TONGJIXUE JIAOCHENG

21世纪远程教育精品教材·经济与管理系列

统计学教程

（第二版）

主编　金勇进

中国人民大学出版社

21世纪远程教育精品教材·经济与管理系列

统计学教程

（第二版）

主编　金勇进

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

统计学教程（第二版）/金勇进主编
北京：中国人民大学出版社，2010
21世纪远程教育精品教材·经济与管理系列
ISBN 978-7-300-12690-6

Ⅰ.①统…
Ⅱ.①金…
Ⅲ.①统计学-远距离教育-教材
Ⅳ.①C8

中国版本图书馆CIP数据核字（2010）第178989号

21世纪远程教育精品教材·经济与管理系列
统计学教程（第二版）
主编　金勇进

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511398（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京东方圣雅印刷有限公司		
规　　格	170 mm×228 mm　16开本	**版　　次**	2004年11月第1版
			2010年10月第2版
印　　张	18.25	**印　　次**	2010年10月第1次印刷
字　　数	333 000	**定　　价**	35.00元

目录

原书缺页

是顺序型数据，是指数据不仅是分类的，而且类别是有序的。例如，满意度调查中的选项有“非常满意”、“比较满意”、“比较不满意”、“非常不满意”等。在这三类数据中，数值型数据由于说明了事物的数量特征，因此可称为定量数据；分类型数据和顺序型数据由于定义了事物所属的类别，说明了事物的品质特征，因此可称为定性数据。区分数据的类型非常重要，这是因为不同类型的数据在不同情况下，需要用不同的统计方法进行处理。

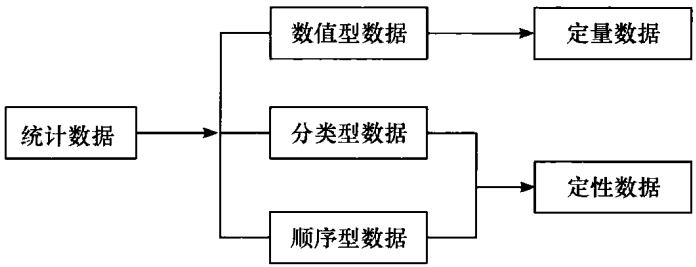

图 1—1　统计数据的分类

除了上述分类方法外，还可以从其他角度对统计数据进行划分。

按照收集方法的不同，可将统计数据分为观测数据和实验数据两类。观测数据指的是在没有对事物进行人为控制的条件下，通过调查或观测而收集到的数据，主要集中在社会经济领域。例如，对商品零售价格变动水平的测量可以得到商品零售价格指数、对股票价格变动水平的测量可以得到股票价格指数。实验数据指的是通过在实验中控制实验对象而收集到的数据，主要集中在自然科学领域。例如，某种新型电池的使用寿命、一种新型降压药疗效的实验数据等。

按照是否与时间相联系，可将统计数据分成截面数据和时间序列数据。表 1—1 中的数据为截面数据，描述了现象在某一时刻的变化情况；表 1—2 中的数据则为时间序列数据，描述了现象随时间变化的情况。

表 1—1　　2008 年我国部分地区国内生产总值及其构成数据　　单位：亿元

地区	国内生产总值	第一产业	第二产业	第三产业
北京	10 488.03	12.81	2 693.15	7 682.07
天津	6 354.38	122.58	3 821.07	2 410.73
河北	16 188.61	2 034.60	8 777.42	5 376.59
山西	6 938.73	302.48	4 265.77	2 370.48
内蒙古	7 761.80	906.98	4 271.03	2 583.79

表 1—2　　2001 年～2008 年我国国内生产总值及其构成数据　　单位：亿元

年份	国内生产总值	第一产业	第二产业	第三产业
2001	109 655.2	15 781.3	49 512.3	44 361.6
2002	120 332.7	16 537.0	53 896.8	49 898.9
2003	135 822.7	17 381.7	62 436.3	56 004.7
2004	159 878.3	21 412.7	73 904.3	64 561.3
2005	183 217.5	22 420.0	87 364.6	73 432.9
2006	211 923.4	24 040.0	103 162.0	84 721.4
2007	249 529.8	28 095.0	121 381.3	100 053.5
2008	300 670.0	34 000.0	146 183.4	120 486.6

