利润
＝12 527－2 370－2 297×50%
＝9 008.5(万元)
个人可支配收入＝个人收入－个人税及非税支付
＝9 008.5－360×45%
＝8 846.5(万元)

第三节　国民经济分析常用指标

一、人均水平指标

(一)人均国内生产总值

国内生产总值是反映一国经济总规模的统计指标，它既表明一国经济规模和发展水平，但也受一国国土面积大小和人口数量的影响。例如，经过改革开放20多年来的努力，我国国民经济得到了前所未有的发展，综合经济实力大大提高，GDP规模已经超过10 000亿美元，位居世界第7位。但以我国近13亿的人口规模计算，人均GDP只有800多美元，仍属于低收入国家。比较瑞士的情况，瑞士GDP总量只有3 000多万美元，但人均GDP超过40 000美元，位居世界前列。所以在进行不同国家经济发展水平的横向对比时，不仅要看总量，也要看人均水平。人均GDP反映了一国国民经济生产总量与本国人口比较的相对强度，其计算公式如下：

$$人均\ GDP=\frac{GDP}{人口总数} \tag{12.13}$$

(二)社会劳动生产率

由于一国总人口中,已经退休的老年人口和儿童人口不直接参与社会生产活动,但直接参与GDP的消费,所以把GDP与人口中直接从事社会生产活动的劳动者相比,即社会劳动生产率指标。计算公式如下:

$$社会劳动生产率=\frac{GDP}{社会劳动者人数} \tag{12.14}$$

社会劳动生产率说明每一个社会劳动者提供的国民经济最终产品数量。社会劳动生产率是各国普遍使用的,用来反映整个国民经济活动效率的指标之一。

(三)人均国民生产总值

如前所述,国民生产总值(国民总收入)是一个收入指标。一国总人口中,虽有一部分人口不参与社会生产活动,但是它们都要通过各种收入分配渠道直接或间接获得收入并进行消费。所以虽然每个人不一定都与国内生产总值有直接关系,但却与国民生产总值存在直接的关系。人均国民生产总值集中反映了一国国民经济的发展水平和人民生活的福利水平,是国际上通行的判定一国经济发展所处阶段和进行横向对比的重要依据。表12-7列出了世界银行《世界发展指标》公布的按照汇率法折算的一些国家和地区的人均GNP。

表12-7 1997年世界一些国家和地区人均GNP的数值 单位:美元

	人均GNP	世界排名
低收入国家	350	
中等收入国家	1 890	
下中等收入国家	1 230	
上中等收入国家	4 540	
高收入国家	25 890	
欧盟国家	23 450	
瑞士	43 060	3
日本	38 160	4
美国	29 080	10
韩国	10 550	53
中国	860	145
中国香港	25 200	19
印度	370	177

资料来源:世界银行《1999世界发展指标》

世界银行的一些相关政策就是根据各国人均 GNP 水平来制定的。如表 12－4 所示，世界银行一年一度的《世界发展报告》，也是以人均国民生产总值水平作为划分低收入国家、中等收入国家和高收入国家的依据。

二、收入分配分析指标

（一）初次分配和再分配

GDP 的分配包括初次分配和再分配两个阶段。经过初次分配和再分配后，形成 GDP 分配的最终格局，即国家、企业和居民个人的最终收入。

1. 初次分配。初次分配在生产领域进行，它与 GDP 的来源和创造相联系，并根据各生产要素在生产过程中对 GDP 的贡献大小来分配，谁的贡献大，获得的收入就多。初次分配可以使市场机制对要素价格的形成和资源的合理配置起到基础性作用。初次分配形成三方面的原始收入：

（1）劳动者原始收入。即劳动者的劳动报酬，包括劳动者获得的各种形式的工资、奖金和津贴，农民自产自用产品作价。既包括货币形式收入也包括实物形式收入。

（2）国家财政原始收入。国家和政府是社会的管理者，是企业生产必不可少的外部条件，它在初次分配中获得生产税净额，即生产税减去国家财政对企业的补贴后的差额。生产税是国家政府为社会生产提供或维持基础设施、社会秩序、社会安全、法律制度、金融保险制度、医疗卫生以及文化科学教育等社会环境而应得的收入。

（3）生产单位原始收入，即营业盈余。企业是生产活动的直接组织者和管理者，开办企业的目的就是赢利。营业盈余是生产单位所创造的增加值中扣除劳动者报酬、生产税净额和固定资产折旧后的余额。它相当于企业的营业利润加上从政府获得的生产税补贴。

2. 再分配。再分配在全社会范围进行，着重解决社会发展效率和社会公平问题。在初次分配的基础上，各个原始收入的获得者按照分配制度的规定，或按照自身的意愿，无偿转让其原始收入的一部分，同时也从其他收入主体那里获得一部分转移性收入，从而完成 GDP 的分配流程。GDP 再分配的收支流量种类繁多，涉及范围广泛，主要包括三大类经常性转移收支：

（1）现期所得税、财产税等税收收支。这是得到初次分配收入的居民、企业、金融机构、行政事业单位等，根据当期原始收入应支付的所得税、利润税、资本收益税和定期支付的财产税及其他经常收入税。

（2）社会保险和社会福利。这是为维持居民当前和未来的福利而发生的经常性转移收支。社会保险是居民为保证在未来各个时期能够获得社会福利金，而对政府组织的社会保险计划或各个单位建立的基金所缴纳的款项，如对失业保险、退休保险、医疗

保险计划的缴款，整个社会保险形成居民部门的转移性支出，政府部门和其他部门的转移性收入。社会福利是居民从政府及其他部门得到的经常性转移收入。包括两部分：一是以失业金、退休金、抚恤金、医疗保险等形式出现的社会保险福利；二是以生活困难补助、救济金、人民助学金等形式出现的社会救济金。社会福利形成居民部门的转移性收入，是政府和其他部门的转移性支出。

3.其他经常转移收支。包括政府内部经常转移收支，本国政府与外国政府、国际组织之间发生的援助、捐赠、会费缴纳等经常转移收支，对私人非营利机构的捐赠、赞助等经常转移以及国际间私人转移等等。

（二）国家、企业、居民个人分配构成分析

初次分配和再分配的最终结果，形成各个分配主体的可支配收入。国家、企业和居民个人三者是GDP分配的三个主体，无论是居民个人部门，企业部门，还是政府部门，他们通过初次分配和再分配两个环节以后，最终获得可支配收入。用公式表示为：

$$可支配收入 = 初次分配收入 + 再分配收入 - 再分配支出 \tag{12.15}$$

各部门的可支配收入形成GDP分配的最终格局，即国家、企业、居民个人可用于消费和储蓄的最终收入的构成。

改革开放以来，我国收入分配的格局发生了很大的变化，具体情况如表12-8所示。

表12-8 国家、企业、个人收入构成 （%）

年份	合计	国家	企业	个人
1978	100	32.7	16.1	51.2
1980	100	23.7	17.7	58.6
1990	100	14.5	21.5	64.0
1995	100	14.1	16.8	69.1
1997	100	17.5	14.4	68.1

资料来源：历年《中国统计年鉴》（中国统计出版社）。

表12-8的资料显示，我国收入分配格局变化的总体情况是：国家得到的份额在改革开放后大幅下降，企业所得份额基本保持不变，个人所得份额大幅上升。

（三）居民收入分配差异分析指标

居民收入分配差异分析是关于收入分配的均衡情况分析。存在收入差距的主要原因有两种：一种是由于居民个人本身素质上的差异所引起的收入差距，如受过高等教育和没有受过高等教育的人之间收入的差距；另一种是由于不同地区和不同行业、部门从事不同职业所引起的收入差距。居民收入分配差距的主要分析方法有五等分法和基尼系数法。

1.五等分法。五等分法的基本思路,是把居民按收入水平高低划分成人数相等的5组,计算每一组的收入占总收入的百分比,通过考察各组的收入百分比来分析收入分配的均衡程度。各组的收入比重越偏离20%的均值,则说明收入分配差异越大;反之,各组的收入比重越接近20%的均值,则社会收入分配越均衡。

具体做法是首先把所有社会成员按照收入水平从低到高排队,收入水平最低的20%为一组,称为"低收入组";第二个20%称为"中等偏下收入组";第三个20%称为"中等收入组";第四个20%称为"中等偏上收入组";第五个20%称为"高收入组"。计算每一组的收入合计占全社会收入的比重来反映社会收入分配的差异程度。因为每一组的人数都占总人数的20%,若某一组的收入合计在社会总收入中的比重超过了20%,就说明这一组的收入水平高于社会平均收入水平;反之,若某一组的收入合计在社会总收入中的比重低于20%,就说明这一组的收入水平低于社会平均收入水平。

表12-9是1999年我国城镇居民家庭收入分配用五等分法分析收入水平差异的实际资料。表中资料显示,1999年城镇居民中低收入户、中等偏下收入户、中等收入户占社会收入总额的比重均低于20%,他们的收入均低于社会平均收入水平。而中等偏上收入户和高收入户占社会收入总额的比重均高于20%,高于社会平均收入水平,其中高收入户占有社会收入总额的34.7%,人均可支配收入达到10 584.96元。

表12-9 1999年中国城镇居民家庭收入分配情况

	户数比重(%)	占总收入比重(%)	人均可支配收入(元)
低收入户	20.00	10.2	3 111.35
中等偏下收入户	20.00	14.3	4 363.78
中等收入户	20.00	18.1	5 512.12
中等偏上收入户	20.00	22.7	6 904.96
高收入户	20.00	34.7	10 584.96
合 计	100.00	100.00	5 854.02*

资料来源:国家统计局.中国统计年鉴2000年.北京:中国统计出版社,2001

* 5 854.02为中国城镇居民人均可支配收入数。

2.基尼系数。这种方法是按照统计学家基尼的名字来命名的。基尼系数是国际上最常用的分析收入分配格局的方法,它是在对居民按照收入水平高低进行排队和等分的基础上,绘制洛伦兹曲线图,然后根据图中的有关部分面积进行计算的。洛伦兹曲线的制作方法我们已在第三章中作了介绍,它的图形如图12-2。

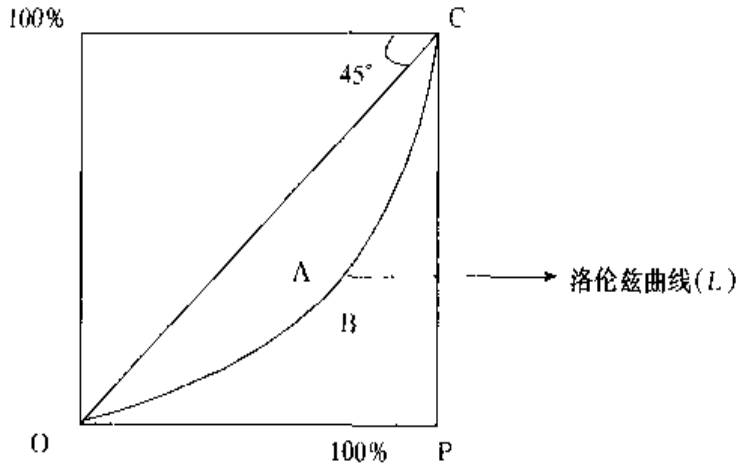


图 12 2

根据洛伦兹曲线的含义，洛伦兹曲线 L 越远离 45°线 OC，收入分配越不平均，洛伦兹曲线越远离 45°线 OC，则与 45°线 OC 之间所围成的面积 S_A 越大。由于三角形 OPC 的面积 S_{A+B}是固定的，基尼系数就是用 S_A 占 S_{A+B}的比重来反映收入分配的均衡程度。用公式表示如下：

$$G = \frac{S_A}{S_{A+B}} \tag{12.16}$$

由于三角形 OPC 的面积 S_{A+B}恒等于 1/2，所以：

$$G = \frac{S_A}{S_{A+B}} = \frac{S_A}{1/2} = 2S_A = 1 - 2S_B \tag{12.17}$$

显然，基尼系数的取值区间为 0～1，基尼系数越接近 1，收入分配的均衡程度越低。世界各国对基尼系数的实际测算表明，大多数国家的基尼系数都在 0.2～0.6 之间。各国通行的判断标准是：基尼系数低于 0.2 表明收入分配高度平均，0.2～0.3 为相对平均，0.3～0.4 大致合理，0.4～0.6 显示收入分配差距过大，0.6 以上则表明收入分配严重向一部分人倾斜。

据有关部门的测算，1980 年以来，我国个人收入分配基尼系数的变化情况是：1981～1984 年，基尼系数在 0.15～0.16 之间，且变化很小，说明我国收入分配处于绝对平均状态。20 世纪 80 年代中期收入差距开始拉大，到 80 年代后期基尼系数达到 0.23。90 年代以后，基尼系数迅速增大，由 1992 年的 0.25 猛增到 1994 年的 0.43。1995 年下半年国家开始采取措施，加大宏观调控力度，使基尼系数在稳定中略有下降。但从 1998 年开始，基尼系数又呈现逐年上升的态势，1998 年为0.456，1999 年为0.457，2000 年又增加到 0.458。

