高等院校应用型人才培养规划教材——统计学类

GAODENG YUANXIAO YINGYONGXING RENCAI PEIYANG GUIHUA JIAOCAI TONGJIXUELEI

# 统计学

# TONGJIXUE

王丙参 刘佩莉 魏艳华 牛晓霞 编著

西南交通大学出版社

高等院校应用型人才培养规划教材——统计学类

# 统计学

王丙参　刘佩莉　魏艳华　牛晓霞　**编著**

西南交通大学出版社

·成　都·

## 内容简介

本书由浅入深，全面、系统地阐述了统计学的基本概念、原理和方法，并结合 Excel 和 SPSS 进行实例分析，读者只需具备高等数学基础即可读懂. 全书共 13 章，第 1～6 章为描述统计，主要有：统计学基本轮廓，数据的收集、整理与显示，数据分布的描述与分析，数据变换，统计指数；第 7 章为概率论，它是统计学的理论基础；第 8～13 章为推断统计，分别为统计量及其抽样分布，参数估计，假设检验，方差分析，相关与回归分析，时间序列分析.

本教材以“培养应用型人才”为最终目标，启迪人生智慧，系统论述基本理论并跟踪学科的发展前沿，力争将社会经济统计与数理统计熔于一炉，侧重理论的软件实现，通过典型案例进行教学，读者可直接以案例为模板处理实际问题.

本书可作为普通高等院校统计类、经济类、管理类各专业的统计学课程教材，也可作为统计工作者及经济管理工作人员的自学、参考用书.

**图书在版编目（CIP）数据**

统计学 / 王丙参等编著. —成都：西南交通大学出版社，2015.2
高等院校应用型人才培养规划教材. 统计学类
ISBN 978-7-5643-3635-6

Ⅰ. ①统… Ⅱ. ①王… Ⅲ. ①统计学－高等学校－教材 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字（2015）第 004893 号

高等院校应用型人才培养规划教材 ——统计学类

**统计学** | 编著 王丙参 刘佩莉 魏艳华 牛晓霞 | 责任编辑 张宝华 封面设计 何东琳设计工作室

**印张** 19.5　　**字数** 487千
**成品尺寸** 185 mm × 260 mm
**版本** 2015年2月第1版
**印次** 2015年2月第1次
**印刷** 四川煤田地质制图印刷厂

**出版 发行** 西南交通大学出版社
**网址** http://www.xnjdcbs.com
**地址** 四川省成都市金牛区交大路146号
**邮政编码** 610031
**发行部电话** 028-87600564　028-87600533

**书号：**ISBN 978-7-5643-3635-6
**定价：**38.00元

课件咨询电话：028-87600533
图书如有印装质量问题　本社负责退换

# 前　言

在信息经济时代，人们所依赖的不只是信息处理手段的先进与科学，更重要的是信息资料的准确收集与可靠分析，而统计学就是通过收集、整理、分析数据等手段，以达到推断所测对象的本质，预测对象未来的一门综合性科学. 所以，我国教育部将“统计学”作为高等院校统计类、经济类、管理类各专业的学科基础课程，作为研究客观社会经济现象总体数量特征和发展规律的实用性方法论学科. 统计学以丰富的背景、巧妙的思维和有趣的结论吸引着读者，使学生在浓厚的兴趣下学习和掌握基本的概念、理论和方法. 在多年教学实践中，作者深刻体会到，目前的统计学教材概念模糊，理论不系统，多偏向于社会经济统计. 为了满足普通本科院校统计专业、经管专业对统计学及培养应用型人才的要求，作者在本课程讲稿的基础上，结合同行专家的优秀成果及作者对本课程的研究，经过多次修订和补充编写了本书，力争将社会经济统计与数理统计熔于一炉.

本书科学地借鉴并整合了传统教材中的精华，力求反映现代统计学的新内容、新特点. 在编写过程中，我们博采百家之长，注重基本理论、概念、方法的叙述，坚持抽象概念形象化，关注应用能力、解题能力的培养，读者只需有高等数学基础即可读懂本书. 由于例题、习题只是加深读者对理论的理解，而不是本书的核心，故本书例题、习题在满足覆盖、巩固、加深理论的要求下尽量减少题量，希望读者把精力集中在对理论的理解上. 为保持独特之处，本书融合了编者最近对相关理论的研究、教学体会及人生体会，并理论联系实际，结合 Excel 与 SPSS 软件进行教学，提高学生运用所学的统计理论和统计方法进行数据收集、整理和分析的基本能力. 本书也参考和吸收了国内外统计学的优秀成果，从中获得了许多启发，采用了一些经典的例子和段落，在这里向这些材料的作者表示衷心的感谢！也非常感谢百度百科提供的平台，它为作者提供了大量帮助，同时也希望读者能正确、合理地利用搜索引擎. 请注意，书中人名均为化名，请勿对号入座.

