理工科大学

经济管理类 研究生教学用书

实用统计分析方法与技术

胡波 郭骊 编著

化学工业出版社

理工科大学经济管理类研究生教学用书

实用统计分析方法与技术

胡 波 郭 骊 编著

化学工业出版社

·北京·

本书内容涵盖了经济管理类研究生可能用到的绝大多数统计分析方法，如一元和多元线性回归分析、违背基本假定的回归分析、虚拟变量回归分析、逻辑回归分析、非参数检验、聚类分析、判别分析、主成分分析和因子分析、时间序列的随机分析、面板分析。同时，本书选择了 SPSS 和 EViews 两个软件作为例题讲解的配套软件，其中涉及时间序列的部分应用 EViews 软件，其他部分应用 SPSS 软件。

本书对统计分析方法的阐述简明扼要，而着重介绍各种方法如何结合经济管理中的实际问题加以应用，以及应用的条件和场合；同时配合每一种统计分析方法，均结合具体经济管理实际问题的例题，并在例题讲解时兼顾统计分析方法和操作实现两个方面，给出详细分析过程的同时，将软件应用的每一步骤均截图介绍，并给出操作依据及输出结果的解释。

本书可作为经管专业研究生课程教材，也可作为相关管理人员和决策人员的参考书。

图书在版编目（CIP）数据

实用统计分析方法与技术/胡波，郭骊编著. —北京：化学工业出版社，2012.10

理工科大学经济管理类研究生教学用书

ISBN 978-7-122-15323-4

Ⅰ.①实… Ⅱ.①胡…②郭… Ⅲ.①实用统计分析方法-高等学校-教材 Ⅳ.①C81

中国版本图书馆 CIP 数据核字（2012）第 215498 号

责任编辑：宋湘玲　　装帧设计：关　飞

责任校对：边　涛

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）

印　　刷：北京云浩印刷有限责任公司

装　　订：三河市宇新装订厂

787mm×1092mm　1/16　印张 15　字数 388 千字　　2013 年 1 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899

网　　址：http://www.cip.com.cn

凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：35.00 元

前　言

数据中包含信息，准确、有效获取数据中的信息已越来越为各学科、各行业所重视。经济现象的研究尤其如此，从国内外经济类学术刊物上，以统计方法作为主要分析方法的论文所占比例越来越高就可见一斑。随着我国社会主义市场经济体制的快速推进，经济现象以及经济变量的随机性已使在定性分析基础上，配合合理的统计定量分析，成为我国经济研究的一种大众化的实证研究方法。正是在这样的背景下，我们编写了这样一本经济管理类硕士研究生学习现代统计分析方法的教材。

在内容安排上，本书涵盖了经济管理类研究生可能用到的绝大多数统计分析方法，如一元和多元线性回归分析、违背基本假定的回归分析、虚拟变量回归分析、逻辑回归分析、非参数检验、聚类分析、判别分析、主成分分析和因子分析，作为经济现象研究越来越不可忽视的时间序列的随机分析，以及结合截面数据和时序数据的面板分析。

统计分析方法的应用离不开统计软件的支持，正是有了统计软件，才使得上述方法的应用成为现实。本书选择了 SPSS 和 EViews 两个软件作为例题讲解的配套软件，其中涉及时间序列的部分应用 EViews 软件，其他部分应用 SPSS 软件。

统计学是研究客观事物内在数量规律的学科，其中运用了大量的数学方法，同时与纯数学相比，又增加了与实际经济管理的结合问题，往往使学习者感到难以掌握。针对这一难题，本书在编写上，仅保留直接影响理解和应用的数学证明，对过于繁琐且不直接应用的数学过程只给出结论，从读者便于理解、应用的角度出发，在编写中力求深入浅出、通俗易懂，侧重统计方法在实践中的应用，为此本书对统计方法的阐述简明、扼要，而着重介绍各种方法如何结合经济管理中的实际问题加以应用，以及应用的条件和场合；同时配合每一种统计分析方法，均结合具体经济管理实际问题的例题，并在例题讲解时兼顾统计分析方法和操作实现两个方面，给出详细分析过程的同时，将软件应用的每一步骤均截图介绍，并给出操作依据及输出结果的解释。

本书配套有立体化教案，为选用本书的教师免费提供，如有需要请登录化学工业出版社教学资源网：www. cipedu. com. cn 下载或联系 1172741428@qq. com。

全书内容全面、结构完整，共十二章，第一至第八章由胡波编写，第九至第十二章由郭骊编写，胡波负责全书的总纂。

本书列入北京科技大学东凌经济管理学院“十二五”研究生教学用书规划。本书出版得到北京科技大学研究生教育发展基金资助，感谢北京科技大学研究生院、北京科技大学东凌经济管理学院、化学工业出版社对本书出版的大力支持！