三、最终消费和投资分析指标

(一)消费率与恩格尔系数

消费是以满足生活的需要为目的而使用货物和服务,这种需求也称作最终消费。从国内生产总值的最终使用角度出发,居民和政府是两大消费主体。居民消费是指常住住户对货物和服务的全部最终消费支出,政府消费是指政府部门为全社会提供公共服务的消费支出和免费或以较低价格向住户提供货物和服务的支出。测定消费规模的统计指标是最终消费总额,它等于居民消费和政府消费的总和,最终消费总额与 GDP 比率为消费率,公式表示如下:

$$消费率=\frac{最终消费总额}{GDP}\times 100\% \tag{12.18}$$

在居民消费分析中,恩格尔系数是一个重要指标:

$$恩格尔系数=\frac{食品支出额}{消费支出总额}\times 100\% \tag{12.19}$$

随着人们收入水平的提高,恩格尔系数不断下降是一个普遍规律。因此,恩格尔系数可以作为判断一国经济发展水平和居民收入水平的辅助指标。

(二)投资率

测定投资水平的指标是社会资本形成总额和投资率。社会资本形成总额是国内生产总值最终使用中用于投资的部分,由年度固定资产投资规模和库存增加两部分组成。投资率是社会资本形成总额与国内生产总值的比值,用公式表示如下:

$$投资率=\frac{资本形成总额}{GDP}\times 100\% \tag{12.20}$$

表 12-10 是改革开放后我国若干年份 GDP 最终消费、资本形成总额的分析数据。由表 12-10 可以看出,改革开放后随着我国 GDP 规模的不断扩大,最终消费规模和资本形成总额的规模都逐年增大,消费率和投资率的波动幅度都不大,消费率稳定在 60%左右,投资率保持在 37%左右。

表 12-10 我国 GDP 最终使用构成分析

年份	最终消费		资本形成总额	
	总量(亿元)	消费率(%)	总量(亿元)	投资率(%)
1978	2 239.1	61.8	1 377.9	38.0
1980	2 976.4	65.9	1 590.0	35.2
1985	5 773.0	64.4	3 386.0	37.8

续表

年份	最终消费		资本形成总额	
	总量（亿元）	消费率（%）	总量（亿元）	投资率（%）
1990	11 365.5	61.3	6 444.0	34.7
1995	34 529.4	59.0	23 877.0	40.8
1998	46 405.9	58.7	29 545.9	37.4
1999	49 684.6	60.3	30 496.3	37.0

资料来源：历年《中国统计年鉴》。

四、其他重要的分析指标

GDP是反映国民经济生产、分配和使用总量的最重要指标，对一些关系国民经济全局的重要经济、社会问题，需要将有关指标与GDP进行比较分析。

（一）外贸依存度

外贸依存度有两种计算公式，从不同的角度出发，说明不同范围的问题：

1. 对外贸易总体依存度

$$\text{对外贸易总体依存度}=\frac{\text{商品和服务的进出口总额}}{\text{GDP}}\times 100\% \tag{12.21}$$

对外贸易总体依存度说明了一个国家，包括进口和出口在内的对外贸易总规模，相当于GDP的百分比，它是国民经济总体的对外依赖程度的反映。

2. 出口贸易依存度

$$\text{出口贸易依存度}=\frac{\text{商品和服务的出口额}}{\text{GDP}}\times 100\% \tag{12.22}$$

出口贸易依存度是说明一个国家的出口贸易额相当于GDP的百分比，它是国民经济对出口贸易依赖程度的反映。

改革开放后，我国积极参与国际竞争，开拓国际市场，对外经济往来越来越密切，对外贸易依存度逐年提高。1978年，我国的对外贸易总体依存度只有9.8%，出口依存度只有4.62%；1985年，我国的对外贸易总体依存度为23.26%，出口贸易依存度为9.02%；1990年，外贸总体依存度为29.89%，出口贸易依存度为16.1%；1994年，外贸总体依存度和出口贸易依存度分别达到40.19%和21.29%的历史最高点；1999年外贸总体依存度为32.84%，出口贸易依存度为19.73%。

(二)财政收入占 GDP 的比重

财政收入是国家财政参与社会产品分配所取得的收入,是实现国家职能的保证,所以财政收入规模必须保持在一个合理的范围内。衡量财政规模的大小一般是计算财政收入占 GDP 的比重,计算公式如下:

$$财政收入占\ GDP\ 的比重=\frac{财政收入}{GDP}\times 100\% \tag{12.23}$$

需要指出的是,我国财政收入即通常所说的预算内收入并不是我国政府的全部收入。除了预算收入外,我国政府收入还包括预算外收入和各种规费收入和其他收入。而国外的财政收入、预算收入和政府收入是同一个概念。

另外,政府为弥补财政收入的不足而举借国债(包括内债和外债),政府债务收入占 GDP 的比重也是一个常用的指标,能够反映政府债务规模的合理性、安全程度等。计算公式如下:

$$政府债务收入占\ GDP\ 的比重=\frac{政府债务收入}{GDP}\times 100\% \tag{12.24}$$

(三)教育科技投入占 GNP 的比重

1.教育投入占 GNP 的比重。教育是社会经济发展的基础性事业,教育是非营利性生产部门,主要由政府投资。目前世界各国采用的测度政府教育经费支出力度的指标是政府公共教育支出占 GNP 的比重。该指标计算公式是:

$$政府公共教育经费占\ GNP\ 的比重=\frac{政府公共教育支出}{GNP}\times 100\% \tag{12.25}$$

我国的公共教育经费支出占 GNP 的比重在世界处于低水平,不仅低于经济发达国家,也低于一些发展中国家。表 12-11 是 1996 年世界一些国家公共教育支出占 GNP 比重的资料。

表 12-11 世界一些国家公共教育支出占 GNP 的比重

	中国	加拿大	日本	韩国	波兰	印度
公共教育支出占 GNP 的比重(%)	2.3	7.0	7.3	3.7	5.2	3.4

资料来源:世界银行《1999 世界发展报告》。

2.科技投入占 GDP 的比重。自 20 世纪 70 年代以来,在世界各国的经济发展中,科技进步日益成为决定性的因素。1976 年经济合作与发展组织正式定义了"以知识为基础的经济",即知识经济。知识经济的产业支柱是高技术产业,发展高技术产业,必须以巨大的产业资本投入作为动力,以大批的高水平科技人才队伍为保证。测度国家科技投入力度的大小一般采用科技投入额占 GDP 的比例来反映。而科技投入最常用且国际统一可比的指标是"研究与发展的经费支出",简称"R&D 经费支出",是指用于基

础研究、应用研究和实验发展三方面课题活动的全部实际支出。测度 R&D 经费投入力度的指标如下所示：

$$\text{R\&D 经费支出占 GDP 的比例} = \frac{\text{R\&D 经费支出额}}{\text{GDP}} \times 100\% \tag{12.26}$$

利用这个公式，可以测度整个国家或各个地区的科技投入强度。当需要测度产业或企业的科技投入强度时，则将公式中的分母 GDP 换成产业的增加值或企业的销售收入。20 世纪 90 年代以来，我国 R&D 经费投入占 GDP 的比例呈徘徊不前的趋势，一直保持在 1% 以下的水平。发达国家的这一比例都在 2% 以上。而根据有关的调查，我国大中型企业 R&D 经费投入占销售收入的比例约为 0.2%～0.3% 左右，发达国家这一比例都在 10% 以上。

（四）能源使用效率

能源包括原煤、原油、天然气、水电等。它既是社会生产的动力，又是社会生活中的必需品。所以能源是经济增长和社会发展的重要保障。

能源使用效率是国民经济技术经济水平和效率的反映，测度国民经济总体的能源使用效率，一般将能源消耗总量及其增长率与 GDP 总量及其增长率相对比。具体有以下几种方法：

1. 单位 GDP 生产所消耗的能源总量：

$$\text{单位 GDP 生产所消耗的能源总量} = \frac{\text{生产用能源消耗总量}}{\text{GDP}} \tag{12.27}$$

该指标反映单位国民经济产出量消耗的能源量，也可以用其逆指标单位能源使用所带来的国民经济产出量来反映。

2. GDP 实际增长率与能源消耗量增长率之差：

$$\text{GDP 实际增长率与能源消耗量增长率之差} = \text{GDP 实际增长率} - \text{同期能源消耗增长率} \tag{12.28}$$

我国国民经济的能源使用效率不断提高，每万元 GDP 消耗的能源 1985 年为 8.6 吨；1990 年为 5.2 吨；1995 年为 2.2 吨。但我国的 GDP 增长率仍然慢于能源消费增长率，这还可以根据能源消费弹性系数来分析，计算公式如下：

$$\text{能源消费弹性系数} = \frac{\text{GDP 增长率}}{\text{能源消耗增长率}} \tag{12.29}$$

我国能源消费弹性系数 1985 年为 0.6；1990 年为 0.47；1995 年为 0.66。从世界各国的比较来看，我国国民经济的能源使用效率也是处在较低的水平。

本章小结

1. 国民经济核算采用一整套的统计表、统计指标、账户体系、平衡表、数学模型对国民经济运行过程进行系统描述和反映。国民经济核算体系是由国内生产总值的生产及使用核算、投入产出核算、资金流量核算、国际收支核算和资产负债核算等5个有机联系的子系统组成的科学体系。

2. 国民经济分类是国民经济核算的基础工作。根据考察对象和研究目的的不同，可以选择不同标志，从不同角度对国民经济进行分类。而国民经济本身的构成也是随着社会生产力和社会分工的发展而发展，随着社会生产关系和社会经济体制的变化而变化的。

3. 目前国际上通行的界定国民经济总体范围的原则是"常住经济单位原则"。所谓常住经济单位，是指在一国的经济领土上具有经济利益中心的经济单位。

4. GDP是国民经济核算体系的核心指标，就其价值构成来看包括：①固定资产折旧；②劳动者报酬；③国家生产税净额收入；④企业营业盈余。就其实物内容来看包括：①居民和社会集团消费的全部消费品和服务；②用作投资的产品，即作为固定资产投资的产品和增加储备的产品；③净出口的产品。

5. 与GDP有关的总量指标体系：

国民生产总值(GNP)＝国内生产总值(GDP)＋来自国外的要素收入净额

国民生产净值＝国民生产总值－固定资产折旧

国内生产净值＝国内生产总值－固定资产折旧

国民可支配收入＝国民总收入＋来自国外的现期转移收入－付给国外的现期转移支出

＝总消费＋总储蓄

国民可支配净收入＝国民可支配总收入－固定资产折旧价值

＝总消费＋净储蓄

个人收入＝国民净收入－生产税净额－企业未分配利润＋政府、企业对个人的债券利息支付

个人可支配收入＝个人收入－个人税及非税支付

6. 五等分法思路就是通过考察各组的收入百分比来说明收入分配的均衡程度，即各组的收入比重越偏离20%的均值，则说明收入分配差异越大。基尼系数 $G = S_A/S_{A+B}$，其值越接近1，收入分配的均衡程度越低；越接近于0，收入分配的均衡程度越高。

7.最终消费总额,它等于居民消费和政府消费的总和;社会资本形成总额是年度固定资产投资规模和库存增加两部分的总和。

思考与练习题

一、选择题:

1.国民经济核算的主体是以(　　)界定的。

a. 一国的地理领土范围　　b. 一国的主权领土范围

c. 一国的经济领土范围　　d. 一国常住经济单位

2.中间产品和最终产品的划分依据是(　　)。

a. 产品的性质　　b. 产品的功能

c. 产品的使用去向　　d. 产品的使用价值

3.生产税中不包括(　　)。

a. 农业税　　b. 营业税

c. 所得税　　d. 销售税

4.可支配收入与消费支出的差额构成的指标是(　　)。

a. 储蓄　　b. 投资

c. 积累　　d. 银行储蓄存款

5.一国的恩格尔系数与基尼系数的关系是(　　)。

a. 恩格尔系数越高基尼系数越高

b. 恩格尔系数越高基尼系数越低

c. 恩格尔系数越低基尼系数越高

d. 不能下结论

二、判断题:

1.第三产业的所有部门属于非物质生产部门。(　　)

2.全世界各国的国内生产总值之和小于国民可支配收入之和。(　　)

3.农业部门创造的增加值就是农业部门生产的最终产品价值。(　　)

4.全社会各部门创造的增加值总和等于全社会最终产品价值总和等于国内生产总值。(　　)

5.一国的经济发展水平主要由一国的国内生产总值规模决定。(　　)

三、问答题：

1.给出以下关于生产、分配和使用的资料(见表12－12、表12－13和表12－14)，试用生产法、分配法和使用法计算GDP，并用适当的方法计算国内生产净值、国民生产总值、国民生产净值、国民可支配总收入和国民可支配净收入。

表12－12　生产和消耗　　单位：亿元

生产部门	总产值	中间消耗率%	折旧率%
农业	3 260	25	5
工业	12 460	45	10
建筑业	2 050	55	10
商业运输业	1 375	30	6
政府公共服务	980	12	3
其他服务	5 820	17	3

表12－13　收入分配　　单位：亿元

项目	数额	项目	数额
劳动者报酬	9 581	国外要素收入净额	－318
自产自用产品价值	489	国外经常转移净收入	192
生产税净额	2 382		
营业盈余	2 523.5		

表12－14　支出使用　　单位：亿元

项目	数额	项目	数额
消费品市场购买额	11 381	固定资产形成总额	3 825
库存增加额	382	净出口额	－181

2.GDP比例分析指标有哪些？请查找《中国统计年鉴》计算最近年份各比例分析指标值。

附录一

Excel 概述

★ 本章学习要点与要求 ★

Microsoft Excel 2000 是美国微软公司 Microsoft Office 2000 软件套件的重要组成部分，它是一个用来组织、计算和分析数据的通用电子表格软件。Excel 普及面广，操作简便，备有多种函数与专门的“分析工具库”，具有很强的数据分析能力与图表功能。在 Excel 中，高级用户还可使用 VBA(Visual Basic for Application)进行宏命令的开发。利用宏命令，可以为特定的统计分析任务制定解决方案。此外，作为办公自动化软件的组件之一，Excel 能够与 Office 中的其他软件相结合，便于制作教学的多媒体课件。因此，本书采用 Excel 作为统计计算与分析的软件。