全书共 13 章. 第 1 ~ 6 章为描述统计，包括统计学基本轮廓，数据的收集、整理与显示，数据分布的描述与分析，数据的变换与统计指数；第 7 章简要介绍概率论，它是统计学的理论基础，属于广义统计学；第 8 ~ 13 章为推断统计，分别为统计量及其抽样分布，参数估计，假设检验，方差分析，相关与回归分析，时间序列分析. 每个专题既有相关理论的简单讲解，也配有实用的案例分析，理论与实践相结合，案例既可以作为读者扩宽视野、提高分析水平的学习资料，也可直接作为模版应用于对实际问题的处理.

本教材初稿在我们内部使用多年，学生反映不错，定稿时，又做了较大的修改与补充，部分内容重写. 我们认为，教材内容要比教学大纲多一些，要比教师在课堂上讲授得多一些，这样能照顾到各类学校各个专业的需要，以满足不同程度学生的学习需要. 本书可作为普通高等院校统计类、经济类、管理类各专业统计学课程的教材，也可作为统计工作者及经济管理工作人员的自学、参考用书，可在 72 学时左右讲完. 若作为统计专业教材，可选择部分内容组织教学，主要侧重描述统计，大致 54 学时讲完. 任课老师可根据实际需要合适安排各章学时.

本书由天水师范学院数学与统计学院王丙参、魏艳华，商学院刘佩莉、牛晓霞共同编著，具体分工为：第 1, 3, 4, 7, 12 章由王丙参编写，第 2, 5, 10 章由魏艳华编写，第 6, 8 章及参考答案由牛晓霞编写，第 9, 11, 13 章及附录由刘佩莉编写. 我们经常讨论，切磋写法，选择例题，相互补充，经过反复讨论和修改后由王丙参定稿.

本书编写得到了天水师范学院中青年教师科研资助项目“Bootstrap 方法中的蒙特卡罗方差减少技术”（TSA1404）与天水师范学院重点建设学科“动态图像中的大数据处理”的资助，得到了学院领导的大力支持. 统计教研室同事为本书提供了很多宝贵的意见和建议，统计专业学生为本书提供了大量数据和题材，如 2009 统计文婉君、董惠玉等，也得到了西南交通大学出版有关各方和同仁的大力支持，特在此一并致以诚挚的谢意！

虽然我们希望编写出一本质量较高、适合当前教学实际需要的教材，但限于作者水平与撰写时间，因而作者搁笔之际，仍有许多创意尚未实施. 况且书中难免存在不妥之处，恳切希望读者批评、指正，使本教材不断得以完善. 为方便广大读者，作者提供支持电子邮箱：

wangbingcan2000@163.com.

读者可通过该邮箱与作者取得联系，获取技术支持和教学资料.

作　者

2014 年 4 月

联系方式：

王丙参：

电话：15809400501 email:wangbingcan2004@163.com QQ:19555965

通信地址：甘肃天水师范学院数学与统计学院

邮编：741001

# 目 录

# 1 导 论

在日常生活中，我们经常与“数”打交道，网络、电视、报纸上的数据无处不在，例如，房价上涨 30%、股票价格指数下跌 8%、就业率升高 6%等. 要使这些数据变为对你有用的信息，帮你决策，就需要对这些数据进行处理和分析. 统计学就是一套处理和分析数据的基本方法和技术，因此具备一些统计学知识是正确阅读并理解数据、图表等的基础. 在很多领域中，统计都有应用且成绩斐然，下面就是通过统计研究得到的一些结论：

（1）吸烟对健康有害，吸烟男性寿命减少 2 250 天；

（2）身材高的父母，其子女的身材也高；

（3）第一个出生的子女比第二个出生的子女聪明；

（4）上课坐在前面的学生，平均考试分数比坐在后面的学生高.

这些结论正确么？你相信么？

理解并掌握一些统计学知识对于普通大众是必要的，因为我们每天都关心生活中的一些事情，这其中就包含了很多统计知识. 比如，在外出旅游时，需要关心一段时间内的详细天气预报；在观看世界杯时，需要了解各球队的技术统计，等等. 理解并掌握一些统计学知识对于制定决策的人更为重要，在他们做出决策时，如果不懂统计，就可能闹出笑话，甚至损失巨大. 例如，在一次政府会议上，统计学者抱怨从其他部门收到的估计值没有给出标准误差，主管领导却问：“误差也有标准吗？”

在终极的分析中，一切知识都是历史；在抽象的意义下，一切科学都是数学；在理性的基础上，所有的判断都是统计学. 总有一天，统计思维会像读与写一样成为每一个有效率公民的必备能力.

本章主要介绍统计学的一些基本问题，包括统计学的含义、统计数据的分类等常用基本概念，最后给出 Excel 简明教程.