资料来源：《中国统计年鉴（2009）》。

二、统计学

统计学是一门收集、整理和分析数据的科学，只要有数据的地方，就有统计学的应用。收集数据并研究如何得到数据，与之对应的是统计学中的抽样调查和试验设计等理论；整理数据指的是将数据用图或表的形式展现出来，与之对应的是统计学中的描述统计的方法；分析数据指的是选择适当的统计方法研究数据，并从数据中提取有用信息进而得出结论，与之对应的是统计学中推断统计的理论与方法，包含了参数估计、假设检验、相关分析、回归分析、时间序列分析等诸多内容。

统计学的应用领域非常广泛，它几乎是所有学科领域的通用数据分析方法，无论是学术研究、政府管理，还是公司或企业的生产经营管理，都离不开统计学的应用。如表 1—3 所示，列出了统计学的一些应用领域。

表 1—3　　统计学的应用领域

Actuarial work（精算）	Hydrology（水文学）
Agriculture（农业）	Industry（工业）
Animal science（动物学）	Linguistics（语言学）
Anthropology（人类学）	Literature（文学）
Archaeology（考古学）	Manpower planning（劳动力计划）
Auditing（审计学）	Management science（管理科学）
Crystallography（晶体学）	Marketing（市场营销学）
Demography（人口统计学）	Medical diagnosis（医学诊断）
Dentistry（牙医学）	Meteorology（气象学）
Ecology（生态学）	Military science（军事科学）
Econometrics（经济计量学）	Nuclear material safeguards（核材料安全管理）
Education（教育学）	
Election forecasting and projection（选举预测和策划）	Ophthalmology（眼科学）
	Pharmaceutics（制药学）
Engineering（工程学）	Physics（物理学）
Epidemiology（流行病学）	Political science（政治学）

原书缺页

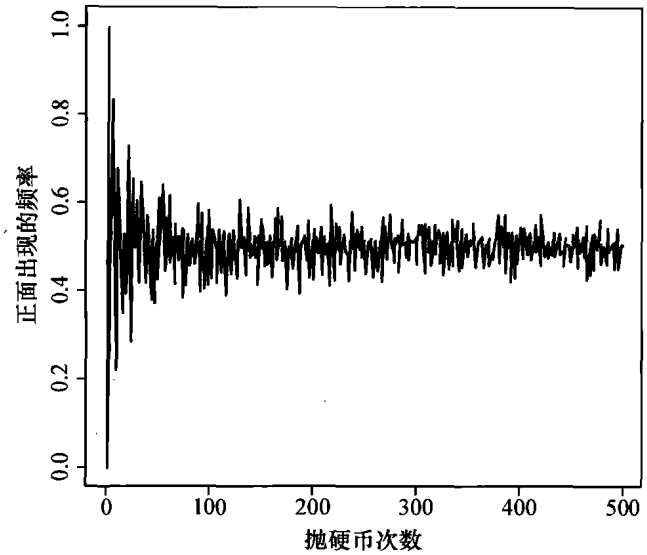

图 1—2　抛一枚硬币，正面出现频率的变化趋势

和决策。例如，要比较甲、乙两所高中的英语教学水平，分别在两所学校各随机抽取 30 名学生进行试卷测试。一般情况下两个学校学生的成绩会有差异，这种差异有可能来自于抽取到水平不同的学生，也可能来自于两所高中英语教学水平的不同。在很多情况下，我们会面临不同背景的观测数据，需要研究这两组不同背景的数据是否来自于同一随机现象，即需要研究两组数据之间的差异是否大到超过随机性本身所能解释的地步，这种研究的理论是概率论。

二、概率和机会

概率是对机会的描述，度量了某件事情发生的可能性，其取值在 0 和 1 之间（也可能是 0 或 1）。概率为 0 对应了那些绝对不可能发生的事情，比如在标准大气压下，水加热到 60℃就沸腾，这是不可能的。概率为 1 对应了那些一定会发生的事情，比如在真空状态下，自由落体在经过 t 秒钟后，落下的距离 s 必定是 $gt^2/2$。现实生活中，概率为 0 或 1 的事件都比较少，绝大部分事件发生的概率都介于 0 和 1 之间，为随机事件，这样的例子有：