限于编者水平，书中难免存在不当之处，欢迎读者批评、指正。

编著者

2012 年 7 月

目　录

第一章　概　　述

第一节　统计学的研究对象

统计学的研究对象是指统计研究所要认识的客体。只有明确了研究对象，才可能根据它的性质特点运用相应的研究方法，达到认识客观规律性的目的。

统计学的研究对象是客观现象的数量特征和数量关系，并通过这些数量方面阐明现象的本质及其内在规律性。

统计学着重于对客观现象的数量方面进行研究，也就是定量分析。定量分析是认识客观事物不可忽视的重要方面，它可以使我们更精确、更具体、更深刻地把握事物的性质、特征及其变化规律。例如，要了解一个企业的基本状况，就要从该企业的职工人数、投资规模，产品的数量、质量、品种，以及劳动生产率、成本、利润等数量方面来具体说明。但是，由于客观事物的质和量是密切联系的，所以，统计研究又离不开事物的质。统计学是在明确了事物质的规定性的基础上，来对其量进行分析研究的。例如，要了解和研究国内生产总值的数量、构成及变化，必须首先阐明国内生产总值这一概念的内涵，然后才能确定国内生产总值的计算范围和计算方法。同时，统计学对客观事物现象量的方面调查研究的最终目的，是为了更深入地阐明事物的本质及其内在规律性。因此，统计学在研究客观现象的数量方面时，必须要和其质的方面结合起来，即定量分析要和定性分析相结合。

统计学研究的是客观现象总体的数量方面。例如，要研究某企业职工的工资状况，不是了解个别职工工资状况，而是要反映该企业全体职工的工资水平及变化情况，其全体职工就组成一个总体。但是，对总体的数量方面进行研究又是从个别单位入手的。同时，统计学研究对象的总体性并不排斥对个别典型单位的深入研究，因为这种研究也是为了更有效地掌握总体现象的规律性。

第二节　统计数据类型

统计数据是统计工作活动过程中所取得的反映社会经济现象的数字资料以及与之相联系的其他资料的总称。统计研究客观事物的数量方面，离不开统计数据，统计数据是对客观现象进行计量的结果。统计数据从不同角度分类可以区分为很多类型，不同类型的数据在进行统计工作各个环节时处理手段是有区别的，因此，需要给统计数据进行适当分类。

一、按照计量尺度不同的分类

统计数据是采用某种计量尺度对事物进行计量的结果。以个体的人为例：由于人的特征和属性表现有的比较简单，是可见的，如头发颜色、个子高矮，有的则比较复杂，特征和属性是不可见的，如偏好和信仰；人的特征和属性有的表现为数量差异，差异可以直接用数字描述，如身高、体重，有的则表现为品质差异，差异不能直接用数字描述，如性别、民族。因此，统计计量也就有定性计量和定量计量的区别，并且可分不同的层次。美国社会学家、

统计学家史蒂文斯（S. S. Stevens）1968 年按照变量的性质和数学运算的功能特点，将统计计量尺度划分为四个层次。

1. 定类尺度

定类尺度又称类别尺度，是按照事物的某种属性对其进行平行的分类或分组，其数据表现为“类别”，但各类之间无法进行比较。例如，将国民经济按其经济类型，可以分为国有经济、集体经济、私营经济、个体经济等，可以按所属经济类型对经济单位进行分组，但每组之间的关系是平等的或并列的，没有等级之分。由于定类尺度各组间是平等或并列的关系，所以各组或各类之间是可以改变顺序的，哪一类放在前面没有数量上的要求，更多的是应用习惯所致，比如经济类型按照“先国有、再集体、再其他”的顺序。但是，在定类尺度中，各组或各类必须符合类别穷尽和互斥的要求，即组别或类别是可以通过列举的方式全部显示出来的，而且每一个数据只能归于其中一类，不应出现某一经济单位既属于国有经济又属于私营经济的现象。

2. 定序尺度

定序尺度又称顺序尺度，是对事物之间等级差别和顺序差别的一种测度。定序尺度也是对事物进行分类，但是不同于定类尺度的是，定序尺度不仅可以测度类别差，还可以测度次序差，即对事物分类的结果是有顺序的，如人可以根据年龄分为幼年、少年、青年、中年、老年；消费者满意度可分为非常满意、比较满意、满意、不满意等。