第一节　Excel 的基本操作

一、熟悉 Excel

(一)Excel 工作界面简介

Excel 第一次启动时，在应用程序工作区中显示一个新的空白工作簿，如附图－1 所示，主要包含如下几项内容：标题栏、菜单栏、工具栏、编辑栏、工作表、工作表标签、滚动栏和状态栏等。

附图-1 Excel的工作界面

1.标题栏。"标题栏"告诉用户正在运行的程序名称和正在打开的文件的名称。标题栏显示"Microsoft Excel-Book1"表示此窗口的应用程序为Microsoft Excel,在Excel中打开的当前文件的文件名为"Book1.xls"。

2.菜单栏。菜单栏按功能把Excel命令分成不同的菜单组,它们分别是"文件"、"编辑"、"视图"、"插入"、"格式"、"工具"、"数据"、"窗口"、"帮助"。当菜单项被选中时,引出一个下拉式菜单,可以从中选取相应的子菜单。

另外,在屏幕的不同地方单击鼠标右键时,"快捷菜单"将出现在鼠标指针处。选取"快捷菜单"中的命令同从菜单栏的菜单上选取相应命令的效果是一样的,但选取速度明显加快。

3.工具栏。Excel显示几种工具栏,这些工具可简化用户的操作。工具栏中的按钮都是菜单中常用命令的副本,当鼠标指向某一按钮后,稍等片刻在按钮右下方会显示该按钮命令的含义。用户可以配置工具栏的内容,通过"视图"菜单中的"工具栏"子菜单

来选择显示不同类型的“工具”或全部显示出来。下面介绍出现在 Excel 开始屏幕中的两种工具栏。

“常用”工具栏中为用户准备了访问 Excel 最常用命令的快捷按钮，如“新建文件”按钮，“打开文件”按钮，“保存文件”按钮等。

“格式”工具栏专门放那些和文本外观有关的命令，如“字体”、“字号”、“对齐方式”及“其他选项”。

4. 编辑栏。编辑栏给用户提供活动单元格的信息。在编辑栏中用户可以输入和编辑公式，编辑栏位于附图－1 中第 5 行。

编辑栏由“名字栏”和“公式栏”组成。位于编辑栏左侧的“名字栏”中显示的是活动单元格的坐标，也可在“名字栏”中直接输入一个或一块单元格的地址进行单元格的快速选定；位于编辑栏右侧的“公式栏”可用于编辑活动单元格的内容，它包含三个按钮和一个编辑区。当向活动单元格输入数据时，公式栏中便出现三个按钮，三个按钮从左至右分别是：“×(取消)”按钮、“√(确认)”按钮和“=(公式指南)”按钮。

通常 Excel 在工作区中显示编辑栏。在“视图”菜单中的“编辑栏”命令是一个开关命令，它可以用于隐藏或显示编辑栏。

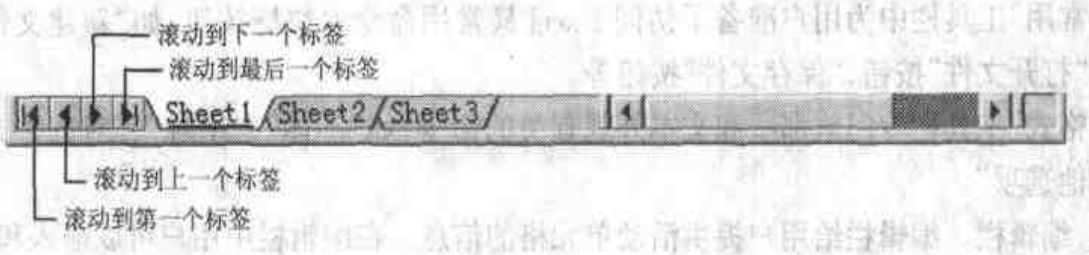
5. 工作表。新的工作簿默认包含三个工作表，名称为“Sheet1”、“Sheet2”和“Sheet3”。当前工作表为 Sheet1，当前工作表只有一张。可以用“工具”菜单中的“选项”对话框来更改新工作簿中所包含的默认工作表的数量。

工作表是一个由行和列组成的表格。行号和列号分别用字母和数字区别。行由上自下范围 1～65 536，列号则由左到右采用字母编号 A～IV。因此，每张表为 256 列×65 536行，若从 Excel 导入的数据超过以上范围，则会被 Excel 自动截去。每一个行、列坐标所指定的位置称之为单元格，如列 A 和行 1 的交叉点的单元格称为单元格 A1。单元格名称也叫“单元格地址”。在单元格中用户可以键入符号、数值、公式以及其他内容。

可以用“插入”菜单中的“工作表”命令来插入新的工作表，用“编辑”菜单中的“删除工作表”命令来删除某个工作表。

6. 工作表标签。工作表标签通常用“Sheet1”，“Sheet2”等名称来表示，用户也可以通过用鼠标右击标签名，选择弹出菜单中“重命名”命令来修改标签名。Excel 一般同时显示工作表队列中的前 3 个标签。利用标签队列左边的一组标签滚动按钮可显示队列中的后续工作表的标签。工作簿窗口中的工作表称之为当前工作表，当前工作表的标签为白色，其他为灰色，如附图－2 所示。

如果工作表数量过多，以致无法同时显示它们，那么就要用左下角的 4 个标签滚动按钮。使用标签滚动按钮可以在工作簿的工作表标签之间进行滚动，并查看工作表的



附图－2　工作表标签

内容。不过这些标签滚动按钮不会激活工作表。要想激活工作表，必须先滚动到所需的工作表，再单击要激活工作表的标签。

7.滚动栏。当工作表很大时，如何在窗口中查看表中的全部内容呢？可以使用工作簿窗口右边及下边的滚动栏（包括滚动框、滚动条和滚动箭头等），使窗口在整张表上移动查看，也可以通过修改"常用"工具栏中"显示比例框"的参数来扩大整个工作表的显示范围。

8.状态栏。状态栏位于Excel窗口底部，它的左端是信息区，右端是键盘状态区。

在信息区中，显示的是Excel的当前工作状态。例如，当工作表准备接受命令或数据时，信息区显示"就绪"；当在编辑栏中键入新的内容时，信息区显示"输入"；当选取菜单命令或工具按钮时，信息区显示此命令或工具按钮用途的简要提示。

在键盘状态区中，显示的是若干按键的开关状态。例如，当按"Caps Lock"键时，状态栏中便显示"CAPS"。

与编辑栏相同，在"视图"菜单中的"状态"命令是一个开关命令，它可以用于隐藏或显示状态栏。

（二）工作表导航

要熟练地使用Excel，必须掌握一些在工作表中移动（或者说是"导航"）的方法。可以使用鼠标、键盘和定位命令在工作表中移动。

1.使用鼠标。用鼠标单击要激活的单元格可将活动单元格从一处移到另一处。要查看在当前工作簿窗口中不可见的工作表部分，可单击垂直或水平滚动条。要注意，使用滚动条滚动时，仅仅改变工作表视图而不改变活动单元格。要在工作簿的工作表之间切换，可单击工作表标签或使用标签滚动按钮在多个工作表标签之间进行切换。

2.使用键盘。利用一些键的组合可在工作表中快速移动。附表－1列出了在工作中最常用的键盘导航键。

附表-1 常用的工作表导航键

键或键的组合	将移至
↑、↓、←、→	箭头方向的下一单元格
Ctrl+↑、Ctrl+↓、Ctrl+←、Ctrl+→	箭头方向的下一包含数据的单元格
Tab	向右一个单元格
Shift+Tab	向左一个单元格
Home	到当前行的列A
Page Up	向上一屏
Page Down	向下一屏
Alt+Page Up	向左一屏
Alt+Page Down	向右一屏
Ctrl+Home	到单元格A1
Ctrl+End	到包含数据的最后一行和最后一列的单元格
Ctrl+Backspace	重新定位工作表的可视部分,显示已滚动到视图之外的活动单元格可选择的范围

3.使用定位命令。单击"编辑"菜单中"定位"命令或按F5键,将显示出"定位"对话框。在"引用位置"文本框中输入单元格的名称并单击"确定"按钮,可以跳转到特定的单元格。

(三)文件基本操作

1.新建文件:执行菜单命令"文件"→"新建",即可创建一个新的文件。

2.打开文件:执行菜单命令"文件"→"打开",调出"打开"对话框,可在Excel中打开一个已经存在的数据文件。它可以是Excel的数据文件,也可是Excel兼容的其他软件的数据文件。可在不同窗口中同时打开多个数据文件,通过"窗口"菜单下方的不同选项,进行不同窗口的切换。

3.保存文件:执行菜单命令"文件"→"保存",可保存当前数据文件。菜单命令"文件"→"另存为",可将当前工作簿存为一个新的文件。文件的格式可以是Excel数据文件,也可是Excel兼容的其他软件的数据文件。

4.文件打印:菜单命令"文件"→"打印",调出"打印"对话框,可打印当前的工作簿文件。打印之前,可以执行菜单命令"文件"→"页面设置"和"打印预览",进行打印前的

页面设置操作和打印效果的预先浏览。

二、输入数据

(一)直接输入

一般来讲,在 Excel 中输入数据,要先选择单元格,然后输入数据并按回车键。输入的数据同时显示在相应的单元格和编辑栏上。如要编辑已输入的数据,可单击编辑栏进行编辑。编辑时,单击"×(取消)"按钮,可放弃编辑(还没按回车键时接受此项),单击"√(确认)"按钮,可接受一个已修改的项。也可双击相应的单元格或选定相应的单元格并按功能键 F2,在单元格中移动插入指针进行编辑。

可以在 Excel 工作表中输入下列形式的信息:

1.数值。数值可以是整数、小数、整分数(如 10, 3/7)或以科学记数法表示的数字(如 2E+9)。可以在数字中使用一些数学符号,包括加(+)、减(-)、百分号(%)、分号(/)、指数符号(E)和货币符号。默认情况下,数值靠单元格的右端对齐。

2.文本。文本可以是任意文字数字字符的组合,包括大小写字母、数字和符号。默认情况下,文本靠单元格的左端对齐。如想要将数值、日期一类的数字保存为文本,在数字前加一个单引号。

3.日期和时间。要在单元格中保存日期和时间,必须使用预定义的日期和时间格式的一种,如,"2002 年 3 月"、"2002-3-3"或"8:00AM"。具体的日期和时间格式请参见 Excel 帮助。

4.公式。公式是一个从现有的数值计算出一个新值的方程式。它可以包括数字、数学运算符、单元格引用和称为"函数"的内部方程式。如"=sum(A5:A8)"。公式的使用将另作介绍。

5.到 Internet 站点或其他文档的超级链接、批注以及电子图片,如剪贴画、图片、地图等。

(二)与其他软件交换数据

除了直接输入数据外,Excel 还可通过复制、公式、与其他软件交换等渠道取得数据。关于复制和公式的操作,我们将在下一节中介绍。这里,介绍如何与其他软件交换数据。

在 Excel 中可以打开其他类型的数据文件,如 DBF 数据库文件,文本文件,lotus1-2-3 的数据文件等。具体操作方法如下:

1. 在"文件"菜单中选择"打开"子菜单。

2. 在"打开文件"对话框中选择所要打开的文件类型即其所在的目录。

3. 用鼠标双击该文件名,然后按 Excel 提示步骤操作即可打开该文件。

三、编辑工作表

(一)选定单元格和区域

一些Excel命令是针对单个单元格或单元格的组合(称为区域)的。活动单元格总是处于选定状态。

要使用鼠标选择一定区域的单元格,可执行下列步骤:

1.鼠标单击要选定的第一个单元格。

2.按下鼠标左键,然后将鼠标拖过选定区,最后放开鼠标左键。被选定的单元格以淡蓝色突出显示,并且外边框为加粗边框。

3.如要选定不相邻的单元格区域,可按住Ctrl键,然后重复步骤1和2直到选定所有的区域。选定后,放开Ctrl键。

还可以通过单击想要选定行的行标,选定一行单元格,或通过单击列标选定一列单元格。通过按下鼠标左键,并拖曳鼠标指针经过希望选定的列或行,则可以选定多个行或列。最后,按下"全选"按钮允许用户在一次操作中选定工作表中的所有单元格,如附图-3所示。

附图-3

要使用鼠标选择一定区域的单元格,可执行下列步骤:

1.使用箭头键移动到要选定的第一个单元格上。

2.按Shift键,然后按相应的箭头键选定其余部分,最后放开Shift键。

3.要选定不相邻的单元格区域,请按下"Shift+F8"组合键(此时状态栏上出现"ADD"),重复步骤1和2。

(二)清除单元格内容

要清除活动单元格的内容,按Delete键,或者右击该单元格,然后从弹出菜单中选

择“清除内容”即可。用户还可以选择一定区域的单元格，然后以同样的方式清除它们的内容。

注意单元格内容并未被完全删除。清除单元格与删除单元格不同。清除单元格只是清除单元格的值，而单元格的格式信息却仍然保留。而删除单元格则是将单元格从电子表格中删除，并用其他单元格来替换该单元格。

（三）插入和删除单元格

要将一个单元格或单元格区域插入到电子表格中，可以选定与要插入新单元格的位置相邻的单元格区域，然后执行菜单命令“插入”→“单元格”→“行”或“列”。用户还可以右击这个单元格，然后从弹出菜单中选择“插入”。如果 Excel 不能确定如何处理其他单元格，则会提示用户是否将单元格下移或右移。它还让用户选择是插入新行还是插入新列。