（1）随意抛掷一颗骰子，出现的点数为 6。

（2）A、B 两队进行一场足球比赛，比赛结果为平局。

（3）在某交易日，上证指数以红盘报收。

（4）某对新婚夫妇生下的是一名男孩。

（5）某天出现雷雨天气。

通过对这些随机事件以概率的形式进行表述，可以为进一步的统计推断提供基础。

三、参数和统计量

为研究某一问题，需要对研究对象进行界定。在统计学中，将包含了所要研究的全部个体（数据的集合）称为总体，其特征的一些概括性数字度量称为参数。例如，某一地区的平均受教育年限、一批袋装牛奶的合格品率等。

作为总体特征概括性数字度量，参数的种类可以有很多，但研究者通常关心的主要有以下几种：总体平均数（记为 μ）；总体方差（记为 σ^2）；总体比例（记为 π）等。

现实中，研究对象的范围虽然容易界定（有时候也不容易），但是要把这些研究对象相关特征的数据全部收集到却通常面临着很大的困难，涉及时间、人力、物力、财力等诸多因素。例如，想要知道全国的总人口数，若逐个进行统计，实施普查，花费的时间和费用是惊人的，现阶段只能是每十年进行一次人口普查。有时候，对应的研究对象还可能是无限总体，即该总体所包括的元素是无限、不可数的，此时收集总体的全部数据变得根本不可行。例如，在科学试验中，每一个试验数据可以看作总体的一个元素，而试验可以无限地进行下去。为此，很自然的一个想法是不必去收集总体的全部数据，而是从总体中随机抽取一小部分元素的集合作为样本，根据样本提供的信息来推断总体的特征。例如，在我国每隔五年进行一次百分之一的人口调查；从出厂的某批次灯泡中随机抽出少量的几个检测其寿命等。

与参数相对应，用来描述样本特征的概括性数字度量为统计量。它根据样本数据计算得出，是样本的函数。常用的统计量和参数类似，主要有样本平均数（记为 $\bar{x}$）、样本方差（记为 s^2）、样本比例（记为 p）等。

统计学中的绝大多数问题都是研究如何根据统计量去推断参数的问题。例如，如何用样本平均数（$\bar{x}$）去估计总体平均数（μ），如何用样本方差（s^2）去估计总体方差（σ^2），如何用样本比例（p）去估计总体比例（π）等。图 1—3 形象地展示了这一过程。

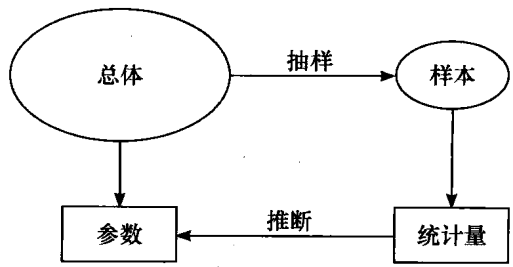

图 1—3　统计问题的分析过程

四、变量

（一）变量及其类型

变量与常量（也叫常数）相对，是说明随机现象某种特征的概念。例如，某个小学每天上课学生的人数是不同的，因为有些人生病，还有一些人因为其他原因没有来，每天上课学生人数就是一个变量，但学校登记注册的学生人数则是常量，是一个固定的已知数目。

事实上，所有随机取值的数据都归属于某个变量，是变量的某一次具体实现。数据可划分为定性数据和定量数据两类，同理变量的类型也分为定性变量和定量变量。性别、受教育程度等，都是定性变量；而商品销售额、职工工资等，则都属于定量变量。统计学所面对的是变量和变量之间的关系。

（二）变量之间的关系

事物是普遍联系的，作为随机现象某种特征的表达，变量与变量之间同样存在着千丝万缕的联系。与区分数据的类型相似，变量同样需要区分不同的类型之间的关系，因为研究不同变量类型之间的关系，对应着不同种类的统计模型。

按照变量所属类型的不同组合，可以将变量之间的关系区分为定性变量之间的关系、定量变量之间的关系和定性与定量变量之间的关系。

性别与文化程度是否有关，不同国家的人对陌生人的态度倾向是否存在差异，居民家庭订阅报纸和开通宽带上网之间是否有联系等，这些都属于定性变量之间的关系。研究定性变量之间关系的统计模型与方法主要有列联分析、对数线性模型等。

广告投入是否会影响销售额，城镇居民人均收入和人均支出相互影响有多大，复习时间和考试成绩之间是否存在必然联系，这些都是定量变量之间的关系。研究定量变量之间关系的统计模型与方法主要有线性回归、非线性回归等。