虽然定序尺度对事物分类的同时给出各类的顺序，但是其数据仍表现为“类别”，并且尽管各类之间是有序的，可以比较优劣，却并不能测量出类别之间的准确差值。

3. 定距尺度

定距尺度也称间隔尺度，是对事物类别或次序之间距离的计量，它通常使用自然或物理单位作为计量尺度。定距尺度是比定序尺度高一层次的计量尺度，它不仅能将事物区分为不同类型并进行排序，而且可以准确地指出类别之间的差距是多少。例如，在进行收入调查时，可以把调查者按一定收入差异区分为不同的组，如 1000～2000 元为较低收入者，3000～5000 元为中等收入者等。显然，定距尺度可以较方便地转换为定序尺度，而通常定序尺度数据不能转换为定距尺度数据。

定距尺度的计量结果表现为数值，可以进行加或减的运算，但却不能进行乘或除的运算，其原因是在等级序列中没有固定的、有确定意义的“零”位。例如，学生甲得分 90 分，学生乙得 0 分，可以说甲比乙多得 90 分，却不能说甲的成绩是乙的 90 倍或无穷大。因为“0”分在这里不是一个绝对的标准，并不意味着乙学生毫无知识；又如气温 40℃比 20℃高 20℃，但是不能说 40℃比 20℃暖和 2 倍。

4. 定比尺度

定比尺度是在定距尺度的基础上，确定可以作为比较的基数，将两种相关的数加以对比而形成新的相对数，用以反映现象的构成、比重、速度、密度等数量关系。由于定比尺度中的“0”表示没有，有绝对零点，或者说有理论上的极限，因此不仅可以进行加减运算，还可以进行乘除运算。由于它是在比较基数上形成的尺度，所以能够显示更加深刻的意义。如将一个国家或地区的国内生产总值与该国或地区的人口数对比，计算人均国内生产总值，可以反映国家或地区的综合经济能力。2009 年我国国内生产总值约占世界生产总值的 7.8%，排列世界第三位，堪称世界经济大国，但我国的人口占世界总人口的约 21%，如果按人均国内生产总值计算，在世界各国中又居于比较落后的位次，说明我国仍属于发展中国家。

上述四种计量尺度对事物的计量层次是由低级到高级、由粗略到精确逐步递进的。高层次的计量尺度具有低层次计量尺度的全部特性，反之则不成立。在统计分析中，一般要求测

量的层次越高越好，因为高层次的计量尺度包含更多的数学特性，所运用的统计分析方法越多，分析时也就越方便，因此应尽可能使用高层次的计量尺度。

采用不同的计量尺度会得到不同类型的统计数据。从上述四种计量尺度计量的结果来看，可以将统计数据分为对应的四种类型，即定类数据、定序数据、定距数据、定比数据。前两类数据说明的是事物的品质特征，不能直接用数字描述，其结果均表现为类别，也称为定性数据或品质数据（oualitative data）；后两类数据说明的是现象的数量特征，能够用数值来表现，因此也称为定量数据或数量数据（quantitative data）。由于定距尺度和定比尺度属于同一测度层次，所以可以把后两种数据看作是同一类数据，统称为定量数据或数值型数据。

二、按照时空维度不同的分类

1. 截面数据

同一时间（时期或时点）某个指标在不同空间的观测数据，称为截面数据。“不同的空间”可以是指不同的地理区域，也可以是指不同的描述对象，如行业、部门或个人。同一时间不同家庭的收入或者消费支出数据、某年各个省（直辖市、自治区）的国内生产总值等都是截面数据。

2. 时间序列数据

把反映某一事物特征和属性的数据，按照一定的时间顺序和时间间隔（如每日、月度、季度、年度）排列起来，这样的统计数据称为时间序列数据。例如，我国逐年的国内生产总值和消费支出数据、某地区逐月的物价指数等。时间序列数据可以是时期数据，也可以是时点数据。

3. 面板数据

面板数据（panel data）是截面数据和时间序列数据相结合的数据，如在居民收支调查中搜集的对各个固定调查户在不同时期的调查数据，又如全国各省、直辖市、自治区不同年份的经济发展状况的统计数据，都是面板数据。

第三节 统计学的分科

一、理论统计和应用统计

理论统计（theoretical statistics）是指统计学的数学原理，是应用纯逻辑推理的方法研究随机现象的数量规律性的科学。理论统计的基础是概率论，此外还包括抽样理论、参数估计和假设检验理论、实验设计、决策理论，以及随机过程等。

应用统计（applied statistics）是指应用统计学方法研究各领域客观现象的数量规律性。应用统计包括一整套统计学方法，如各领域通用的参数估计、假设检验、方差分析、相关与回归分析，以及专门适用于某一领域的分析方法，如经济统计学中的指数法等。应用统计涉及的领域很广，无论自然科学还是社会科学，都离不开统计分析方法，因此形成了各种应用统计学，如生物统计学、经济统计学、教育统计学等。