## 1.1 统计与统计学

很多人误认为“统计”与“统计学”是等价的，其实，两者的区别很大，侧重点不一样. 本节详细介绍了什么是统计与统计学，并给出了统计学的研究对象、分类及统计的应用领域.

### 1.1.1 统计与统计学的概念

人们在日常生活中经常接触“统计”一词，但很多人对此一知半解，甚至有很多误区. 统计一词起源于国情调查，最早意为国情学，历史悠久，可以说，自从有了国家就有了统计实践活动. 最初，统计也只是为了统治者管理国家的需要而搜集资料，弄清国家的人力、物力和财力，并作为国家管理的依据.

（1）中国：公元前 22 世纪的夏禹时代，中国分为九州，人口约 1 352 万，由此可见人口

统计的久远；《书经·禹贡篇》记述了九州的基本土地情况，被西方经济学家推崇为“统计学最早的萌芽”；西周建立了较为系统的统计报告制度；秦时《商君书》中提出“强国知十三数”，其中包括粮食储备、各国人数、农业生产资料及自然资源等.

（2）外国：公元前 27 世纪，埃及为了建造金字塔和大型农业灌溉系统，曾进行过全国人口和财产调查；公元前 15 世纪，犹太人为了战争的需要进行了男丁的调查；公元前约 6 世纪，罗马帝国规定每 5 年进行一次人口、土地、牲畜和家奴的调查，并以财产总额作为划分贫富等级和征丁课税的依据.

一般认为，统计学的学理研究始于古希腊的亚里士多德，迄今已有 2 300 多年的历史. 今天，“统计”一词已被人们赋予了多种含义，很难给出一个简单的确切定义. 人们给统计学下的定义很多，比如，“统计学是收集、分析、表述和解释数据的科学”“统计学是一组方法，用来设计试验、获得数据，然后在这些数据的基础上组织、概括、演示、分析、解释和得出结论”. 综合来说，**统计学（statistics）就是收集、处理、分析、解释数据并从数据中得出结论的科学**. 这一定义揭示了统计学是一套处理数据的方法和技术.

在不同场合，统计一词具有不同的含义，它可以指统计数据的搜集活动，即统计工作；也可以指统计活动的结果，即统计数据资料；还可以指分析统计数据的方法和技术，即统计学.

（1）**统计工作**是指搜集、整理、分析和研究统计数据资料的工作过程. 统计工作在人类历史上出现得比较早，随着历史的发展，统计工作逐渐发展和完善起来，使统计成为国家、部门、事业和企业、公司和个人及科研单位认识与改造客观世界和主观世界的一种有力工具. 我们的各级政府机构基本上都有统计部门，比如统计局，其主要职能就是从事统计数据的收集. 大多数企业也都有专门从事统计工作的人员，负责企业生产和销售数据的记录、积累以及向上级部门报送数据的任务. 统计工作，可以简称为统计. 例如，某统计师在回答自己的工种时，会说我是干统计的，这里所说的统计指的是统计工作.

（2）**统计数据**是统计工作中进行搜集、整理、分析和研究的主体及最终成果. 我们经常会看到专门出版统计数据的出版物，如《统计年鉴》，在网络、报纸、杂志上也会见到大量统计数据. 当你看到或听到“据统计……”这样的说法时，这里的统计一词是指统计数据. 由于统计数据（statistics）在英文中以复数形式出现，表明统计数据不是指单个数字，而是指同类的较多数据. 因为单个数据如果不和其他数据比较，是不能说明问题的. 例如，某学生在统计学考试中得了 86 分，如果仅凭这一数字，我们很难对这位学生的知识和能力做出判断和评价，因为 86 分在班级中可能是最高分，也可能是中间分，还可能是低分.

（3）**统计学**是对研究对象的数据资料进行搜集、整理、分析和研究，以显示其总体的特征和规律性的学科，亦可简称为统计. 例如，我们所学的课程 ——统计，实际指的是“统计学”课程.

正确理解“统计”概念是十分必要的. 一提到“统计”，就想到统计工作的思维习惯是片面的、狭隘的，要针对具体情况进行具体分析，三者相辅相成，不可分割.

（1）统计数据和统计学的基础是统计工作，统计工作的成果是统计数据，统计学既是统计工作经验的理论概括，又是指导统计工作的原理、原则和方法. 原始的统计工作，即人们收集数据的原始形态，已经有几千年的历史，而它作为一门科学，则是从 17 世纪开始. 在英语中，统计学家和统计员是同一个单词，但统计学并不是直接产生于统计工作的经验总结. 每

一门科学都有其建立、发展的客观条件，统计学是统计工作经验、社会经济理论、计量经济方法融合、提炼、发展而来的一门边缘性学科.