手机品牌和手机销售量之间的关系，是否违约和信用卡用户年龄、月收入之间的关系，上市公司所属行业与其八大基本财务指标之间的关系等，都可归为定性变量与定量变量之间的关系。研究此类关系的统计模型与方法主要有方差分析、逻辑回归、判别分析等。

上述统计模型与方法，有些是较为基础的内容，会在本书的相关章节予以介绍，比如列联分析、方差分析和回归分析；有些则属于更深层次的内容，请参阅相关专业统计书籍。

第三节　统计软件

随着科技的飞速发展和计算机的普及，原本枯燥、庞大的统计计算等工作，如今都可以通过统计软件由计算机完成了，这为统计应用的普及提供了良好的条件。

统计软件的种类很多，我们在这里介绍常见的几种。

（1）Excel。

虽然严格说来，Excel 并不是一款统计软件，但它自带了一些统计计算功能。在 Excel 中设计了种类十分齐全的统计函数，并且通过加载宏安装数据分析的功能，能够实现一些诸如方差分析、线性回归等简单的统计功能。

（2）SPSS。

SPSS 的全称是 Statistical Product and Service Solutions，即统计产品与服务解决方案，是目前非常受欢迎的一款统计软件。它囊括了各种成熟的统计方法与模型并提供了各种数据准备与整理技术。

（3）SAS。

SAS 系统的全称为 Statistical Analysis System。在数据处理和统计分析领域，SAS 系统也是一款权威统计软件。SAS 系统以编程为主，在编程操作时需要用户最好对所使用的统计方法有较清楚的了解，非统计专业人员掌握起来较为困难。

（4）S-plus。

S-plus 是由美国 MathSoft 公司开发的一种基于 S 语言的统计软件，是世界上公认的三大统计软件之一，主要用于数据挖掘、统计分析和统计作图等。该软件最大的特点在于它可以交互地从各方面发现数据中的信息，并可以很容易地实现一个新的统计方法，兼容性好。

（5）R。

R 是一款国际自由统计软件，由一群致力于推广统计应用的志愿者组织管理。它完全免费，其统计功能的实现源自不断加入的由各个研究方向的统计学家编写的软件包，是目前更新速度最快的软件。R 同样需要编程，但与 SPSS 和 SAS 中的编程语言相比，R 语言是彻底面向对象的统计编程语言，十分简洁和高效。R 的官方网站是 http://www.r-project.org，在这个网站上可以下载到各种程序包及相关资料。

（6）Eviews。

在时间序列数据的分析和处理上，Eviews 是一款非常专业的软件，擅长

于多种常用的计量经济模型。它通过建立序列间的统计关系式，实现预测和模拟等功能。该软件在科学数据分析与评价、金融分析、经济预测、销售预测和成本分析等领域应用广泛。

值得注意的是，统计软件的使用必须建立在熟悉相关统计理论与方法的基础上，否则容易导致误用。而学习统计软件的最好方式是需要时在使用中学习，并多看帮助和说明。由于 Excel 是 Office 常用软件，普及较广，本书后面将以 Excel 为例介绍各种统计方法在统计软件上的实现。

本章小结

统计学是一门收集、整理和分析数据的学科，正因为如此，统计学和统计数据密不可分。本章从现实中的统计数据出发，讲述了统计学和统计数据的概念以及它们之间的关系。统计学中一些基本概念包括随机性和规律性、概率和机会、参数和统计量、变量等。这里既有统计思想，如随机性和规律性的关系；也有统计术语，如参数、统计量、变量等，这些思想和术语将贯通全书。