自 20 世纪 60 年代以来，统计学的发展出现了三个明显的趋势：一是越来越广泛的应用数学方法；二是形成了以统计学为基础的边缘学科，如经济计量学、工程统计学等；三是计算机及其软件的应用加快了统计的计算速度，扩大了统计学的应用范围。

二、统计学和数学的关系

统计学所研究的量是具体事物在一定时间、地点、条件下的数量表现，这是它和数

学的区别所在。数学所研究的量是脱离了具体对象的抽象数量关系，统计学所研究的是具体事物的数量方面。但是，统计学毕竟是研究客观事物的数量关系的，因此也要遵守数学原则，可以用数学模型来表现现象之间的数量关系。统计学的发展越来越依靠数学，其数量关系有的表现为函数关系，有的表现为随机性的统计关系，这就需要应用数理统计学进行分析。

三、应用统计在经济管理领域中的作用

现代经济与管理领域中正越来越多地运用统计学的分析研究方法，总的来说，统计学在经济与管理中的作用可归纳为反映作用、决策作用、控制作用和监督作用。

反映作用简单地说就是提供信息，是从统计学产生之日起就具有的一种基本作用，也是其他作用的基础。

决策作用是统计作用发展的产物，是根据统计反应和预测分析的信息及有关资料，运用一定的决策方法，为各级管理者提供最优的方案，然后在此基础上进行规划，确定各方面目标，用以指导当前和未来的活动。统计分析方法具有科学性，因此，在经济与管理中，正确运用统计分析方法往往可以提高决策的准确性，降低决策风险。例如，在宏观经济管理中，政府部门需要经常进行不同规模的统计调查，来了解当前的经济形势及未来的经济趋势，以此作为制定和修订经济政策的依据；在微观经济管理中，各级各类管理人员或多或少地需要应用统计分析方法，比如，决策层经常面对未来不确定条件做出选择，基层管理者更是要面临各种决策问题，在这些问题的解决中，适当应用统计方法，定性分析与定量分析相结合，往往起到事半功倍的效果。

控制作用是根据决策、规划所确定的各项目标，对预期可能发生或已经发生的进度和状态进行监测和对比，发现偏差并及时反馈矫正信息，以便在事前或事中进行调节，保证目标的实现。我国的经济景气监测预警系统，即是通过分析经济周期波动情况，及时发现非正常波动，解释其出现的原因，以尽可能避免出现经济的剧烈收缩和过快增长，影响国民经济的正常发展。

监督作用是以法律及有关法令、政策、计划和制度等为准绳，通过检查与分析，找出问题，提出对策，评价好坏，从而促进监督对象的良性循环。比如，我国国家质检部门为保证群众的健康，经常对市场上的食品质量进行抽样检验，对不合格食品的生产销售部门依法进行处罚。

第四节　统计学的研究方法

一、统计工作过程

统计是一种调查研究活动，是从质出发，经过对量的调查研究，达到对事物本质和规律的认识，是从量到质的认识过程的飞跃。一个完整的统计过程，可分为统计设计、统计调查、统计整理和统计分析四个阶段。

统计设计是根据统计的任务与目的和统计对象的特点，对统计工作的内容和程序做出通盘的考虑和安排，确定统计指标和统计指标体系、调查方案、汇总整理方案以及分析应用方案等。由于统计研究客观现象的数量关系是通过特定的指标和指标体系来完成的，因此，统计设计的主要任务是确定指标和指标体系，以便既能有效地反映研究对象，又能以较少的人力、物力去完成。只有搞好统计设计，才能使统计工作正常而有序地进行。

统计调查是根据统计设计所确定的指标体系，拟定调查纲要，搜集资料，这是认识事物

的起点，也是统计整理与分析的基础。统计调查的关键是取得准确的统计资料，如果统计调查失真，就会导致错误的分析结论。

统计整理是对调查资料加以汇总综合，使之系统化、条理化，便于进行统计分析。

统计分析是将加工整理好的统计资料加以分析研究，采用各种分析方法，及各种分析指标，揭示被研究对象的基本特征和发展的规律性，根据研究目的做出科学的判断和结论。统计分析是决定性环节，使统计资料发挥更大的作用。

二、统计研究方法

统计在调查、整理、分析的各个阶段，使用各种专门的统计方法，这些方法不论具体步骤如何，方法的核心性原理基本是一致的，主要包括大量观察法、统计分组法、统计指标法、统计模型法和归纳推断法等。