（2）统计数据的收集是取得统计数据的过程，是进行统计分析的基础. 离开了统计数据，统计方法就失去了用武之地，因此如何取得所需的统计数据是统计学研究的基本内容之一. 整理数据是对统计数据的加工处理过程，目的是使统计数据系统化、条理化，符合统计分析的需要，它是介于数据收集与数据分析之间的一个必要环节.

（3）统计数据的分析是统计学的核心内容，它是通过统计描述和统计推断的方法探索数据内在规律的过程. 如果不用统计方法分析，统计数据仅仅是一堆数据，甚至杂乱无章，不能得出任何有益的结论.

因此，可将统计研究的过程描述如下，如图 1.1.1 所示.

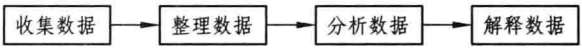


**图 1.1.1 统计研究的过程**

### 1.1.2 统计学的研究对象与分类

#### 1）统计学的研究对象

一般来说，统计学的研究对象是自然、社会客观现象总体的数量关系. 不论是自然领域，还是社会经济领域，客观现象总体的数量方面都是统计学所要分析和研究的. 从某种意义上说，统计学是寄生的，要依靠研究其他领域内的工作而生存. 这并不是对统计学的轻视，因为对很多寄主来说，如果没有寄生虫就会死，而对有的动物来说，如果没有寄生虫，它就不能消化食物. 因此，可以说，在人类奋斗的很多领域，如果没有统计学，虽然不会死亡，但它一定会变弱，这也从一个侧面要求统计工作者要知识渊博，最好是某一领域的行家.

统计学研究对象的特点有：

（1）**数量性**. 统计学的研究对象是自然、社会经济领域中现象的数量方面，故统计学是定量分析学科. 数字是统计的语言，数据资料是统计的原料. 事物的数量是我们认识客观现实的重要方面，通过分析研究统计数据资料，方可研究和掌握统计规律性，进而达到统计分析的研究目的.

（2）**总体性**. 统计的数量研究是对总体普遍存在着的事实进行大量观察和综合分析，得出反映现象总体的数量特征和资料规律性. 一般统计分析的目的并不局限于某个个体或者小团体，而是反映更大范围内的群体在某个方面的特征和属性，比如通过抽查部分大学毕业生的就业来推断大学生的就业状况. 在一般情况下，自然、社会经济现象的数据资料和数量关系等是在一系列复杂影响因素下形成的，有些因素起主要作用，有些因素起次要作用. 由于各种原因，每个个体都具有一定的随机性质，而对于有足够多个体的总体来说又具有相对稳定的共同趋势，显示出一定的统计规律性. 例如，人的身高有高低之分，各不相同，但经分析可发现：高个子父母的子女一般比矮个子父母的子女要高.

（3）**具体性**. 统计研究的是具体现象的数量方面，而不是纯数量的研究，它具有明确的现实涵义. 这一特点也是统计学与数学的分水岭. 数学是研究事物的抽象空间和抽象数量的科学，而统计学研究的数量是客观存在的、具体实在的数量表现. 统计研究的这一特点也决定了它的实用性.

（4）**变异性**. 统计研究对象的变异性是指构成统计研究对象的总体各单位，除了在某一方面必须是同质的以外，在其他方面又要有差异，而且这些差异并不是由某种特定的原因事先给定的，否则就没有必要进行统计分析研究了. 例如，学生作为统计数据资料对象，每个学生在性别、年龄、身高等方面会有不同的表现，这样统计分析研究才能对其表现出来的差异探索统计规律性.

思考：什么是社会现象？什么是自然现象？

提示：社会现象与自然现象是一个相对的概念. 太阳东升西落，四季轮回等都称为自然现象，而社会现象强调人的参与，是指与人类活动的产生、发展、变化密切联系的现象，例如农村儿童留守现象，欺诈现象，腐败现象等.

### 2）统计学的分类

统计方法已被应用到自然科学和社会科学的众多领域，统计学也发展成为由若干分支学科组成的学科体系，依照不同的标准，它的分类也不同.

（1）从统计方法的构成来分可分为描述统计学与推断统计学.

① **描述统计学**是通过图表或数学方法，对数据资料进行整理、分析，并对数据的分布状态、数字特征和随机变量之间关系进行估计和描述的方法. 描述统计是一套处理和分析数据的基本方法和技术，可分为集中趋势分析、离中趋势分析和相关分析三大部分.

② **推断统计学**是研究如何根据样本数据去推断总体数量特征的方法，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断. 例如，我们想研究教育背景是否会影响人的智力测验成绩，可以找 100 名 24 岁大学毕业生和 100 名 24 岁初中毕业生，采集他们的一些智力测验成绩. 用推断统计方法进行数据处理，最后会得出类似这样的结论："研究发现，大学毕业生组的成绩显著高于初中毕业生组的成绩，两者在 0.05 水平上具有显著性差异". 这说明大学毕业生组的智力测验成绩优于中学毕业生组.

