普通高等教育“十一五”国家级规划教材

教育部经济管理类核心课程教材

ECONOMICS AND MANAGEMENT

# 统计学

（第三版）

贾俊平　编著

Statistics

(Third Edition)

中国人民大学出版社

普通高等教育“十一五”国家级规划教材

教育部经济管理类核心课程教材

ECONOMICS AND MANAGEMENT

# 统计学

（第三版）

贾俊平　编著

Statistics

(Third Edition)

中国人民大学出版社

·北京·

**图书在版编目（CIP）数据**

统计学/贾俊平编著. 3版.
北京：中国人民大学出版社，2008
普通高等教育“十一五”国家级规划教材. 教育部经济管理类核心课程教材
ISBN 978-7-300-09913-2

Ⅰ. 统…
Ⅱ. 贾…
Ⅲ. 统计学-高等学校-教材
Ⅳ. C8

中国版本图书馆CIP数据核字（2008）第171583号

普通高等教育“十一五”国家级规划教材
教育部经济管理类核心课程教材
**统计学（第三版）**
贾俊平　编著

| | | | |
|---|---|---|---|
| **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** | 北京东君印刷有限公司 | | |
| **规　　格** | 185 mm×235 mm　16开本 | **版　　次** | 2003年6月第1版 |
| | | | 2008年11月第3版 |
| **印　　张** | 21.5插页2 | **印　　次** | 2008年11月第1次印刷 |
| **字　　数** | 412 000 | **定　　价** | 32.00元 |

# 《教育部经济管理类核心课程教材》

# 出 版 说 明

按照购买力平价标准衡量，中国已被世界银行列为世界第二大经济体，仅次于美国。但是，我们不能因此沾沾自喜。成为经济大国并不意味着就是经济强国，中国的强国之路依然漫长而曲折。我们应该清醒地认识到，面对新的发展形势，我们自身还存在着许多短板，如果不能及时将这些短板补齐，我们将会在前进的道路上失去平衡而摔跤。最重要的短板之一，是我们在经济管理高等教育与实践方面的落后和不足。中国现代经济管理实践比西方国家晚几十年甚至上百年，很多理论知识和实践经验最初是从西方“拿来”的，这导致中国的经济管理类人才在知识储备上总是落后于人，缺乏领先的理念来引导实践。

基于以上认识，中国人民大学出版社近年来不断深化教材的层次和结构，无论是引进版还是本版，都从多个维度进行开发和建设，以适应新的发展要求。作为国内最早引进国外优秀经济管理类图书的出版社之一，我们最初引进的一批经典欧美经济管理类图书造就了一大批成功的管理者。借鉴引进版的成功经验，在本土教材开发方面，除了及时吸纳国内外经济管理领域的先进思想和理念，还提供尽可能多的案例，特别是本土案例。这一点在《教育部经济管理类核心课程教材》系列中体现得十分充分。

本套教材的开发思路得到了对外经济贸易大学国际商学院院长张新民教授、中国人民大学商学院王化成教授以及上海交通大学安泰经济与管理学院执行院长徐飞教授的极大认同和支持。感谢这些老师投入极大的热情，与我们共同设计整套教材的方案，选定每本教材的主编，制定教材开发原则和体例。每位参编老师都是各自领域的佼佼者，并且无论其身居何职，都依然站在教学第一线。我们尽力做到教材从内容到形式都具有独特的风格；同时，我们还为每本书配备了案例集或学习指导书，并提供一些教学辅助资料供教师免费下载，为使用教材的老师和学生们提供尽可能周到的服务。

作为新中国成立后最早建立的一家大学出版社，中国人民大学出版社一直秉承“出教材学术精品，育人文社科英才”的宗旨。如今同类经济管理类教材充斥市场，我们更觉得有责任紧跟时代脉搏，不断推出精品，提升教材的质量和层次，一方面，为选择教材的广大师生节约选书的时间成本，另一方面，也希望为提升中国的经济管理教育和实践水平做出贡献。我们期待着广大使用者的建议和鞭策，促使我们不断对本套教材进行改进和完善，使之长远传承，经久不衰。

中国人民大学出版社

# 前　言

统计作为数据分析的一种通用语言，已被越来越多的领域所应用。对很多人而言，掌握统计技术可以在竞争日趋激烈的就业市场中占据优势。统计学作为研究数据的一门科学，为使用者提供了一套获取数据、分析数据并从数据中得出结论的原则和方法。

多数人都把统计学作为一门难学的课程来看待，但统计学其实并不像人们想象的那么难，关键是看你怎么学。如果在学习过程中把注意力放在公式上，放在数据的计算过程上，而忽视对统计思想的理解，不仅难以学会，也难以将统计用到实处。这样的学习方法无形中把统计给复杂化了。如果抛开复杂公式的表象，静心想一想，特别是把简单而繁杂的计算过程交给计算机来完成的时候，就会发现统计其实很简单。