删除单元格区域情况相同，只是单元格按相反的方向移动。用户可以将单元格左移或上移，或者用户可以删除受影响的行或列。如果用户在选择插入或删除命令之前，选定了一行或多行（或列），那么 Excel 将不会向用户提示额外信息，它只是向前执行所需的操作。

（四）“撤消”、“恢复”和“重复”命令

如果误操作删除了不该删除的区域，可以通过“编辑”菜单的“撤消”命令来恢复被删除的内容。该命令的键盘快捷键是“Ctrl + Z”。

“重复”命令处于“编辑”菜单的“撤消”命令下面，使用该命令可以在工作表的其他位置上重复刚才执行的命令。例如，你刚刚给 A3 单元格设置了某些格式（如加了边框、字体设为一号），则选定 B5 单元格并应用“重复”命令，可以为 B5 单元格设置与 A3 单元格一样的格式。该命令的键盘快捷键是“Ctrl + Y”。

在使用“撤消”命令后，Excel 将“编辑”菜单中“重复”命令的名称更改为“恢复”，可以再次运行上一个命令。“恢复”命令具有与“撤消”命令相反的功能。该命令的键盘快捷键是“Ctrl + Y”。

（五）使用“剪切”和“粘贴”命令移动数据

当使用“剪切”命令剪切某个单元格区域时，该区域周围显示点线矩形框，表示这是要移动的单元格，然后将该单元格区域内容（包括批注和格式）存放在一个临时的存储空间，该存储空间就是 Windows 剪贴板。当选定放置数据的新位置并执行“粘贴”命令，单元格以及它们的格式将从剪贴板复制到新的位置上，同时原始单元格区域的数据和格式将被清除。如果在出现矩形框后要取消移动，请按 Esc 键。

操作步骤如下：

1. 选定要移动的单元格区域。

2.执行“编辑”菜单中的“剪切”命令。也可单击“常用”工具栏上的“剪切”按钮或按快捷键“Ctrl+X”。

3.选定要移动数据的目的单元格。

4.执行“编辑”菜单中的“粘贴”命令。也可单击“常用”工具栏上的“粘贴”按钮或按快捷键“Ctrl+V”。

(六)使用“复制”和“粘贴”命令复制数据

如果只是要复制工作表中的单元格区域,而不是移动它们,可用“编辑”菜单中的“复制”命令。这个命令将选定的单元格副本保存到剪贴板中,然后使用“粘贴”命令可以将这些单元格任意多次地粘贴到工作表中。

其操作步骤基本上与“剪切”和“粘贴”命令相同,只是第2步应改为:执行“编辑”菜单中的“复制”命令。也可单击“常用”工具栏上的“复制”按钮或按快捷键“Ctrl+C”。

(七)使用自动填充功能创建序列

Excel中最有用的特性之一就是它具有自动完成一个序列的能力。一个序列可以是一组数字(如1,2,3或2,4,6)或字母(如A,B,C),也可以是一组日期,如Jan-99,Feb-99,Mar-99或工作日,用户甚至可以定义自己的序列(Excel已有的自定义序列有11个),Excel将自动填充。

自动填充功能通过将“填充柄”拖过新单元格时产生作用。填充柄是位于活动单元格或选定单元格区域右下角的小黑框。当指针停留在填充柄上时,指针形状变成加号(+),这表明自动填充功能已经启用。

要使用自动填充功能,可以按下列步骤进行:

1.在电子表格中输入序列的前几个值。

2.在保持按下鼠标左键的同时,拖动鼠标指针选定这些值。

3.将指针移动到选定区域的填充柄上,然后等待指针变为加号。

4.在保持按下鼠标左键的同时,拖动填充柄覆盖希望填充的单元格区域。

5.释放鼠标左键将利用突出显示的序列填充选定的单元格区域。

用户可以执行菜单命令“工具”→“选项”,在调出的“选项对话框中”选择“自定义序列”标签创建自己的序列。要创建自己的序列条目,可以选择“自定义序列”框中的“新序列”,然后在“输入序列”窗格中键入自己的序列。一次输入一个条目,按回车键转到另一行。完成时,单击“添加”按钮使得用户自己的序列可用,然后单击“确定”按钮关闭对话框。

注意,如果要将公式或数据复制到临近的单元格内,使用“填充柄”会非常方便。

四、使用公式和函数

Excel 具有很强的公式功能，对于没有现成分析工具的统计分析任务，要用 Excel 的公式来完成。公式在数据分析时主要有以下几方面的优点：

1. 公式的复制功能适应大批量数据处理，极大地加快了运算速度。

2. 用 Excel 的公式，可以记住分析的具体步骤，非常便于检查和修改。

3. Excel 具有自动更新功能，修改公式引用的单元格，公式可立刻进行重新计算并显示出结果。

4. 利用 Excel 的函数，可以省去众多手动计算时的中间步骤，提高效率。

Excel 在内部保存公式（可以在编辑栏上看到公式），在放置公式的单元格中显示结果。Excel 中的所有公式都以一个等号（=）开头，公式中可以包括数字、数学运算符、单元格引用和函数。下面介绍应用公式的具体操作。

（一）创建公式

选择要输入公式的单元格，输入一个等号和要进行计算的公式，按加回车键完成公式并显示结果。有两种方式输入公式：一是直接在单元格中输入公式，二是使用“公式选项板”输入公式。要显示“公式选项板”，可单击“编辑栏”上的“编辑公式”按钮。

数组公式的输入：

1. 如果数组公式返回一个结果，请单击需要输入数组公式的单元格。

如果数组公式将返回多个结果，请选定需要输入数组公式的单元格区域。

2. 键入数组公式。

3. 按“CRTL + SHIFT + ENTER”组合键。

在输入数组公式时，Excel 自动在大括号｛｝之间插入公式。

（二）在公式中使用单元格引用

单元格引用指的是在公式中包含单元格的名称。可以将单元格引用与数字、数学运算符和内置函数一起使用。要在公式中指定单元格引用，可以输入单元格名称，使用鼠标或键盘单独突出显示。

例如，要将单元格 B5 和 C5 的内容相加，结果放至 D5 中去，应创建公式：“=B5 + C5”，这时可使用如下 2 种方法：

1. 通过键盘输入创建：选择单元格 D5，输入“=B5 + C5”，按回车键。

2. 用鼠标创建：选择单元格 D5，输入等号（=），单击单元格 B5，然后按加号键（+），再单击单元格 C5，最后按回车键。

在使用单元格引用时，要注意相对引用与绝对引用的区别。在 Excel 中，如果希望当公式复制到别的区域之时，公式中引用的单元格地址不会随之相对变动，那么你必须

在公式中使用单元格的绝对引用。如果你希望当公式复制到别的区域时,公式引用的单元格地址也会随之相对变动,那么你必须在公式中使用单元格的相对引用。在公式中如果直接输入单元格的地址,那么默认的是相对引用单元格,如果在单元格地址的行标和列标前都加上"$"符号,那么意味着绝对引用单元格。

有时,在复制公式时,要求只保持行或列不变,这种情况下要使用混合引用。如果只在单元格地址的行标前或只在列标前加上"$"符号,那么意味着混合引用单元格,即在一个单元格引用中,同时包含相对引用和绝对引用。

Excel有3种方式表示单元格的引用,即冒号、逗号和空格。

冒号表示连续区域的引用,即引用一个矩形区域内的单元格;逗号表示联合引用,即引用若干个单元格或单元格区域,这些单元格或单元格区域并不一定是在一起的。空格表示交叉引用,即引用两个单元格区域的公共部分。

如SUM(A1:B5)表示求A1到A5及B1到B5单元格的和;SUM(A1,A3,B1:B4)表示求A1,A3和B1到B4单元格的和;SUM(A1:B4 A2:B5)求A2到A4及B2到B4单元格的和。

(三)公式的复制

公式复制是Excel数据成批计算的重要操作方法。可以使用自动填充功能来实现公式的复制,也可以用"复制"和"粘贴"命令来实现公式的复制。但要注意的是公式的复制会产生单元格地址的变化,会对计算结果产生影响。

在复制包含相对引用的公式时,Excel会自动调整公式的引用,以便引用相对于当前公式位置的其他单元格。例如,在单元格A1和B1中已输入了数值数据,用鼠标选定C1单元格,然后输入公式"=A1+B1",此公式引用的便是两个相对的单元格A1,B1,也就是说,如果将该公式复制到C2的单元格,公式所引用的单元格的地址将随着发生变化,公式将变为"=A2+B2",如果将该公式复制到单元格F100,那么公式将变为"=D100+E100"。

在复制包含绝对引用的公式时,不论将公式复制到何处,公式的内容都不会发生变化。

在复制包含混合引用的公式时,如果单元格地址的行标前加"$",则保持"行"不变动,但"列"会随着公式位置的变化而变动;如果单元格地址的列标前加"$",则保持"列"不变动,但"行"会随着公式位置的变化而变动。例如,将单元格C1中的公式"=A$1+B$1"复制到单元格F100,那么公式将变为,"=D$1+E$1"。

在复制公式时,可以作这样的归纳:公式中"$"符号后面的单元格坐标不会随着公式位置的变化而变动,而不带"$"符号后面的单元格坐标会随着公式位置的变化而变动。

（四）复制公式的计算结果

可以仅将公式的计算结果复制到目标区域，并不复制公式本身，可用选择性粘贴来实现。操作步骤如下：

1.拖动鼠标选定待复制区域，并执行“复制”命令。

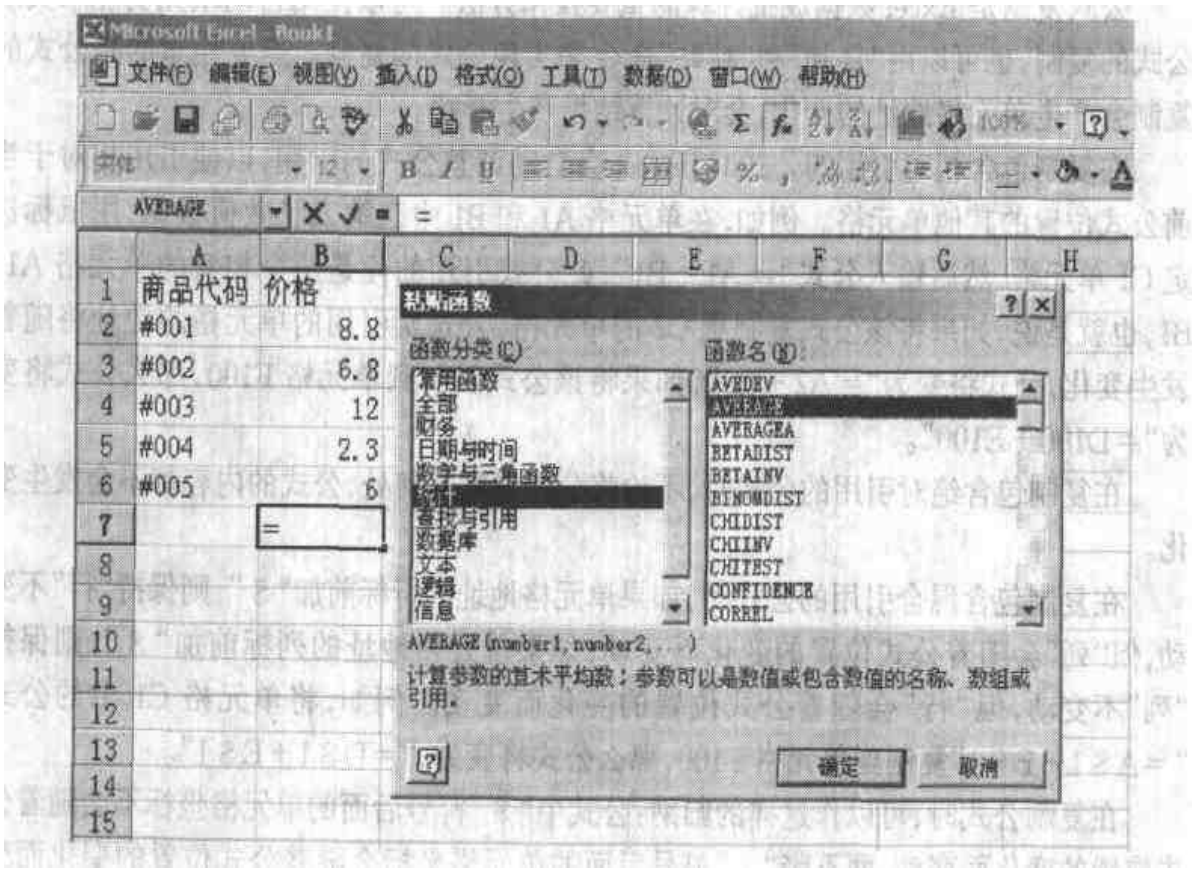
2.用鼠标右击目标区域，执行“选择性粘贴”命令。出现“选择性粘贴”对话框，选定“数值”选项，然后用鼠标单击“确定”按钮。

（五）使用函数

函数是一些预定义的公式，利用函数不仅能提高效率，而且可以减少由于人为原因导致的错误，同时也能减少工作表占用的内存空间，提高 Excel 工作速度。在 Excel 中使用函数，必须将它们输入到工作表的公式中。输入函数比较方便的方法是使用“粘贴函数”对话框。下面用一个例子来说明如何使用“粘贴函数”插入函数。