值得注意的是，智力测试成绩是智商的主要标志，在相同条件下，成绩越高，智商也越高，如果条件不同，则没有可比性. 因此，虽然大学毕业生组的智力测试成绩比初中毕业生组的高，但我们不能说大学毕业生的智商比初中毕业生的智商高，因为他们的教育背景不同.

（2）从统计方法的研究和应用来分可分为理论统计学与应用统计学.

① **理论统计学**是指统计学的数学原理，主要研究统计学的一般理论和统计方法的数学理论. 现代统计学几乎用到了所有方面的数学知识，由于概率论是统计推断的数学和理论基础，因而广义地讲，统计学也应该包括概率论在内. 理论统计学是统计方法的理论基础，没有理论统计学的发展，统计学就不可能发展成为像今天这样一个完善的科学知识体系. 因此，作为从事统计理论和方法研究的人员必须要有坚实的数学基础，否则将寸步难行，这也是普通院校的学生不喜欢理论统计学的原因之一.

② **应用统计学**是研究如何应用统计方法去解决实际问题的. 在统计研究领域，相对来说，从事理论统计学研究的人占很少一部分，而大部分则从事应用统计学研究.

没有概率论，我们就无法真正理解统计；没有理论统计学，我们也无法真正理解和应用统计学. 目前，一些教育工作者一味地强调应用，不按教学规律上课，东讲一点，西讲一点，简直就是为了应用而应用. 在当时，学生感觉挺实用的，可过了几天，就忘记了，况且稍微复杂一点的应用也学不会，原因就在于知识不系统，理论薄弱. 为此，本书非常重视统计学

的理论学习，力争逻辑清晰，顺序得当，力争将理论统计学与应用统计学熔于一炉. 建议读者先修概率论，如果没学过概率论，凭借高中的概率与统计基础也可阅读本书，在前面章节中，实在遇到不甚理解的概念，可放下，我们在后面章节中会给出严格定义.

### 1.1.3 统计规律

统计学是探索数据内在规律的一套方法. 那么，什么是统计数据的内在规律呢？为什么统计方法能通过对数据的分析找到其内在的数量规律呢？下面通过几个例子进行说明.

**1）人口性别**

众所周知，就单个家庭而言，新出生婴儿的性别可能是男性，也可能是女性. 如果不限制生育，多个子女的家庭可能全部是男孩，也可能全部是女孩. 表面上看，新生婴儿的性别好像没有任何规律，但如果对大量家庭的新生婴儿进行统计分析，就会发现：新生婴儿的男孩略多于女孩，男女比例大概为 107 ∶ 100. 为什么男婴的出生率会高于女婴呢？拉普拉斯从概率论的观点解释说：这是因为含 X 染色体的精子与含 Y 染色体的精子进入卵子的机会不完全相同. 其实，女性中 XX 染色体比男性中 XY 染色体的可靠性高，这是因为 XX 可以看作并联系统，而 XY 可以看作串联系统. 另外，由于雄性激素的作用，男人更易具有危险动作，如打架斗殴、酗酒. 这样，在自然状态下，男人的死亡率会略高于女人，即使男婴出生率高一点，但到了结婚年龄，两者比例很接近. 进入中老年后，男性的死亡率仍然高于女性，导致男性的平均寿命低于女性，老年男性反而少于女性.

由于生育人口在性别上保持了大致平衡，才保证了人类社会的发展和延续，所以对人口性别的研究是统计学的起源之一，也是运用统计方法探究的数量规律之一.

**2）掷硬币游戏**

在投掷一枚硬币时，既可能出现正面，也可能出现反面，因此预先做出确定的判断是不可能的. 但是假如硬币均匀，直观上看，出现正面与反面的机会应该相等，即在大量的试验中出现正面的频率应接近 50%. 历史上有不少人做过抛硬币试验，其结果见表 1.1.1，从表中的数据可看出：出现正面的频率逐渐稳定在 0.5.

**表 1.1.1 抛掷硬币试验记录**

| 实验者 | 抛硬币次数 | 出现正面的次数 | 频率 |
|---|---|---|---|
| 德莫根（De Morgan） | 2 048 | 1 061 | 0.518 1 |
| 蒲丰（Buffon） | 4 040 | 2 048 | 0.506 9 |
| 费勒（Feller） | 10 000 | 4 979 | 0.497 9 |
| 皮尔逊（Person） | 12 000 | 6 019 | 0.501 6 |
| 皮尔逊 | 24 000 | 12 012 | 0.500 5 |

**3）英语字母的频率**

在生活实践中，人们逐渐认识到，英语中某些字母出现的频率要高于其他字母. 有人对各类典型的英语书刊中字母出现的频率进行统计，发现各个字母的使用频率相当稳定. 这项

研究对计算机键盘的设计（在操作方便的地方安排使用频率较高的字母键）、信息的编码（用较短的码编排使用频率最高的字母键）等都是十分有用的.