现代统计应用与统计软件紧密相连，统计方法需要利用统计软件实现。本章介绍了常用的几种统计软件，在平时的使用中，结合帮助和说明熟练掌握软件的使用以及功能。

思考与练习

1. 什么是统计学？怎样理解统计学与统计数据之间的关系？

2. 统计数据可分为哪几种类型？不同类型的数据各有什么特点？

3. 指出下面数据的类型：

(1) 体重；

(2) 民族；

(3) 空调销量；

(4) 购买商品时的支付方式（现金、信用卡）；

(5) 学生对某教学改革措施的态度（赞成、中立、反对）。

4. 一项调查表明，北京市大学生每学期在网上购物的平均花费是 500 元，他们选择在网上购物的主要原因是“价格实惠”。试问：

(1)“大学生在网上购物的原因”是分类型变量、顺序型变量还是数值型变量？

(2) 在这个问题中，总体参数指的是什么？

(3)“北京市大学生每学期在网上购物的平均花费是 500 元”是参数还是统计量？

5. 一家研究机构从IT从业者中随机抽取800人作为样本进行调查，其中70%回答他们的月收入在5 000元以上，40%的人回答他们的消费支付方式是信用卡。试问：

（1）月收入和消费支付方式分别属于哪一种变量，分类型、顺序型还是数值型？

（2）这一研究涉及的是截面数据还是时间序列数据？

6. 举例说明总体、样本、参数、统计量、变量这几个概念。

7. 为什么要定义数据和变量的类型？为什么对变量之间的关系进行划分？

8. 试举出一个日常生活发生的，反映随机之中又有规律性的例子。

第二章　数据的搜集

学习导航

- 统计数据的两种不同来源：直接来源和间接来源
- 数据误差的种类
- 抽样误差的含义与影响因素
- 数据文件的一般格式

在当今信息爆炸的时代，每天翻开报纸、打开电视或浏览网页，就会有铺天盖地的数据迎面而来。例如，股票行情、物价指数、外汇汇率、离婚率、房价、流行病等的相关数据，当然还有国家统计局定期发布的各种宏观经济数据、海关发布的进出口贸易数据等。从这些数据中，我们可以提取对自己有用的信息。此外，人们购买住房是喜欢大户型还是小户型，学习成绩的好坏与性别是否有关，公众在购买汽车时，倾向于选择国内品牌还是国外品牌？这些都是我们感兴趣却又不知道答案的问题。为了回答这些问题，需要搜集相关的数据进行分析。换句话说，当研究的问题确定之后，就要考虑为进行研究所需要的数据，包括我们从哪里获得数据？谁为我们提供数据？如果需要调查，调查对象是谁？如何从众多的潜在被调查者中抽样？怎样实施调查？还有一些研究问题可能需要通过实验的方法获得数据，怎样使用实验方法获得数据等。

一位睿智的统计学家说过，世上有两种数据：好数据和坏数据。好数据是指根据合理、正确的统计原理收集得到的数据；坏数据是指通过其他方法收集得到的数据。我们所得到数据的准确性如何？如果不准确，误差是怎样产生的？应当如何控制误差以获得高质量的数据？这些工作都是统计研究活动所不可缺少的环节。本章将对上述有关问题加以讨论。

第一节　数据的来源

从统计数据本身的来源看，它最初都是来源于直接的调查或实验。但是，从使用者的角度看，数据则主要来源于两种渠道：一是来源于直接的调查和科学实验，这是数据的直接来源，我们称之为第一手数据或直接数据；二是来源于其他人的调查或实验的数据，这是数据的间接来源，我们称之为第二手数据或间接数据。本节将从使用者的角度，对获取数据的这两种渠道分别加以介绍。

一、数据的直接来源

数据的直接来源主要有两种渠道：一是调查（或观察）；二是实验。调查是取得社会经济数据的重要手段，其中包括政府统计部门进行的调查，如经济普查、人口普查；其他部门或机构为特定的目的而进行的调查，如市场调查等。实验则是取得自然科学数据的主要手段。我们把通过调查方法获得的数据称为调查数据，把通过实验方法得到的数据称为实验数据。

（一）调查及调查数据

调查通常是对社会现象而言的。例如，经济学家们通过搜集经济现象的数据来分析经济形势、某种经济现象的发展趋势、经济现象之间的相互联系和影响；心理学家们通过搜集有关人心理测试的数据，以了解人的心理及其行为；管理学家们通过搜集生产、经营、销售等各方面的数据，分析整个企业运行的状况。调查数据通常取自有限总体，即总体所包含的个体单位是有限的。调查包括以下三种类型：普查、抽样调查和统计报表。

1. 普查（census）

普查是为某一特定目的而专门组织的一次性全面调查，如我国定期进行的人口普查、农业普查、经济普查等。

世界各国一般都定期进行各种普查，以便掌握有关国情、国力的基本统计数据。普查是适用于特定目的、特定对象的一种调查方式，它主要用于搜集处于某一时间点状态上的、不能够或者不适宜定期用全面统计报表搜集的社会经济现象的数据，目的是掌握特定社会经济现象的基本全貌，为国家制定有关政策或措施提供依据。