要使用 Average 函数计算数值列表的平均值，请执行以下步骤：

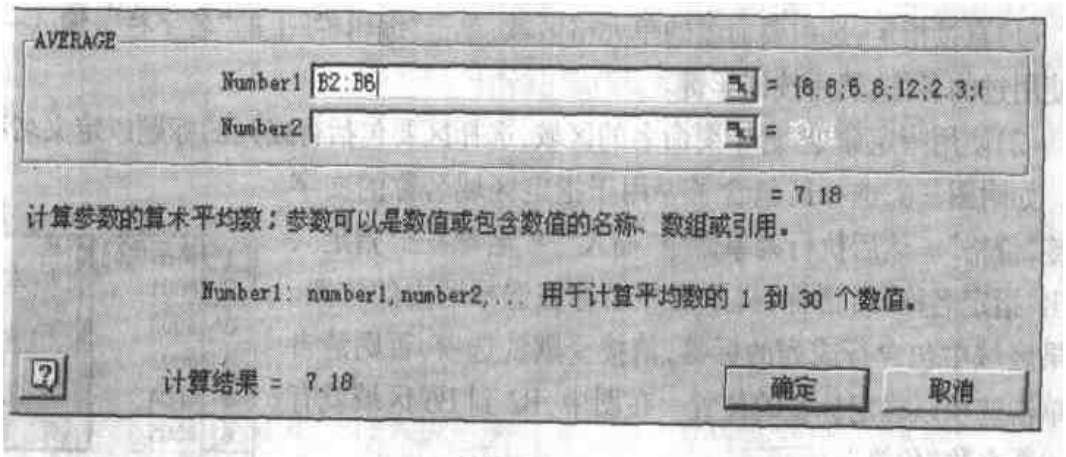
1.选定放置 Average 函数的单元格，然后执行“插入”菜单中的“函数”命令或单击“常用”工具栏上的“粘贴函数”按钮显示“粘贴函数”对话框，见附图－4。在“函数分类”列表框中显示了 9 类函数与“全部”和“常用函数”选项。



附图－4

2.单击“统计”类,“函数名”列表框中将出现统计类的函数。选定“Average”,这时图的下方会出现关于该函数功能的说明,参见附图－4。

3.单击确定。这时出现第二个对话框,询问函数的参数,如附图－5所示。



附图－5

4.单击“折叠对话框”按钮,对话框收缩为要填写数值的文本框。

5.选中要计算平均值的单元格。在本例中,选择单元格B2到B7。

6.放开鼠标左键,然后按下回车键。对话框恢复正常尺寸,同时选中的单元格区域出现在对话框和公式栏的Average函数中,对话框的外观如附图－4。最后单击“确定”按钮。

(六)公式与错误值

如果公式不能正确计算出结果,Excel将显示一个错误值。常见的错误值见附表－2。

附表－2 Excel中常见错误值

错误值	简单说明
#DIV/0!	公式中出现0为除数的情况
#N/A	没有可用数值
#NAME?	公式中使用了Excel不能识别的文本
#NULL!	试图为两个不相交的区域指定交叉点
#NUM!	公式中某个数字有问题
#REF!	公式包含的区域引用无效时
#VALUE!	使用错误的参数或运算对象类型,或公式自动更正功能不能更正公式
######	计算结果太宽,无法在单元格中显示

(七)在公式中使用区域名称

要使公式和函数更具有可读性和更易于输入,可以在工作表中命名一个单元格区域,然后在公式和函数中使用该区域名称代替此区域的单元格引用。

1.指定区域名称:

(1)直接指定:选中要命名的单元格区域,单击"编辑栏"上的"名字栏",输入一个没有使用过的名称,然后按回车键。

(2)使用指定命令:选中要命名的区域,选择区要包括行或列的标题以定义名称。

如附图-6,选择区包含了要用于指定区域名称的文字标签"价格"。然后执行菜单命令"插入"→"名称"→"指定",调出"指定名称"对话框,提示区域名称在选择区内的位置。如果区域中包含行或列的标签,请接受默认选项,否则请自行单击选项以确定标签的位置。在图中,B2 到 B6 区域就有了一个名称"价格"。

	A	B
1	商品代码	价格
2	#001	8.8
3	#002	6.8
4	#003	12
5	#004	2.3
6	#005	6

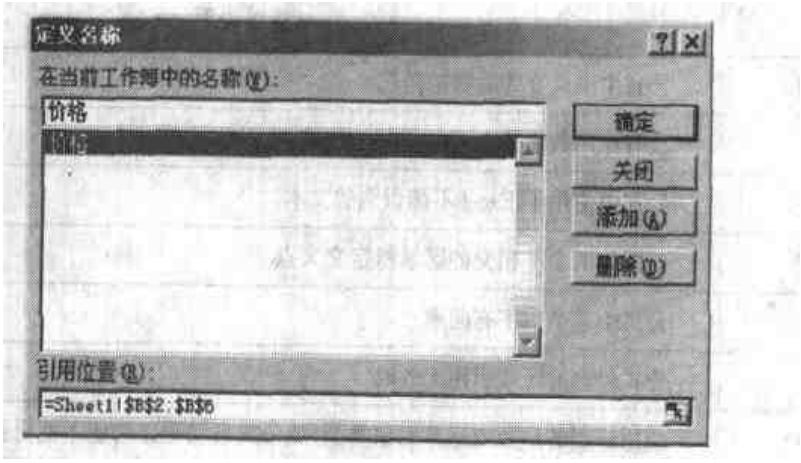
附图-6

(3)使用定义命令:选中要命名的区域。然后执行菜单命令"插入"→"名称"→"定义",调出"定义名称"对话框(见附图-7),在该对话框的"在当前工作簿中的名称"文本框中输入相应的名称,再单击"确定"按钮。

2.使用区域名称。可以在公式和函数中使用区域名称。如本例要计算单元格 B2 到 B6 的平均值,可以用公式"=AVERAGE(价格)"

3.修改区域。当修改区域时,Excel 将自动更新所有的公式,便于修改公式。要修改包含在带名称区域内的单元格,步骤如下:

(1)执行菜单命令"插入"→"名称"→"定义",调出"定义名称"对话框。



附图-7

(2)单击要修改的区域名称。

(3)在"引用位置"文本框修改单元格引用。也可以直接在工作表中选择新的单元格区域。

(4)最后单击"确定"按钮保存所作的修改。

4.删除区域名称。调出"定义名称"对话框,单击要从工作表中删除的区域名称,再单击"删除"按钮,最后按"确定"按钮。

在下面的几节,我们将对Excel作的简要介绍,基本上涵盖进行统计分析所需的Excel基本操作知识和必备的统计分析工具与函数的相关知识。实际应用中,众多的统计分析任务形式多样,有的还十分复杂,在用Excel处理这些分析任务时,必须利用Excel适于处理大批量数据的特点,充分发挥Excel公式和函数的强大功能,灵活地组合应用分析工具库以及各类函数和公式,才能达到统计分析任务的要求。这一点在本书各章的Excel分析介绍中得到了充分的体现。当然,其中的许多统计分析任务也需要我们精通统计理论,熟练掌握统计分析的具体步骤。统计工作表函数用于对数据区域进行统计分析。例如,统计工作表函数可以提供由一组给定值绘制出的直线的相关信息,如直线的斜率和y轴截距,或构成直线的实际点的数值。

第二节 使用分析工具库与统计函数

一、分析工具

Excel提供了一组数据分析工具,称为"分析工具库"。在建立复杂的统计分析时,使用现成的数据分析工具,可以节省很多时间。只需为每一个分析工具提供必要的数据和参数,该工具就会自动使用适宜的统计或数学函数,并在输出表格中显示相应的结果。其中的一些工具在生成输出表格时还能同时产生图表。如果要浏览已有的分析工具,可以单击"工具"菜单中的"数据分析"命令。如果"数据分析"命令没有出现在"工具"菜单上,则必须运行安装程序来加载"分析工具库"。安装完毕之后,通过"工具"菜单中的"加载宏"命令,在"加载宏"对话框中选择并启动它。

使用各项分析工具的操作步骤如下:

1.打开相应的Excel文件。

2.执行菜单命令"工具"→"数据分析",在调出的"数据分析"对话框中双击相应的数据分析工具选项,如"描述统计"、"抽样"等。

4. 在打开的各种分析工具的对话框中，按各种分析工具的要求填写相应的信息，最后按“确定”按钮。

以下我们给出本书所使用的主要分析工具的一览表，如附表－3所示，具体的使用方法在本书各有关章节中说明。

附表－3　Excel主要数据分析工具一览表

分析工具名称	简单说明
F-检验：双样本方差分析	此分析工具可以进行双样本F-检验，用来比较两个样本总体的方差。
t-检验：双样本均值分析	此分析工具及其公式可以进行成对双样本学生氏t-检验，用来确定样本均值是否不等。此t-检验并不假设两个总体的方差是相等的。
t-检验：双样本等方差假设	此分析工具可以进行双样本学生氏t-检验。此t-检验先假设两个数据集的方差相等，故也称作齐次方差t-检验。可以使用t-检验来确定两个样本均值实际上是否相等。
t-检验：双样本异方差假设	此分析工具及其公式可以进行双样本学生氏t-检验。此t-检验先假设两个数据集的方差不等，故也称作异方差t-检验。可以使用t-检验来确定两个样本均值实际上是否相等。
z-检验：双样本均值分析	此分析工具可以进行方差已知的双样本均值z-检验。此工具用于检验两个总体均值之间存在差异的假设。
抽样分析	此分析工具以输入区域为总体构造总体的一个样本。当总体太大而不能进行处理或绘制时，可以选用具有代表性的样本。如果确认输入区域中的数据是周期性的，还可以对一个周期中特定时间段中的数值进行采样。例如，如果输入区域包含季度销售量数据，以4为周期进行取样，将在输出区域中生成某个季度的样本。
回归分析	此工具通过对一组观察值使用“最小二乘法”做直线拟合，进行线形回归分析。本工具可用来分析单个因变量是如何受一个或几个自变量影响的。
描述统计	此分析工具用于生成对输入区域中数据的单变值分析，提供有关数据趋中性和易变性的信息。
排位和百分比排位	此分析工具可以产生一个数据列表，其中将给定数据按大小次序排位和相应的百分比排位。用来分析数据集中各数值间的相互位置关系。

续表

分析工具名称	简单说明
随机数发生器	此分析工具可以按照用户选定的分布类型，在工作表的特定区域中生成一系列独立随机数字，可以通过概率分布来表示主体的总体特征。
相关系数	此分析工具及其公式可用于判断两组数据集(可以使用不同的度量单位)之间的关系，可以使用"相关系数"分析工具来确定两个区域中数据的变化是否相关。
协方差	此分析工具及其公式用于返回各数据点的一对均值偏差之间的乘积的平均值，可以使用协方差工具来确定两个区域中数据的变化是否相关。
移动平均	此分析工具及其公式可以基于特定的过去某段时期中变量的均值，对未来值进行预测。
直方图	在给定工作表中数据单元格区域和接收区间的情况下，计算数据的个别和累积频率，用于统计某个数值元素的出现次数。
指数平滑	此分析工具及其公式基于前期预测值导出相应的新预测值，并修正前期预测值的误差。此工具将使用平滑常数 a，其大小决定了本次预测对前期预测误差的修正程度。
Anova:单因素方差分析	此分析工具通过简单的方差分析(Anova)，对两个以上样本均值进行相等性假设检验(抽样取自具有相同均值的样本空间)。此方法是对双均值检验(如 t 检验)的扩充。
Anova:可重复双因素分析	此分析工具是对单因素 Anova 分析的扩展，即每一组数据包含不止一个样本。
Anova:无重复双因素分析	此分析工具通过双因素 Anova 分析(但每组数据只包含一个样本)，对两个以上样本均值进行相等性假设检验(抽样取自具有相同均值的样本空间)。此方法是对双均值检验(如 t-检验)的扩充。

二、常用数学与统计函数简表

Excel 软件中常用的数学与统计函数，如附表－4 中所示。

附表－4

函数名称	函数功能
ABS	返回参数的绝对值
AVERAGE	求算术平均值
BINOMDIST	求二项式分布的概率值
CHIDIST	求 X^2 分布的单侧概率值
CHIINV	求 X^2 分布给定单侧概率的逆函数值，如果 p = CHIDIST(X，自由度)，则 CHIINV(p，自由度) = X
COMBIN	计算从给定数目的对象集合中提取若干对象的组合数，Combin(n，k) = C_n^k
CONFIDENCE	求总体平均值的置信区间，即求 $z_{\alpha/2}\cdot(\sigma/\sqrt{n})$
CORREL	求相关系数
COUNT	求变量个数
COUNTIF	计算给定区域内满足特定条件的单元格的数目
COVAR	求协方差
CRITBINOM	求使累积二项分布概率值大于等于临界值 α 的最小值
EXP	返回 e 的 n 次幂
EXPONDIST	求指数分布累积分布($1-\lambda e^{-\lambda x}$)或概率密度($\lambda e^{-\lambda x}$)
FACT	返回数的阶乘，即 n!
FACTDOUBLE	返回参数的半阶乘，即 n!!
FDIST	求 F 分布的概率值
FINV	求 F 概率分布的逆函数值，如果 p = FDIST(X，...)，则 FINV(p，...) = X
FREQUENCY	求一组数的频率分布
IF	执行真假值判断，根据逻辑测试的真假值返回不同的结果