**4）最佳施肥量**

在进行农作物试验时，如果其他试验条件相同，我们会发现某种粮食作物的产量会随着某种施肥量的增加而增加. 在最初增加施肥量时，粮食产量增加得比较快，以后增加同量的施肥量，粮食产量的增加逐渐减少. 当施肥量增加到一定量时，粮食产量最高，这时如果继续增加，粮食产量反而会减少. 粮食产量与施肥量的这种数量关系（边际效用递减）就是我们探索的数量规律. 如果从大量的试验数据中，用统计方法找到粮食产量与施肥量之间的数量关系，就可得到最佳施肥量，进而达到最大效益.

上述例子说明：就一次观察或试验而言，其结果往往是随机的，但在大量试验中却呈现出某种规律性，这种规律性称为**统计规律性**. 利用统计方法可以探索出其内在的数量规律，因为客观事物本身是必然性与偶然性的对立统一，必然性反映了事物的本质特征和规律，偶然性反映了事物表现形式上的差异. 如果客观事物仅有必然性的一面，则它的表现形式就会很简单. 也正是偶然性的存在，才使得事物的表现形式和必然的规律性之间产生偏差，从而形成了表面的千差万别，使得事物的必然性被掩盖在表面的差异中. 这正如恩格斯所指出的："在表面上是偶然性在起作用的地方，这种偶然性始终是受内部隐藏着的规律支配的，而问题只是在于发现这些规律". 概率论的任务就是要透过随机现象的随机性揭示其统计规律性；统计学的任务则是通过分析带随机性的统计数据来推断所研究的事物或现象固有的规律性. 两者的研究目的都是随机现象的统计规律，但其研究方法存在一定差异，概率论主要利用演绎方法，统计主要利用归纳方法.

### 1.1.4　统计学的应用领域

目前，统计方法已被用到自然科学与社会科学的众多领域，统计学已发展为由若干学科分支组成的学科体系. 可以这样说，统计学几乎被用到所有研究领域，比如政府部门、学术研究领域、日常生活中、企业管理等，进而也形成了众多的具有统计学应用性质的学科，如社会统计学、工业统计学、农业统计学、物理统计学、生物统计学、医药统计学、人口统计学、空间统计学，等等. 现在，就连纯文科领域的法律、历史、语言、新闻等也越来越重视对统计数据的分析，国外的人文和社会学科普遍开设了统计学课程. 可以这样说，统计方法与数学、哲学一样成为了所有学科的基础.

**例 1.1.1**　用统计识别作者.

1787—1788 年，Alexander Hamilton，John Jay 和 James Madison 三位作者为了说服纽约人认可宪法，匿名发表了著名的 85 篇论文. 这些论文中的大多数作者已经得到了识别，但是，其中的 12 篇论文的作者身份引起了争议. 通过对不同单词的频数进行统计分析，得出的结论是：James Madison 最有可能是这 12 篇论文的作者. 现在，对于这些存在争议的论文，认为 James Madison 是原创作者的说法占主导地位，而且几乎可以肯定这种说法是正确的.

"红楼梦"后 40 回是否为曹雪芹所写？1985—1986 年，复旦大学李贤平教授带领他的学生做了这项工作. 他们的创造性想法是将 120 回看成 120 个样本，然后确定与情节无关的虚词作为变量. 为什么要抛开情节？这是因为在一般情况下，同一情节大家描述得都差不多，

但由于个人写作特点和习惯的不同，所用的虚词是不一样的. 然后，数出每一回里变量出现的次数，并作为数据，用多元分析中的聚类分析法进行分类. 分析结果果然表明，120 回分属两类，即前 80 回为一类，后 40 回为一类，这有力地证实了全书不是出自同一人的手笔. 那么后 40 回是否为高鹗所写？同样，论证结果推翻了“后 40 回是高鹗一个人所写”. 这个论证在红学界轰动很大，也支持了红学界观点，使红学界大为赞叹.

下面给出统计在工商管理中的部分应用：

（1）**产品质量管理**：质量是企业的生命，是企业持续发展的基础，统计质量管理已成为质量管理的重要手段. 在一些知名的跨国公司中，$6\sigma$ 准则已经成为一个重要的管理概念，统计质量控制图已广泛应用于生产过程的检测中.

（2）**市场研究**：企业要在激烈的市场竞争中取得优势，首先必须了解市场，要了解市场，就需要做广泛的市场调查，取得所需的信息，并对这些信息进行科学的分析，以便作为生产和营销的依据，而这些都需要统计的支持.