1. 大量观察法

某些个别现象通常有其特殊性和偶然性，而总体现象则具有相对的普遍性和确定性，是有规律可循的。因此，统计研究客观现象和过程的规律，是从现象总体加以考察，就总体中的全部或足够多数单位进行调查和综合分析，这种研究方法称为大量观察法。

大量观察法的数理根据是大数定律。大数定律的逻辑意义是：由偶然性因素发生作用而产生的随机现象也是具有规律性的，但这种规律性不表现在个体上，而是从总体上才表现出来。因为每个偶然因素对总体的影响都相对较小，通过大量观察法对各方面的综合平均，偶然因素的影响相互抵消，而显现出现象的稳定性质。所以必须采用大量观察法。

2. 统计分组法

根据研究对象内在的性质和统计研究的要求，将研究对象按照一个或几个特定对象分成若干组成部分，相同的归并在一起，不相同的区分开，这种通过分组来认识总体现象的方法称为统计分组法。例如：将人口按照性别划分为男性人口和女性人口两组；将企业按照所有制划分为国有、集体、私营等若干组。

3. 统计指标法

统计指标法是指运用各种统计指标来反映和研究客观现象总体的数量特征和数量关系的研究方法。统计指标法通过对大量的原始数据进行整理汇总，计算各种指标，可以表明现象在具体时间、地点、条件下的规模、水平和数量对比关系。在统计分析中，人们广泛运用各种指标来研究总体内部的各种数量关系，揭露矛盾，发现问题，进一步寻找解决问题的方法。

4. 统计模型法

统计模型法是根据一定的专业理论和假定条件，用数学方程或方程组去模拟客观现象相互关系的一种研究方法。利用统计模型可以进行数量依存关系及其发展变化的评估和预测。

统计模型包括变量、基本关系式和模型参数三个基本要素，将总体中一组相互联系的统计指标作为变量，其中有些变量被描述为其他变量的函数，这些变量称为因变量，它们所依存的其他变量称为自变量。现象的基本关系式通常用一组数学方程来表示，方程可以是线性的也可以是非线性的，可以是二维的也可以是多维的，模型参数表明方程式中自变量对因变量的影响程度，它由一组实际观察数据来确定。

5. 归纳推断法

在统计研究过程中，由观察总体中各单位的特征得出关于总体的某种信息，这种由个别到一般、从事实到概括的推理方法，称为逻辑归纳法。但是，常常存在这种情况，我们所观察的只是部分或者有限的单位，而所需要判断的总体范围却是大量的，甚至是无限的，这就

产生根据部分单位的资料判断总体数量特征的置信度问题。这种以一定的置信标准，根据部分单位数据来判断总体数量特征的归纳推理方法称为统计推断法。统计推断法是逻辑归纳法在统计推理中的应用，因此，也称为归纳推断法。

本书从第二章开始，根据研究对象数量特征的属性，选择经济管理领域常用的统计分析方法，从原理和应用两个方面进行介绍，具体包括截面数据、时序数据以及面板数据的回归分析，非参数检验，以及聚类、判别、主成分和因子分析等几种多元统计分析方法。

本教材的结构框架如图 1-1 所示。

图 1-1 教材结构框架

三、统计分析软件

统计分析软件有很多，如 SPSS，SAS，R，STAT，EViews 等，它们都可以方便地进行常规的统计分析，但是又各有特点。本书结合具体的统计分析方法，选择其中的两种软件配合介绍：SPSS 和 EViews。涉及时间因素的部分包括第七章时间序列分析和第八章面板数据分析应用 EViews 软件，其他部分应用 SPSS 软件。

1. SPSS 简介

SPSS（statistics package for social science for Windows，简记 SPSS）是 1968 年由美国斯坦福大学开发、运行在 Windows 系统下的社会科学统计软件软件包。SPSS 软件包集数据整理、分析过程、结果输出等功能为一体，采用窗口操作界面，统计分析方法涵盖面广，用户操作使用方便，输出数据表格图文并茂，并且随着它的功能不断完善，统计分析方法不断充实，问世几十年来，已经拥有全球数以万计的用户，分布在通信、医疗、银行、证券、保险、制造、商业、市场研究、科学教育等众多的行业领域，成为世界上应用最广泛的专业统计软件之一。

SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等，具体内容包括描述统计、列联分析，总体的均值比较、相关分析、回归模型分析、聚类分析、主成分分析、时间序列分析、非参数检验等多个大类，每个类中还有多个专项统计方法。SPSS 设有专门的绘图系统，可以根据使用者的需要将给出的数据绘制成各种图形，能够满足用户的不同需求。