续表

函数名称	函数功能
INT	求一个数向下舍位到最接近的整数
KURT	求峰度
LN	求一个数的自然对数
LOG	按所指定的底数,求一个数的对数
LOG10	求以10为底的对数
MAX	求一组数的最大值
MEDIAN	求中位数
MIN	求一组数的最小值
MINVERSE	求逆矩阵
MMULT	求两个矩阵的乘积
MODE	求众数
NORMDIST	求正态分布的概率值(给定平均值和标准偏差)
NORMINV	求给定平均值和标准偏差的正态分布的累积函数的逆函数
NORMSDIST	求标准正态分布的概率值
NORMSINV	求标准正态分布累积函数的逆函数
PERMUT	返回从给定数目的对象集合中选取的若干对象的排列数,Permut$(n,k)=P_n^k$
POISSON	求泊松分布的概率值
POWER	求给定数字的乘幂
PRODUCT	将所有以参数形式给出的数字相乘,再求和
ROUND	求某个数字按指定位数舍入后的数字

续表

函数名称	函数功能
SKEW	求偏度
SQRT	求正平方根
STDEV	求样本的标准差
STDEVP	求总体的标准差
SUMX2MY2	求$\sum(x^2-y^2)$
SUMX2PY2	求$\sum(x^2+y^2)$
SUMXMY2	求$\sum(x-y)^2$
TDIST	求T分布的概率值
TINV	求指定自由度的T分布的逆函数
TRANSPOSE	求矩阵的转置
VAR	求样本方差
VARP	求总体方差

附录二

常用统计表

附表 1　二项分布临界值表

在 $p = q = \frac{1}{2}$ 下，x 或 $n - x$（不论何者为大）的临界值

n	单侧检验(α)		双侧检验(α)	
	0.05	0.01	0.05	0.01
5	5	—	—	—
6	6	—	6	—
7	7	7	7	—
8	7	8	8	—
9	8	9	8	9
10	9	10	9	10
11	9	10	10	11
12	10	11	10	11

续表

n	单侧检验(α)		双侧检验(α)	
	0.05	0.01	0.05	0.01
13	10	12	11	12
14	11	12	12	13
15	12	13	12	13
16	12	14	13	14
17	13	14	13	15
18	13	15	14	15
19	14	15	15	16
20	15	16	15	17
21	15	17	16	17
22	16	17	17	18
23	16	18	17	19
24	17	19	18	19
25	18	19	18	20
26	18	20	19	20
27	19	20	20	21
28	19	21	20	22
29	20	22	21	22
30	20	22	21	23

附表 2 正态分布概率表

$F(Z)=P(|x-\bar{x}|/\sigma<z)$

Z	F(Z)	Z	F(Z)	Z	F(Z)	Z	F(Z)
0.00	0.000 0	0.20	0.158 5	0.40	0.310 8	0.60	0.451 5
0.01	0.008 0	0.21	0.166 3	0.41	0.318 2	0.61	0.458 1
0.20	0.016 0	0.22	0.174 1	0.42	0.325 5	0.62	0.464 7
0.03	0.023 9	0.23	0.181 9	0.43	0.332 8	0.63	0.471 3
0.04	0.013 9	0.24	0.189 7	0.44	0 340 1	0.64	0.477 8
0.05	0.039 9	0.25	0.197 4	0.45	0.347 3	0.65	0.484 3
0.06	0.047 8	0.26	0.205 1	0.46	0.354 5	0.66	0.490 7
0.07	0.055 8	0.27	0.212 8	0.47	0.361 6	0.67	0.497 1
0.08	0.063 8	0.28	0.220 5	0.48	0.368 8	0.68	0.503 5
0.09	0.071 7	0.29	0.228 2	0.49	0.375 9	0.69	0.509 8
0.10	0.079 7	0.30	0.235 8	0.50	0.382 9	0.70	0.516 1
0.11	0.087 6	0.31	0.243 4	0.51	0.389 9	0.71	0.522 3
0.12	0.095 5	0.32	0.251 0	0.52	0.396 9	0.72	0.528 5
0.13	0.103 4	0.33	0.258 6	0.53	0.403 9	0.73	0.534 6
0.14	0.111 3	0.34	0.266 1	0.54	0.410 8	0.74	0.540 7
0.15	0.119 2	0.35	0.273 7	0.55	0.417 7	0.75	0.546 7
0.16	0.127 1	0.36	0.281 2	0.56	0.424 5	0.76	0.552 7
0.17	0.135 0	0.37	0.288 6	0.57	0.431 3	0.77	0.558 7
0.18	0.142 8	0.38	0.296 1	0.58	0.438 1	0.78	0.564 6
0.19	0.150 7	0.39	0.303 5	0.59	0.444 8	0.79	0.570 5

续表

Z	F(Z)	Z	F(Z)	Z	F(Z)	Z	F(Z)
0.80	0.576 3	1.05	0.706 3	1.30	0.806 4	1.55	0.878 9
0.81	0.582 1	1.06	0.710 9	1.31	0.809 8	1.56	0.881 2
0.82	0.587 7	1.07	0.715 4	1.32	0.813 2	1.57	0.883 6
0.83	0.593 5	1.08	0.719 9	1.33	0.816 5	1 58	0.885 9
0.84	0.599 1	1.09	0.724 3	1.34	0.819 8	1.59	0.888 2
0 85	0.604 7	1.10	0.728 7	1.35	0.823 0	1.60	0.890 4
0.86	0.610 2	1.11	0.733 0	1.36	0.826 2	1.61	0.892 6
0.87	0.615 7	1.12	0.737 3	1.37	0.829 3	1.62	0.894 8
0.88	0.621 1	1.13	0.741 5	1.38	0.832 4	1.63	0.896 9
0.89	0.626 5	1.14	0.745 7	1.39	0.835 5	1.64	0.899 0
0.90	0.631 9	1.15	0.749 9	1.40	0.838 5	1.65	0.901 1
0.91	0.637 2	1.16	0.754 0	1.41	0.841 5	1.66	0.903 1
0.92	0.642 4	1.17	0.758 0	1.42	0.844 4	1.67	0.905 1
0.93	0.647 6	1.18	0.762 0	1.43	0.847 3	1.68	0.907 0
0.94	0.652 8	1.19	0.766 0	1.44	0.850 1	1.69	0.909 0
0.95	0.657 9	1.20	0.769 9	1.45	0.852 9	1.70	0.910 9
0.96	0.692 9	1.21	0.773 7	1.46	0.855 7	1.71	0.912 7
0.97	0.668 0	1.22	0.777 5	1.47	0.858 4	1.72	0.914 6
0.98	0.672 9	1.23	0.781 3	1 48	0.861 1	1.73	0.916 4
0.99	0.677 8	1.24	0.785 0	1.49	0.863 8	1.74	0.918 1
1.00	0.682 7	1.25	0.788 7	1.50	0.866 4	1.75	0.919 9
1.01	0.682 5	1.26	0.792 3	1.51	0.869 0	1.76	0.921 6
1.02	0.692 3	1.27	0.795 9	1.52	0.871 5	1.77	0.923 3
1.03	0.697 0	1.28	0.799 5	1.53	0.874 0	1.78	0.924 9
1.04	0.701 7	1.29	0.803 0	1 54	0.876 4	1.79	0.926 5

续表

Z	F(Z)	Z	F(Z)	Z	F(Z)	Z	F(Z)
1.80	0.928 1	2.10	0.964 3	2.60	0.990 7	4.00	0.999 94
1.81	0.929 7	2.12	0.966 0	2.62	0.991 2	4.50	0.999 994
1.82	0.931 2	2.14	0.967 6	2.64	0.991 7	5.00	0.999 999
1.83	0.932 8	2.16	0.969 2	2.66	0.992 2		
1.84	0.934 2	2.18	0.970 7	2.68	0.992 6		
1.85	0.935 7	2.20	0.972 2	2.70	0.993 1		
1.86	0.937 1	2.22	0.973 6	2.72	0.993 5		
1.87	0.938 5	2.24	0.974 9	2.74	0.993 9		
1.88	0.939 9	2.26	0.976 2	2.76	0.994 2		
1.89	0.941 2	2.28	0.977 4	2.78	0.994 6		
1.90	0.942 6	2.30	0.978 6	2.80	0.994 9		
1.91	0.943 9	2.32	0.979 7	2.82	0.995 2		
1.92	0.945 1	2.34	0.980 7	2.84	0.995 5		
1.93	0.946 4	2.36	0.981 7	2.86	0.995 8		
1.94	0.947 6	2.38	0.982 7	2.88	0.996 0		
1.95	0.948 8	2.40	0.983 6	2.90	0.996 2		
1.96	0.950 0	2.42	0.984 5	2.92	0.996 5		
1.97	0.951 2	2.44	0.985 3	2.94	0.996 7		
1.98	0.952 3	2.46	0.986 1	2.96	0.996 9		
1.99	0.953 4	2.48	0.986 9	2.98	0.997 1		
2.00	0.954 5	2.50	0.987 6	3.00	0.997 3		
2.02	0.956 6	2.52	0.988 3	3.20	0.998 6		
2.04	0.958 7	2.54	0.988 9	3.40	0.999 3		
2.06	0.960 6	2.56	0.989 5	3.60	0.999 68		
2.08	0.962 5	2.58	0.990 1	3.80	0.999 86		

附表 3　t-分布临界值表

$P[|t(v)|>t_\alpha(v)]=\alpha$

单侧 双侧	$\alpha=0.10$ $\alpha=0.20$	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01
$v=1$	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	2.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

续表

单侧 双侧	$\alpha=0.10$ $\alpha=0.20$	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
125	1.288	1.657	1.979	2.357	2.616
150	1.287	1.655	1.976	2.351	2.609
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

附表 4　x^2 分布临界值表

$P[x^2(v) > x_\alpha^2(v)] = \alpha$

v	显著性水平(α)												
	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.0002	0.0006	0.0039	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.403	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.068
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	13.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.304	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141

续表

v	显著性水平(α)												
	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.413	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.035	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	37.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.897	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.930	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.593	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

附表 5　F 分布临界值表($\alpha=0.05$)

$P[F(v_1,v_2)>F_\alpha(v_1,v_2)]=\alpha$

v_2 \ v_1	1	2	3	4	5	6	8	10	15
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	241.9	245.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.38	2.23

续表

v_2 \ v_1	1	2	3	4	5	6	8	10	15
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.32	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.30	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.27	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.24	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.22	2.07
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.20	2.06
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.19	2.04
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.18	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	2.03	1.87
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.84
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.97	1.81
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.95	1.79
90	3.95	3.10	2.71	2.47	2.32	2.20	2.04	1.94	1.78
100	3.94	3.09	2.70	2.46	2.31	2.19	2.03	1.93	1.77
125	3.92	3.07	2.68	2.44	2.29	2.17	2.01	1.91	1.75
150	3.90	3.06	2.66	2.43	2.27	2.16	2.00	1.89	1.73
200	3.89	3.04	2.65	2.42	2.26	2.14	1.98	1.88	1.72
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.83	1.67

续表

v_2 \ v_1	1	2	3	4	5	6	8	10	15
1	4052	4999	5403	5625	5764	5859	5981	6065	6157
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.40	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.23	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.55	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	10.05	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.87	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.62	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.26	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.54	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	4.10	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.66
15	8.86	6.36	5.42	4.89	4.56	4.32	4.00	3.80	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.41
17	8.40	6.11	5.19	4.67	4.34	4.10	3.79	3.59	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.43	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	3.09
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.31	3.03
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.26	2.98

续表

v_2 \ v_1	1	2	3	4	5	6	8	10	15
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.21	2.93
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.17	2.89
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	3.13	2.85
26	7.72	5.53	4.64	1.14	3.82	3.59	3.29	3.09	2.81
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	3.06	2.78
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	3.03	2.75
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	3.00	2.73
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.52
50	7.17	5.06	4.20	3.72	3.41	3.19	2.89	2.70	2.42
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.35
70	7.01	4.92	4.07	3.60	3.29	3.07	2.78	2.59	2.31
80	6.96	4.88	4.04	3.56	3.26	3.04	2.74	2.55	2.27
90	6.93	4.85	4.01	3.53	3.23	3.01	2.72	2.52	2.42
100	6.90	4.82	3.98	3.51	3.21	2.99	2.69	2.50	2.22
125	6.84	4.78	3.94	3.47	3.17	2.95	2.66	2.47	2.19
150	6.81	4.75	3.91	3.45	3.14	2.92	2.63	2.44	2.16
200	6.76	4.71	3.88	3.41	3.11	2.89	2.60	2.41	2.13
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.23	2.04

附表 6　秩和检验表

表中列出了秩和下限 $T_1(\alpha)$及秩和上限 $T_2(\alpha)$的值

$\alpha=0.05$								$\alpha=0.025$							
n_1	n_2	$T_1(\alpha)$	$T_2(\alpha)$	n_1	n_2	$T_1(\alpha)$	$T_2(\alpha)$	n_1	n_2	$T_1(\alpha)$	$T_2(\alpha)$	n_1	n_2	$T_1(\alpha)$	$T_2(\alpha)$
2	4	3	11	5	5	19	36	2	6	3	15	5	6	19	41
2	5	3	13	5	6	20	40	2	7	3	17	5	7	20	45
2	6	4	14	5	7	22	43	2	8	3	19	5	8	21	49
2	7	4	16	5	8	23	47	2	9	3	21	5	9	22	53
2	8	4	18	5	9	25	50	2	10	4	22	5	10	24	56
2	9	4	20	5	10	26	54	3	4	6	18	6	6	26	52
2	10	5	21	6	6	28	50	3	5	6	21	6	7	28	56
3	3	6	15	6	7	30	54	3	6	7	23	6	8	29	61
3	4	7	17	6	8	32	58	3	7	8	25	6	9	31	65
3	5	7	20	6	9	33	63	3	8	8	28	6	10	33	69
3	6	8	22	6	10	35	67	3	9	9	30	7	7	37	68
3	7	9	24	7	7	39	66	3	10	9	33	7	8	39	73
3	8	9	27	7	8	41	71	4	4	11	25	7	10	43	83
3	9	10	29	7	9	43	76	4	5	12	28	8	8	49	87
3	10	11	31	7	10	46	80	4	6	12	32	8	9	51	93
4	4	12	24	8	8	52	84	4	7	13	35	8	10	54	98
4	5	13	27	8	9	54	90	4	8	14	38	9	9	63	108
4	6	14	30	8	10	57	95	4	9	15	41	9	10	66	114
4	7	15	33	9	9	66	105	4	10	16	44	10	10	79	131
4	8	16	36	9	10	69	111	5	5	18	37				
4	9	17	39	10	10	93	127								
4	10	18	42												