（3）**经济预测**：企业家要对未来的市场状况进行预测，经济学家要对宏观经济或某一行业进行预测，他们在进行预测的时候，最常用的方法就是利用各种统计信息和统计方法. 比如，企业家要对市场的潜力进行预测，以便调整生产计划，使利润最大化，这就需要通过市场调查取得数据，并进行统计分析.

（4）**人力资源管理**：利用统计方法对员工的年龄、性别、受教育程度、工资等进行分析，并作为制定工资计划、奖惩制度的依据.

当然，统计不仅在工商管理中有用，而且已经渗透到自然科学和社会科学的各个领域，为多个学科提供通用的数据分析方法. 从某种意义上说，统计仅仅是数据分析方法，必须与其他学科结合才能发挥自己的作用，于是统计工作者的知识面一定要宽广，最好擅长某一方面，比如生物、医学、农业、教育、经济等.

下面列出统计的一些应用领域，让读者简单浏览并形成一个概念：统计学非常有用.

精算　金融　农业　动物学　考古学　医学　生态学　教育学　计量经济学　管理学
人类学　博彩　地质学　心理学　质量控制　工业　工程　水文学　军事科学
物理学　市场营销学　地理学　语言学　分类学　宗教研究

统计应用上的两个极端情况是：不用或几乎不用统计；简单问题复杂化. 在统计应用中，这两种情况都是不可取的. 简单的方法不一定没用，复杂的方法也不一定有用. 正如有的学者所说的，最简单的模型往往是最有用的，统计应该恰当地应用到它能起作用的地方. 同时，我们不能把统计神秘化，更不能歪曲统计，把统计作为掩盖事实的陷阱. 统计是一种从数量上认识客观世界的有力工具，但是若运用不当，即使是科学的技术也有可能得出错误的结论，甚至会成为谬误的护身符. 实际上，在我们周围，误用统计数字或滥用统计方法的现象是不乏其例的，这不能不影响到统计科学的严肃性和统计分析的准确性.

下面举例介绍统计欺骗中的常用手法，以供读者参考.

（1）**有偏样本**：有一个装着红、白两色小球的箱子，如果你想要准确地知道这个箱子中两种小球的数量，你唯一能做的就是一颗一颗地数小球. 然而用一种更简单的方法也可以估计红球的数量：抓一把小球，假定手中红球所占的比例与箱子中红球所占的比例相同，只要数一数手中的小球即可. 如果你的样本足够大且选择方法正确，那么在大多数情况下，它能

够很好地代表整体. 但是，如果以上两个条件均不满足，利用这样的样本进行推断比猜想好不到哪儿去，除了能够营造科学、精确的假象之外，其他则根本不值一提. 不幸的是，我们所看到的，或者我们自以为了解的许多事物，往往都是根据类似样本所得出的结论，这种样本由于选择方式的不合理或者容量过小，抑或两种情况同时存在，均导致样本有偏. 现在通过一个极端的例子可以马上看到有偏样本是如何形成的. 假设你向同学发放问卷，问卷中包含这样一个问题："你乐意回答调查问卷吗？"整理所有的答案，你很有可能得到下面的结论：压倒多数的人选择了"乐意". 为了能有说服力，你还可以详细列出这个比例，直至最后一位小数. 事实上，大多数持否定意见的人已经随手将你的问卷丢进垃圾篓中，即从样本中自动除名了. 哪怕最初的样本中，100 个里面有 88 个会当这种"投手"，在宣布你的结果时，你仍然会遵从惯例，忽略他们. 第二次世界大战期间，英国空军希望增加飞机的装甲厚度，但如果将装甲全部加厚，则会降低灵活性，所以最终决定只增加受攻击最多部位的装甲. 后来，工作人员经过对中弹飞机的统计，发现大部分飞机的机翼弹孔较多，所以决定增加机翼的装甲厚度. 后来一个专家说："可是机头中弹的那些飞机就没有飞回来". 这个故事里本应是对全部飞机进行分析，但统计样本没有包含已经损毁的飞机，所以得出的结论只是根据部分数据，或者说是根据具有同样特征（受伤）的某一类数据推出的，并不能代表全部类型的数据，所以得出的结果很可能是错误的.

（2）**数字欺骗**：主要是通过刻意强调"有利指标数据"，而回避"不利指标数据"，以隐瞒事实. 有一个"数字欺骗"的经典段子：西门庆正与潘金莲鬼混，恰逢武松归来. 武松怒目圆睁要抓墙上的戒刀！武大忙拉住弟弟的袖子说道，其实我已经赢了！今年 4 月，西门官人才来咱家 20 次，同比下降 17%，环比虽然上升 20%，但是增幅明显放缓； 昨天和你嫂子在楼上的时间只有 68 分钟，同比下降 24%，明显是害怕我了.