2. EViews 简介

EViews（econometrics views，简记 EViews），直译为计量经济学观察，通常称为计量经济学软件包。EViews 的前身是 1981 年第 1 版的 Micro TSP，是美国 QMS 公司研制的在 Windows 下专门从事数据分析、回归分析和预测的工具。

EViews 处理的基本数据对象是时间序列，也能够处理截面数据以及面板数据。EViews 的基本功能包括：建立 ARMA 模型；执行普通最小二乘法、带有自回归校正的最小二乘法、两阶段最小二乘法和三阶段最小二乘法、非线性最小二乘法、广义矩估计法、ARCH 模型估计法等；进行协整检验、Granger 因果检验；对二元选择模型进行 Probit、Logit 和 Gompit 估计；估计和分析向量自回归系统等。

思 考 题

1. 按照不同的分类标准，如何对统计数据进行分类？各类别数据之间有什么明显不同？
2. 一个完整的统计工作过程大体包括哪几个阶段？每一阶段主要解决什么问题？

第二章　一元线性回归分析

回归分析是应用统计方法寻找一数学方程，建立自变量与因变量之间的关系，并据以利用自变量的给定值来推算或估计因变量的值。一元线性回归分析是描述两个变量之间相关关系的最简单的回归分析方法。一元线性回归虽然简单，但通过模型的建立过程，可以了解回归分析方法的基本思想及它在实际问题研究中的应用。本章将讨论一元线性回归分析的建模思想、参数估计、显著性检验，以及预测与控制的理论和应用。

第一节　回归分析概述

一、相关关系与回归分析

自然界和社会中的许多事物或现象，彼此之间都是有机地相互联系着、相互依赖着、相互制约着，离开周围的现象和条件而孤立存在的现象是没有的。因此，统计学不光要研究某一现象的数量变动规律，而且还要对事物间相互联系的数量关系加以探讨。

事物或现象之间的相互依存关系有两种：即函数关系和统计相关关系。函数关系是一种确定性的关系，即随着一个数值的变化，其他数值也发生着一定的对应变化。例如，通过具有一定电阻 R 的电路中的电流 I 与电路两端的电压 V 之间有 $I=\frac{V}{R}$，当 V 与 R 的数值一经确定之后，I 的数值也随之确定了。

另一种是统计相关关系，这是一种不完全确定的关系。例如，钢的化学成分与力学性能的关系，我们只能说某种元素的加入能使力学性能提高（或降低），而我们却不能绝对地说假如 a 千克某元素一定能使力学性能提高 b，因为还有许多偶然因素的影响，使得同样加 a 千克某元素可取得各种不同结果。所以相关关系是指现象之间确实存在的，但关系数值不固定的相互依存关系。相关关系虽然是不确定的，但从概率的意义上说，它也是具有规律性的。对相关关系进行研究就是通过对具有相关关系的变量的大量观察，排除随机性干扰，寻找这些变量之间的规律。

现象之间相关关系的分析是从两方面进行的：一方面研究现象之间变量关系的密切程度，直线相关用相关系数（correlation coefficient）表示，曲线相关用相关指数（correlation index）表示，这种研究称为相关分析（correlation analysis）；另一方面研究自变量和因变量之间的变动关系，并将这种关系用数学方程表达，称为回归分析（regression analysis）。

在相关关系中，在相互联系的现象之间，有时表现为因果关系，即一类为因变量（independent variable，用 Y 表示），另一类为自变量（dependent variable，用 X 表示），此时，我们可以通过回归分析来研究它们间的相关关系。例如，经济学理论中的消费支出与可支配收入之间密切相关，存在着因果关系，即可支配收入的变化往往是消费支出变化的原因。这时，不仅可以通过相关分析研究两者间的相关程度，而且可以通过回归分析研究两者间的具体依存关系。但是在另一些相关关系中，变量之间只存在相互联系，并不存在明显的因果关系。例如，每万元增加值的耗电量与工业增加值之间存在着一定的联系，但是在二者之间难以指出哪一个是原因、哪一个是结果，即哪一个是自变量、哪一个是因变量，在这种情况

下，主要根据研究目的而定。例如，为了研究在一定耗电量水平下，工业增加值可能是多少时，就把耗电量看作是自变量，而把工业增加值看作是因变量；为了研究在一定工业增加值下，可能耗用多少电，就把工业增加值看作自变量，而把耗电量看作因变量。

应该指出，不论在哪种情况下，作为研究现象与现象之间的关系，必须是真实的、具有内在联系的关系，而决不能是臆造的，或只不过是形式上的偶然巧合。因此，统计在研究相关关系时，应当根据有关的科学理论，通过观察和实验，在对现象进行深入分析的基础上建立这种联系，并且还要通过理论上与实际上的检验，只有这样，才能通过研究，得出有科学意义的结论。