附表 7 游程检验 R 临界值表

表中对应于 n_1 与 n_2 的有两行数值。若 R 等于小于上行数值，或等于大于下行数值，则在 $\alpha = 0.025$（单侧检验）或 $\alpha = 0.05$（双侧检验）水平上判定序列为非随机的。

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4				2 9	2 9	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2 9	2 10	3 10	3 11	3 11	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2 9	3 10	3 11	3 12	3 12	4 13	4 13	4 13	4 13	5	5	5	5	5	5	6	6
7		2	2	3 11	3 12	3 13	4 13	4 14	5 14	5 14	5 14	5 15	5 15	6 15	6	6	6	6	6
8		2	3	3 11	3 12	4 13	4 14	5 14	5 15	5 15	6 16	6 16	6 16	6 16	6 17	7 17	7 17	7 17	7 17
9		2	3	3	4 13	4 14	5 14	5 15	5 16	6 16	6 16	6 17	7 17	7 18	7 18	7 18	8 18	8 18	8 18
10		2	3	4	4 13	5 14	5 15	6 16	6 16	7 17	7 17	7 18	8 18	8 18	8 19	9 19	9 19	9 20	9 20

续表

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
11		2	3	4	4 13	5 14	5 15	6 16	6 17	7 17	7 18	7 19	8 19	8 19	8 20	9 20	9 20	9 21	9 21
12	2	2	3	4	4 13	5 14	6 16	6 16	7 17	7 18	7 19	8 19	8 20	8 20	9 21	9 21	9 21	10 22	10 22
13	2	2	3	4	5	5 15	6 16	6 17	7 18	7 19	8 19	8 20	9 20	9 21	9 21	10 22	10 22	10 23	10 23
14	2	2	3	4	5	5 15	6 16	7 17	7 18	8 19	8 20	9 20	9 21	9 22	10 22	10 23	10 23	11 23	11 24
15	2	3	3	4	5	6 15	6 16	7 17	7 18	8 19	8 19	9 20	9 20	10 21	10 21	11 22	11 22	11 23	12 23
16	2	3	4	4	5	6	6 17	7 18	8 19	8 20	9 21	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 25
17	2	3	4	4	5	6	7 17	7 18	8 19	9 20	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 26	13 26
18	2	3	4	5	5	6	7 17	8 18	8 19	9 20	9 21	10 22	10 23	11 24	11 25	12 25	12 26	13 26	13 27
19	2	3	4	5	6	6	7 17	8 18	8 20	9 21	10 22	10 23	11 23	11 24	12 25	12 26	13 26	13 27	13 27
20	2	3	4	5	6	6	7 17	8 18	9 20	9 21	10 22	10 23	11 24	12 25	12 25	13 26	13 27	13 27	14 28

附表 8 累计法平均增长速度查对表

递增速度　　间隔期:1～5年

平均每年增长 %	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
0.1	100.10	200.30	300.60	401.00	501.50
0.2	100.20	200.60	301.20	402.00	503.00
0.3	100.30	200.90	301.80	403.00	504.50
0.4	100.40	201.20	302.40	404.00	506.01
0.5	100.50	201.50	303.01	405.03	507.56
0.6	100.60	201.80	303.61	406.03	509.06
0.7	100.70	202.10	304.21	407.03	510.57
0.8	100.80	202.41	304.83	408.07	512.14
0.9	100.90	202.71	305.44	409.09	513.67
1.0	101.00	203.01	306.04	410.10	515.20
1.1	101.10	203.31	306.64	411.11	516.73
1.2	101.20	203.61	307.25	412.13	518.27
1.3	101.30	203.92	307.87	413.17	519.84
1.4	101.40	204.22	308.48	414.20	521.40
1.5	101.50	204.52	309.09	415.23	522.96
1.6	101.60	204.83	309.71	416.27	524.53
1.7	101.70	205.13	310.32	417.30	526.10
1.8	101.80	205.43	310.93	418.33	527.66
1.9	101.90	205.74	311.55	419.37	529.24
2.0	102.00	206.04	312.16	420.40	530.80

续表

平均每年增长 %	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
2.1	102.10	206.34	312.77	421.44	532.39
2.2	102.20	206.65	313.40	422.50	534.00
2.3	102.30	206.95	314.01	423.53	535.57
2.4	102.40	207.26	314.64	424.60	537.20
2.5	102.50	207.56	315.25	425.63	538.77
2.6	102.60	207.87	315.88	426.70	540.40
2.7	102.70	208.17	316.49	427.73	541.97
2.8	102.80	208.48	317.12	428.80	543.61
2.9	102.90	208.78	317.73	429.84	545.20
3.0	103.00	209.09	318.36	430.91	546.84
3.1	103.10	209.40	319.00	432.00	548.50
3.2	103.20	209.70	319.61	433.04	550.10
3.3	103.30	210.01	320.24	434.11	551.74
3.4	103.40	210.32	320.88	435.20	553.41
3.5	103.50	210.62	321.49	436.24	555.01
3.6	103.60	210.93	322.12	437.31	556.65
3.7	103.70	211.24	322.76	438.41	558.34
3.8	103.80	211.54	323.37	439.45	559.94
3.9	103.90	211.85	324.01	440.54	561.61
4.0	104.00	212.16	324.65	441.64	563.31
4.1	104.10	212.47	325.28	442.72	564.98
4.2	104.20	212.78	325.92	443.81	566.65
4.3	104.30	213.08	326.54	444.88	568.31
4.4	104.40	213.39	327.18	445.98	570.01
4.5	104.50	213.70	327.81	447.05	571.66

续表

平均每年增长%	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
4.6	104.60	214.01	328.45	448.15	573.36
4.7	104.70	214.32	329.09	449.25	575.06
4.8	104.80	214.63	329.73	450.35	576.76
4.9	104.90	214.94	330.37	451.46	578.48
5.0	105.00	215.25	331.01	452.56	580.19
5.1	105.10	215.56	331.65	453.66	581.89
5.2	105.20	215.87	332.29	454.76	583.60
5.3	105.30	216.18	332.94	455.89	585.36
5.4	105.40	216.49	333.58	456.99	587.06
5.5	105.50	216.80	334.22	458.10	588.79
5.6	105.60	217.11	334.86	459.29	590.50
5.7	105.70	217.42	335.51	460.33	592.26
5.8	105.80	217.74	336.17	461.47	594.04
5.9	105.90	218.05	336.82	462.60	595.80
6.0	106.00	218.36	337.46	463.71	597.54
6.1	106.10	218.67	338.11	464.84	599.30
6.2	106.20	218.98	338.75	465.95	601.04
6.3	106.30	219.30	339.42	467.11	602.84
6.4	106.40	219.61	340.07	468.24	604.61
6.5	106.50	219.92	340.71	469.35	606.35
6.6	106.60	220.24	341.38	470.52	608.18
6.7	106.70	220.55	342.03	471.65	609.95
6.8	106.80	220.86	342.68	472.78	611.73
6.9	106.90	221.18	343.35	473.95	613.56
7.0	107.00	221.49	343.99	475.07	615.33

续表

平均每年增长 %	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
7.1	107.10	221.80	344.64	476.20	617.10
7.2	107.20	222.12	345.31	477.37	618.94
7.3	107.30	222.43	345.96	478.51	620.74
7.4	107.40	222.75	346.64	479.70	622.61
7.5	107.50	223.06	347.29	480.84	624.41
7.6	107.60	223.38	347.96	482.01	626.25
7.7	107.70	223.69	348.61	483.15	628.05
7.8	107.80	224.01	349.28	484.32	629.89
7.9	107.90	224.32	349.94	485.48	631.73
8.0	108.00	224.64	350.61	486.66	633.59
8.1	108.10	224.96	351.29	487.85	635.47
8.2	108.20	225.27	351.94	489.00	637.30
8.3	108.30	225.59	352.62	490.19	639.18
8.4	108.40	225.91	353.29	491.37	641.05
8.5	108.50	226.22	353.95	492.54	642.91
8.6	108.60	226.54	354.62	493.71	644.76
8.7	108.70	226.86	355.30	494.91	646.67
8.8	108.80	227.17	355.96	496.08	648.53
8.9	108.90	227.49	356.63	497.26	650.41
9.0	109.00	227.81	357.31	498.47	652.33
9.1	109.10	228.13	357.99	499.67	654.24
9.2	109.20	228.45	358.67	500.87	656.15
9.3	109.30	228.76	359.33	502.04	658.02
9.4	109.40	229.08	360.01	503.25	659.95
9.5	109.50	229.40	360.69	504.45	611.87

续表

平均每年增长%	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
9.6	109.60	229.72	361.37	505.66	663.80
9.7	109.70	230.04	362.05	506.86	665.72
9.8	109.80	230.36	362.73	508.07	667.65
9.9	109.90	230.68	363.42	509.30	669.62
10.0	110.00	231.00	364.10	510.51	671.56
10.1	110.10	231.32	364.78	511.72	673.50
10.2	110.20	231.64	365.47	512.95	675.47
10.3	110.30	231.96	366.15	514.16	677.42
10.4	110.40	232.28	366.84	515.39	679.39
10.5	110.50	232.60	367.52	516.61	681.35
10.6	110.60	232.92	368.21	517.84	683.33
10.7	110.70	233.24	368.89	519.05	685.28
10.8	110.80	233.57	369.60	520.32	687.32
10.9	110.90	233.89	370.29	521.56	689.32
11.0	111.00	234.21	370.97	522.77	691.27
11.1	111.10	234.53	371.66	524.01	693.27
11.2	111.20	234.85	372.35	525.25	695.27
11.3	111.30	235.18	373.06	526.52	697.32
11.4	111.40	235.50	373.75	527.76	699.33
11.5	111.50	235.82	374.44	529.00	701.33
11.6	111.60	236.15	375.15	530.27	703.38
11.7	111.70	236.47	375.84	531.52	705.41
11.8	111.80	236.79	376.53	532.76	707.43
11.9	111.90	237.12	377.24	534.03	709.48
12.0	112.00	237.44	377.93	535.28	711.51

续表

平均每年增长 %	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
12.1	112.10	237.76	378.62	536.52	713.53
12.2	112.20	238.09	379.34	537.82	715.63
12.3	112.30	238.41	380.03	539.07	717.67
12.4	112.40	238.74	380.75	540.37	719.78
12.5	112.50	239.06	381.44	541.62	721.82
12.6	112.60	239.39	382.16	542.92	723.94
12.7	112.70	239.71	382.85	544.17	725.98
12.8	112.80	240.04	383.57	545.47	728.09
12.9	112.90	240.36	384.26	546.72	730.14
13.0	113.00	240.69	384.98	548.03	732.28
13.1	113.10	241.02	385.70	549.33	734.40
13.2	113.20	241.34	386.39	550.59	736.46
13.3	113.30	241.67	387.11	551.89	738.59
13.4	113.40	242.00	387.83	553.20	740.73
13.5	113.50	242.32	388.53	554.48	742.83
13.6	113.60	242.65	389.25	555.79	744.98
13.7	113.70	242.98	389.97	557.10	747.13
13.8	113.80	243.30	390.67	558.38	749.23
13.9	113.90	243.63	391.39	559.69	751.38
14.0	114.00	243.96	392.11	561.00	753.53
14.1	114.10	244.29	392.84	562.34	755.74
14.2	114.20	244.62	393.56	563.65	757.89
14.3	114.30	244.94	394.26	564.93	760.01
14.4	114.40	245.27	394.99	566.27	762.21
14.5	114.50	245.60	395.71	567.59	764.39

续表

平均每年增长%	各年发展水平总和为基期的%				
	1年	2年	3年	4年	5年
14.6	114.60	245.93	396.43	568.90	766.55
14.7	114.70	246.26	397.16	570.24	768.76
14.8	114.80	246.59	397.88	571.56	770.94
14.9	114.90	246.92	398.61	572.90	773.16
15.0	115.00	247.25	399.34	574.24	775.38
15.1	115.10	247.58	400.06	575.56	777.56
15.2	115.20	247.91	400.79	576.91	779.80
15.3	115.30	248.24	401.52	578.25	782.02
15.4	115.40	248.57	402.25	579.60	784.26
15.5	115.50	248.90	402.98	580.94	786.48
15.6	115.60	249.23	403.71	582.29	788.73
15.7	115.70	249.56	404.44	583.64	790.97
15.8	115.80	249.90	405.19	585.02	793.26
15.9	115.90	250.23	405.92	586.36	795.49
16.0	116.00	250.56	406.65	587.71	797.74

部分习题参考答案

第一章

一、判断题：

1.× 2.× 3.× 4.√ 5.× 6.×7.× 8.√

二、选择题：

1.a 2.不变标志:a 数量标志:b, c 3.a

三、问答题：

例如考察全国居民人均住房情况，全国所有居民构成统计总体，每一户居民是统计单位，抽查其中5 000户，这被抽查的5 000户居民构成样本。

第二章

一、选择题：

1.c 2.a 3.a 4.a,b,d

二、问答题：

1.(略)

2.不合适。因这样的分组不符合“互斥原则”。

3.(略)

4.