（3）**误导性图表**："图表欺骗"主要是通过对图表中"不利指标数据"进行隐藏、过滤、更换颜色、更改图表类型实现、更改比例等手段实现. 比如，改变图表宽度与高度，影响性能趋势理解；改变图表默认显示方式，干扰读者理解数据图表.

（4）**故意曲解**：事实上，统计数据本身并无罪，可人们往往喜欢滥用统计工具来支撑自己的立场，而不是反映真实情况. 统计最重要的功能是数据分析，不同的人对数据分析的理解大不一样，曲解数据是常有的现象. 在某些人的心目中，数据分析就是寻找证据支持自己的结论，这恰恰曲解了数据分析的本质. 数据分析的真正本质是从数据中寻找规律，从数据中寻找启发，而不是寻找支持. 真正的数据分析是没有事先结论的，而是通过数据分析得出某种结论.

切记，统计不是万能的，它不能解决你面临的所有问题. 统计是分析数据的一门通用工具，虽然可以帮助你进行数据分析，并从分析中得出某种结论，但是如果要对统计结论进一步分析，则需要你的专业知识. 比如，吸烟可以增加患肺癌的概率，但要解释吸烟为什么能引起肺癌，这不是统计学家的任务，需要更多的医学知识才能做到.

## 1.2　数据的计量与类型

统计研究的是客观事物的数量方面，它离不开统计数据. 统计数据是对客观现象进

行计量的结果，比如，对股票价格变动水平进行计量可以得到股票价值指数；对学生考试结果进行计量可得考试分数；对经济活动总量的计量可以得到国内生产总值（GDP）；等等. 因此，在收集数据之前，我们总是要先对现象进行计量或测度，这就涉及计量尺度的问题.

### 1.2.1 数据的计量尺度

**计量尺度**（levels of measurement）是指对计量对象量化时采用的具体标准，如千克、米、美元、人民币等. 由于客观事物有的比较简单，有的比较复杂；有的特征和属性是可见的（如人的外貌体征），有的则是不可见的（如人的偏好、道德、信仰）；有的表现为数量差异，有的表现为品质差异，所以有些事物只能对它的属性进行分类，比如人口性别和文化程度等；有的可以采用比较精确的数字加以计量，比如物体的长度、产品的质量等. 因此，对于不同事物，我们能够计量或测度的程度是不同的，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. 从对事物计量的精确程度来看，采用数字计量比采用分类计量精度更高. 按照计量的精度程度，从低级到高级，从粗略到精确可分为四个层次：定类尺度、定序尺度、定距尺度和定比尺度. 采用不同的计量尺度可以得到不同的统计数据，进而采用的统计分析方法也是不同的.

#### 1）定类尺度

**定类尺度**（nominal scale），也称为**类别尺度**，它将数字作为现象总体中不同类别或不同组别的代码，这是最低层次的计量尺度，也是其他计量尺度的基础. 这种计量尺度只能按照事物的某种属性对其进行平行的分类或分组，不同的数字仅表示不同类（组）别的品质差别，而不表示它们之间量的顺序或量的大小. 这种尺度的主要数学特征是“=”或“≠”. 例如，人口按照性别可分为男（1）、女（0）两类，不能因为 $1>0$，就说男人大于女人，而应该说男人不等于女人. 国民经济按其经济性质可以分为国有经济、集体经济、私营经济、个体经济等类，并用代码（01）表示国有经济，（02）表示集体经济，（03）表示私营经济，（04）表示个体经济. 并且用（011）代表国有经济中的国有企业，（012）代表国有联营企业；用（021）表示集体经济中的集体企业，（022）表示集体联营企业；用（031）表示私营经济中的私营独资企业，（032）表示私人合伙企业，（033）表示私营有限责任公司；用（041）表示个体经济中的个体工商户，（042）表示个人合伙等. 其中，前两位代码表示经济大类，而第三位代码则表示各类中的构成，代码的位数与类别数目有关，如果类别数目多，则代码位数也多. 本例中，国民经济最多分为 99 大类，每一大类中最多分为 9 小类. 不同代码反映同一水平的各类（组）别，并不反映其大小顺序. 各类中虽然可以计算它的单位数，但不能反映第一类的一个单位可以相当于第二类的几个单位等. 对于定类尺度的计量结果，我们可以通过计算每一类别中各个元素或个体出现的频数或频率来进行分析.

在使用定类尺度对事物进行分类时，必须符合**穷尽**和**互斥**的要求，即对事物进行无交完备分解. **类别穷尽**是指在所有全部的分类中，必须保证每一个元素或个体都归属于某一类别，不能遗漏；**类别互斥**是指每一个元素或个体只能归属于一个类别，不能在其他类别中重复出现. 比如，按照自然两分法，一个人要么是男性，要么非男性，总有所归属，而且只能属于其中一个类别.