二、回归分析的特点

对于回归分析来说，要确定哪个是自变量、哪个是因变量，如人的身高与体重的关系，以身高为自变量，则以体重为因变量；反之，若以体重为自变量，则以身高为因变量。但是有些现象的两个变量之间不能互换。例如，炉膛温度和出铁量，只能以炉膛温度为自变量，出铁量为因变量，分析炉膛温度对出铁量的影响，而反过来分析出铁量对炉膛温度的影响则没有意义。在回归分析中，要求因变量是随机变量，自变量是非随机变量，是给定的数值。

回归分析可分为线性回归（linear regression）与非线性回归（curvilinear regression），对线性回归与非线性回归的区分有两种理解：一种按回归变量本身是否线性，即是否一次式来划分，例如，$Y=\beta_0+\beta X_1+\beta_2 X_2+\beta_3 X_3+\varepsilon$ 为三元线性回归模型，而 $Y=\beta_0+\beta_1 X+\beta_2 X^2+\beta_3 X^3+\varepsilon$ 为一元三次非线性回归模型；二是按回归变量的参数即回归系数（regression coefficient）是否线性来划分，例如，上例两式都是线性方程，因为它们的回归系数 β_1、β_2、β_3 都是线性的（一次式），而 $Y=\beta_0 X^{\beta_1}+\varepsilon$ 是非线性回归，因 Y 不是两个参数 β_0、β_1 的线性函数，β_0 与 β_1 是用乘法和指数方法连在一起的。在应用研究中，常见到的是按变量是否一次性来划分线性与非线性回归方程，因此本书沿用这种观点。

在线性回归分析中，对一个因变量与一个自变量的回归称一元线性回归（linear regression），而一个因变量与多个自变量的回归称多元线性回归（multiple Linear regression）。本章讨论一元线性回归。

第二节　一元线性回归模型的参数估计

如果随机变量 Y 随自变量 X 的变化而变化，且呈简单线性关系，则 Y 依 X 变化的规律可用一元线性回归方程表示。由于随机因素的干扰，Y 与 X 的线性关系中包含随机误差项 ε，即有一元线性回归模型：

$$Y=\beta_0+\beta_1 X+\varepsilon \tag{2-1}$$

式中，β_0、β_1 为待估的模型参数。

一、普通最小二乘估计

统计学对模型参数进行估计的方法很多，应用最广泛的是普通最小二乘估计（ordinary lease square estimation，OLSE），为保证普通最小二乘估计量 $\hat{\beta}_0$，$\hat{\beta}_1$ 的优良性质，对模型有如下主要约束条件。

条件 1　自变量 X 是确定性变量，不是随机变量。

条件 2　随机误差项 ε 具有 0 均值、同方差及序列不相关性，即

$$E(\varepsilon_i)=0 \qquad i=1,2,\cdots,n$$

$$\mathrm{Var}(\varepsilon_i)=\sigma^2 \qquad i=1,2,\cdots,n$$

$$\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=0 \quad i\neq j \quad i,j=1,2,\cdots,n$$

条件 3 随机误差项 ε 服从正态分布。即

$$\varepsilon_i\sim N(0,\sigma^2) \qquad i=1,2,\cdots,n$$

严格地说，第三个条件如果不满足，并不影响普通最小二乘估计量的优良性，但是为了接下来对模型进行检验顺利进行，需要满足随机误差项 ε 服从正态分布。

上述条件称为经典假设条件，满足经典假设条件，有 $\varepsilon\sim N(0,\sigma^2)$，即 $E(\varepsilon)=0$，则对于给定的 X，各次 Y 值会有所波动。但平均说来，应有：

$$E(Y)=\beta_0+\beta_1 X_i \tag{2-2}$$

这就是总体回归直线方程，β_0 为截距，β_1 为回归系数。一般来说，我们是从总体中抽取部分单位来观察 Y 依 X 的变化规律，所以我们通过样本观察值求出样本回归直线方程，用它对总体回归情况进行估计：

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_i \tag{2-3}$$

下面举例说明普通最小二乘估计方法的原理和步骤。

【例 2-1】 钢铁工业固定资产投资总额与钢产量之间有较密切的关系。现将某钢铁公司 2001～2010 年的有关资料列于表 2-1，据此进行回归分析。

表 2-1 某钢铁公司固定资产投资总额及钢产量资料

年份/年	固定资产投资总额/万元	钢产量/万吨
2001	9232	52.20
2002	13193	56.28
2003	15888	59.43
2004	14726	61.59
2005	12746	66.35
2006	16339	71.00
2007	21789	77.94
2008	32743	85.56
2009	45431	95.61
2010	57675	98.36