按消费品支出额分组	f	$\frac{f_i}{\sum f_i}$(%)	向上累计		向下累计	
			频数	频率(%)	频数	频率(%)
800以下	1	2	1	2	50	100
800~900	16	32	17	34	49	98
900~1 000	26	52	43	86	33	66

续表

按消费品支出额分组	f	$\frac{f_i}{\sum f_i}$(%)	向上累计		向下累计	
			频数	频率(%)	频数	频率(%)
1 000~1 100	5	10	48	96	7	14
1 100 以上	2	4	50	100	2	4
合　计	50	100				

第三章

二、问答题：

1. 表 3-21 中的主词：国内生产总值及其按三次产业分组；宾词：2000，2001 年产值及速度数据。该表属于主词简单分组表。

表 3-22 中的主词：社会商品零售额及其分组；宾词：月总额和速度数据。该表属于主词复合分组表。

表 3-23 中的主词：饮食业按经济类型和县市分组；宾词：机构和人员的数据。该表属于主词复合分组表。

3.（略）

4. 否。在两者研究的总体一致和所选择的权数合适的条件下两者计算结果一致。

6. 甲市场农产品平均价格 1.375，乙市场农产品平均价格 1.325。

8. $\bar{x} = 525, \sigma = 164.544\ 8, M_e = 522.857\ 1, M_o = 537.5$

10. 销售额：$\bar{x} = 536.25, \sigma = 289.22, V = 0.539\ 3$

利润：$\bar{x} = 32.51, \sigma = 21.6, V = 0.664\ 4$

12. 优秀率：$\sigma_p^2 = 0.127\ 5, \sigma_p = 0.357\ 1$

合格率：$\sigma_p^2 = 0.09, \sigma_p = 0.3$

第四章

一、单项选择题：

1.d　2.a　3.b　4.b

三、问答题：

1.

(1) $\mu = 180$(元)，$\sigma = 28.28$(元)

(2)样本的抽样分布如下：

样本平均(元)	频数	频率
140	1	1 / 25
150	2	2 / 25
160	3	3 / 25
170	4	4 / 25
180	5	5 / 25
190	4	4 / 25
200	3	3 / 25
210	2	2 / 25
220	1	1 / 25
合计	25	1

(3)$\overline{X} = 180 = \mu$

(4)$\sigma_{\bar{x}} = 20$

(5)$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 20$,与(4)结果同

(6)计算方法与上同,略。

3.

(1) $\overline{X} = 560, \sigma_{\bar{x}} = 4.635$

(2) $\mu = 560$,区间估计:[550.8, 569.2]

5.

(1)合格品率 $p = 95\%, \sigma_p = 0.015$

(2)$\Delta_p = z \cdot \sigma_p = 2 \times 0.015 = 0.03$,置信区间:[92%,98%]

合格品数量的区间估计为:[1 840,1 960]

(3)$z = \frac{\Delta_p}{\sigma_p} = 1.54, F(z) = 87.64\%$

第五章

一、单项选择题：

1.b 2.b 3.d 4.c

三、问答题：

1.检验统计量：$t = 1.33$

结论：$\alpha = 0.01$ 与 $\alpha = 0.05$ 都接受原假设，认为这批产品的平均重量与 800 克没有显著差异

3.

(1) 检验统计量：$z = 4.69$。结论：拒绝原假设；

(2) p-值 $= 2p(|z| > 4.69) = 1 - F(4.69) = 0.000\,006$。结论拒绝原假设。

(3) 检验统计量：$z - 1.25$。结论：接受原假设；

(4) p-值 $= p(z < -1.25) = [1 - F(1.25)]/2 - 0.105\,7$。结论与(3) 同；

(5) 这时样本容量 $n = 25$，是小样本，检验统计量改为 t 统计量，其他与上同，略

5. 检验统计量：$t = 2.204$。结论：接受原假设；

p-值 $= 1 - p(-2.204 < t < 2.204) = 6.61\%$ 结论相同。

7. $n^+ = 6, n^- = 4, n = 10$。(双侧) 临界值是 $9, n^+ = 6 < 9$。接受原假设。

第六章

一、单项选择题：

1.d 2.b 3.a

三、问答题：

1. SST 自由度是 $rnm - 1$；SSA 自由度是 $r - 1$，SSB 自由度是 $n - 1$；SSAB 的自由度是 $(r-1)(n-1)$，$SSE = \sum_i \sum_j \sum_l (X_{ijl} - \overline{X}_{ij.})^2$ 的自由度是 $rn(m-1)$

3. 判断结论：拒绝原假设，认为品种对产量有显著影响

第七章

一、判断题：

1.× 2.× 3.√ 4.× 5.√ 6.√

二、选择题：

1.b, c, d 2.a 3.a, b, d 4.a, b, c, d

三、问答题：

3. 证：设 $\tilde{\beta}_2 = \sum a_t Y_t$ 为 β_2 的任意线性无偏估计量，则有：

$$E(\tilde{\beta}_2) = \sum a_t E(\beta_1 + \beta_2 X_t + u_t)$$

$$= \beta_1 \sum a_t + \beta_2 \sum a_t X_t + \sum a_t E(u_t) - \beta_2$$

也就是说,作为 β_2 的任意线性无偏估计量,必须满足下列制约条件:

$$\sum a_t = 0; \qquad \sum a_t X_t = 1$$

按照书中介绍的推导估计量方差的方法,可推导出:

$$\mathrm{Var}(\tilde{\beta}_2) - E(\sum a_t u_t)^2 = \sigma^2 \sum a_t^2$$

书中已经推导出最小二乘估计量的方差 $Var(\hat{\beta}_2) = \sigma^2 \sum w_t^2$

而 $\sum a_t^2 = \sum (a_t - w_t + w_t)^2$

$$= \sum (a_t - w_t)^2 + 2\sum w_t(a_t - w_t) + \sum w_t^2$$

$$= \sum (a_t - w_t)^2 + \sum w_t^2 \geqslant \sum w_t^2$$

证毕。

以上推导过程中,利用了前面关于线性无偏估计量的制约条件和 $w_t = \dfrac{X_t - \overline{X}}{\sum (X_t - \overline{X})^2}$。

5.

(1) $\hat{\beta}_2 = \dfrac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = 0.786\,3$

$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} = 549.8 - 0.786\,3 \times 647.88 = 40.372\,0$

(2) $r^2 = \dfrac{[\sum (X - \overline{X})(Y - \overline{Y})]^2}{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2} = 0.999\,835$

$\sum e_t^2 = (1 - r^2) \sum (Y - \overline{Y})^2 = 44.685\,4; S_e = \sqrt{\dfrac{\sum e_t^2}{n-2}} = 2.113\,9$

(3) $H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$

$S_{\beta_2} = \dfrac{S_e}{\sqrt{\sum (X - \overline{X})^2}} = 0.003\,24; t_{\beta_2} = 242.685 t_{0.05}(10) = 2.228$

(4) $Y_f = 40.372\,0 + 0.786\,3 \times 800 = 669.41$(万元)

$S_{ef} = \sqrt{1 + \dfrac{1}{12} + \dfrac{(800 - 647.83)^2}{425\,053.73}} = 1.066\,7$

区间预测:$667.03 \leqslant Y_f \leqslant 671.19$

7.

(1) $C_t = 466.80 + 0.447\,1 Y_t + 0.264 C_{t-1}$

(2) $S = 442.22$;

(3) $\bar{R}^2 = 0.999\,4$

(4) $t_{\beta_1} = 3.35$; $t_{\beta_2} = 15.66$; $t_{\beta_3} = 4.94$

(5) $F = 16\,484$,整个方程非常显著

(6) $56\,375 \leqslant Y_f \leqslant 58\,658$

第八章

一、选择题:

1.d 2.d 3.d 4.a 5.c

二、判断题:

1.× 2.× 3.√ 4.× 5.×

三、问答题:

1. 1 260 人

3. 15 053.6 百万

5. 4.25%

7.1 969.33 亿元;12.26%

9. $y_c = 861.50 + 12.22t$

11.普通最小二乘法:$y_c = 0.87 + 0.63t$;加权最小二乘法:$y_c = 0.915 + 0.615t$

第九章

一、选择题:

1.a,b,c,d 2.a,b 3.b,c

二、问答题:

2.

(1)不会购买彩票。因为,其期望净收益必然为负数;

(2)其采用的是满意准则。买彩票中奖的概率虽然很小,但要想用极小的代价获得非常高的收益这可能是惟一的机会。

4.

(1)

收益矩阵表

θ_j:天气状况		天气好	天气坏
先验概率 $P(\theta_j)$		0.6	0.4
方案	9月施工 a_1	1 000	-500
	10月施工 a_2	0	0

$E(Q(a_1)) = 1\,000 \times 0.6 - 500 \times 0.4 = 400$ 万元; $E(Q(a_2)) = 0$

(2) $EVPI = 0 \times 0.6 + 500 \times 0.4 = 200$ 万元

(3)

天气状况 θ_j	发出天气好预报的后验概率 $P(\theta_j/e_1)$	发出天气坏预报的后验概率 $P(\theta_j/e_2)$
天气好 θ_1	0.818	0.176 5
天气坏 θ_2	0.182	0.823 5

(4)

① 利用后验概率 $P(\theta_j/e_1)$ 计算期望收益:

$E(Q(a_1)) = 1\,000 \times 0.818 - 500 \times 0.182 = 727$ 万元

$E(Q(a_2)) = 0 \times 0.818 + 0 \times 0.182 = 0$ 万元

根据期望值准则应选择方案一 a_1。

② 利用后验概率 $P(\theta_j/e_2)$ 计算期望收益:

$$E(Q(a_1)) = 1\,000 \times 0.176\,5 - 500 \times 0.823\,5 = -235.25 \text{ 万元}$$

$$E(Q(a_2)) = 0 \times 0.818 + 0 \times 0.182 = 0 \text{ 万元}$$

根据期望值准则应选择方案二 a_2。

(5) 计算后验完全信息价值

发出天气好预报的后验完全信息价值

后验 $EVPI(e_1) = 0 \times 0.818 + 500 \times 0.182 = 91$ 万元

发出天气坏预报的后验完全信息价值

后验 $EVPI(e_2) = 1\,000 \times 0.176\,5 + 0 \times 0.823\,5 = 176.5$ 万元

(6) 补充信息价值的期望值 $EVAI = 86.53$ 万元

(7) 为提高决策的效益,该水利工程公司应购买气象情报,以便更准确地把握气象水文状况。只要该项情报要价低于 86.53 万元,平均来看购买情报是有利的。如果气象预报天气好,则应在 9 月份施工,如果气象预报天气坏,则应在 10 月份施工。

第十章

一、选择题：

1.c 2.c 3.2 4.a 5.c

二、问答题：

1.拉氏 102.9%；帕氏 102.47%

4.110.28%

6.106.25%

8.工业总产量指数－104.87%；增加绝对量＝43.3 万件

工人人数变动影响：相对数 104.39%，绝对数 39 万件

工作日变动影响：相对数 96.15%，绝对数－35.7 万件

作时变动影响：相对数 93.28%，绝对数－60 万件

单位劳动生产率影响：相对数 112.01%，绝对数 100 万件

10.

(1)104.00%

(2)423.08；444.23；480.77；519.23

(3)312；328.64；374.4

(4)105.33%；120.00%；133.33%；141.03%；148.08%；160.26%；173.08%

12.

(1)4.74%

(2)拉氏：156.05%；帕氏：154.85%

第十一章

一、判断题：

1.× 2.× 3.√ 4.√ 5.×

2.综合分：7.512

第十二章

一、选择题：

1.d 2.c 3.c 4.a 5.d

二、判断题：

1.× 2.× 3.× 4.✓ 5.×

三、问答题：

GDP＝16 876；国内生产净值＝14 975.5；国民生产总值＝16 558；国民生产净值＝14 657.5；国民可支配总收入＝16 750；国民可支配净收入＝14 849.5。

主要参考书目

1 黄良文,曾五一.统计学原理.北京:中国统计出版社,2000

2 袁卫,庞皓,曾五一.统计学.北京:高等教育出版社,2000

3 陈珍珍,罗乐勤.统计学.厦门:厦门大学出版社,2002

4 庞皓,杨作廪.统计学(修订本).成都:西南财经大学出版社,2001

5 黄良文,陈仁恩.统计学原理(修订本).北京:中央广播电视大学出版社,1997

6 刘汉良.统计学教程.上海:上海财经大学出版社,1998

7 Terry Sincich 著.陈鹤琴、罗明安译.例解商务统计学.北京:清华大学出版社,2001

8 徐国祥,胡青友.统计预测和决策.上海:上海财经大学出版社,1998

9 苏为华.多指标综合评价理论与方法研究.北京:中国物价出版社,2001

10 钱伯海.国民经济统计学.北京:中国统计出版社,2000

11 朱之鑫主编.国家统计报表制度主要指标解释(第一版).北京:中国统计出版社,2001

12 Mark Dodge. Craig Stinson. Excel 2000 中文版使用大全.北京:清华大学出版社,2000

21世纪高等院校经济与管理核心课经典系列教材

工商管理专业

管理学

宏观经济学

微观经济学

市场营销学

统计学概论

经济法概论

会计学原理

财务管理

企业战略管理

企业运营管理

人力资源管理

管理信息系统