【解】 以固定资产投资总额各值为自变量（X_i），以钢产量各值为因变量（Y_i），画出散点图（图 2-1）：

从散点图中可以看出各点的分布大概呈一直线趋势，所以我们用线性方程式 $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_i$ 来表示 Y 与 X 之间的关系。方程式中 $\hat{\beta}_0$、$\hat{\beta}_1$ 是模型参数 β_0、β_1 的估计量，需要根据样本数据计算其估计值。

计算 $\hat{\beta}_0$、$\hat{\beta}_1$ 估计值可有不同的方法，使用最多的普通最小二乘估计就是通过要求各散点到回归线的距离平方和最小来求得 $\hat{\beta}_0$、$\hat{\beta}_1$ 估计值，从而确定回归线，这时所求的回归线是样本数据的最适线，即：

$$Q=\sum(Y_i-\hat{Y}_i)^2=\text{最小值} \tag{2-4}$$

将回归方程 $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_i$ 带入式(2-4) 有：

$$Q=\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i)^2 \tag{2-5}$$

求 Q 对 $\hat{\beta}_0$、$\hat{\beta}_1$ 的偏导数并令其为 0，即：

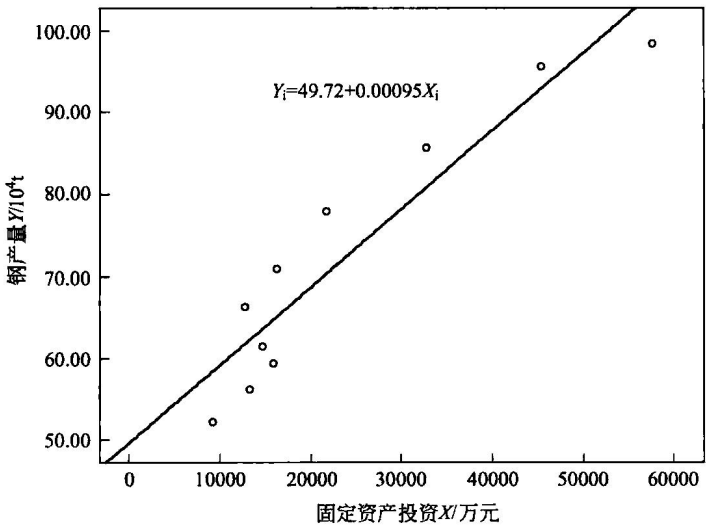

图 2-1　固定资产投资总额与钢产量的散点图

$$\begin{cases}\dfrac{\partial Q}{\partial \hat{\beta}_0}=-2\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0\\[2ex]\dfrac{\partial Q}{\partial \hat{\beta}_1}=2\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)(-X_i)=0\end{cases}\tag{2-6}$$

从以上联立方程组解出$\hat{\beta}_0$和$\hat{\beta}_1$：

$$\begin{cases}\hat{\beta}_1=\dfrac{n\sum X_iY_i-(\sum X_i)(\sum Y_i)}{n\sum X_i{}^2-(\sum X_i)^2}\\[2ex]\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}\end{cases}\tag{2-7}$$

若将式(2-7) 中$\hat{\beta}_1$式的分子项、分母项分别除以 n，则

$\hat{\beta}_1$式分子项：

$$\frac{n\sum X_iY_i-(\sum X_i)(\sum Y_i)}{n}$$
$$=\sum X_iY_i-n\overline{X}\,\overline{Y}$$
$$=\sum X_iY_i-n\overline{X}\,\overline{Y}-n\overline{X}\,\overline{Y}+n\overline{X}\,\overline{Y}$$
$$=\sum X_iY_i-n\overline{X}\frac{\sum Y_i}{n}-n\overline{Y}\frac{\sum X_i}{n}+n\overline{X}\,\overline{Y}$$
$$=\sum X_iY_i-\overline{X}\sum Y_i-\overline{Y}\sum X_i+n\overline{X}\,\overline{Y}$$
$$=\sum(X_iY_i-\overline{X}Y_i-\overline{Y}X_i+\overline{X}\,\overline{Y})$$
$$=\sum(X_i-\overline{X})(Y_i-\overline{Y})=L_{XY}$$

$\hat{\beta}_1$式分母项：

$$\frac{n\sum X_i{}^2-(\sum X)^2}{n}$$
$$=\sum X_i^2-n\overline{X}^2$$
$$=\sum X_i^2-n\overline{X}^2-n\overline{X}^2+n\overline{X}^2$$