统计的精髓是使复杂问题简单化，而不是把简单问题复杂化；统计的真谛在于它所体现的思想，在于它所提供的思维方式；学好统计的关键是掌握如何运用统计思维来思考问题，而不是简单地记住那些死的统计知识。本书试图体现作者的这些想法。与前两版相比，第三版有如下变化和新颖之处。

第一，首次将部分多元统计方法纳入《统计学》教材。第三版增加了三章新内容，其中两章为多元统计方法，包括主成分分析和因子分析、聚类分析，另一章为非参数检验。这样做的原因有二：一是计算机应用的普及使这些方法的实际应用成为可能，其实也是必然；二是越来越多的读者已经提出了这种要求，一些院校已经将统计教学完全与计算机结合起来，使这些方法的教学和学习都变得容易起来。

第二，实现教材内容与计算机的完全结合。除部分为展示方法的计算过程外，多数统计计算都实现了计算机化。在讲清楚统计方法原理和思想的基础上，基本上都给出了由统计软件实现计算的详细操作步骤，并对输出结果做了

详细解读。考虑到读者对统计软件的接近程度和熟悉程度，本书结合使用了 Excel 和 SPSS 两个软件。凡是 Excel 能够解决的，仍然以 Excel 为主，Excel 不能解决的，使用了多数人熟悉的 SPSS。对每种软件都给出了具体的操作步骤，读者按此操作就会得到所需的统计分析结果。

第三，书稿的篇幅大大压缩了。在增加了统计方法和内容的基础上，书稿的篇幅没有增加。写作时尽可能删除一些不必要的表述，使内容更加简明、易懂。

第四，本书配有内容丰富的教学和学习资源库，内容包括教学和学习用 PPT、教材案例和练习题的数据文件、教学案例、模拟试题、教师用试题题库、学生用各章自测题等，放在中国人民大学出版社工商管理出版分社的网站上，网址为：www. rdjg. com. cn/zyk/tj。此外，还专门配有与本书配套的学习指导书，书中给出了每一章的知识结构、学习要点、选择题及其答案，以及教材后所附练习题的详细解答。

本书可作为高等院校经济管理类专业本科生统计学课程的教材使用，也可作为研究生和 MBA 的教材或参考书。在使用中，对有些章节可根据教学需要和教学时数酌情选讲。希望本书能对您有所帮助，也希望您提出更多的修改建议，感谢您对本书的支持。同时，也要感谢中国人民大学出版社的编辑，他们对本书写作的支持以及对书稿的认真校对和颇有效率的工作，使得本书能尽快与读者见面。

贾俊平

# 目 录

# 第1章 统计和统计数据

统计思维总有一天会像读与写一样成为一个有效率公民的必备能力。

——H. G. Wells

## 你相信这样的一些统计结论吗?

每天你都会看到各种统计数字，但你或许没有仔细想过它们意味着什么。看看下面的一些统计研究结果，你会有怎样的看法呢?

- 吸烟对健康是有害的，吸烟的男性减少寿命2 250天。
- 不结婚的男性会减少寿命3 500天，不结婚的女性会减少寿命1 600天。
- 身体超重30%会使寿命减少1 300天。
- 每天摄取500毫升维生素C，生命可延长6年。
- 身材高的父亲，其子女的身高也较高。
- 第二个出生的子女没有第一个聪明，第三个出生的子女没有第二个聪明，依此类推。
- 学生们在听了莫扎特钢琴曲10分钟后的推理测试会比他们听10分钟娱乐磁带或其他曲目做得更好。
- 上课坐在前面的学生平均考试分数比坐在后面的学生高。

看懂这些结论并不困难，但这些结论是怎样得出来的?你相信这些结论吗?学点儿统计学知识你就会正确理解它们。

在日常生活中，经常会接触到统计数据或一些统计研究结果。比如，在电视、报纸、网络等媒体中就会经常看见一些报道使用统计数据、图表等。作为一门科学的统计学研究什么呢?怎样获得所需要的统计数据呢?这就是本章将要介绍的问题。

# 1.1 统计及其应用领域

每个人都离不开统计，了解一些统计学知识对每个人都是必要的。比如，在外出旅游时，你需要关心一段时间内的详细天气预报；在投资股票时，你需要了解股票市场价格的信息，了解某只特定股票的有关财务信息；在观看足球比赛时，除了关心进球的多少外，你还要知道各支球队的技术统计；等等。要正确阅读并理解统计数据，就需要具备一些统计学知识。

## 1.1.1 统计学研究什么

### 1. 什么是统计学

在你的工作或管理中，总会面对各种各样的数据。你需要分析这些数据，从中得出某些结论以帮助你作出决策。统计就是用来处理数据的，它是关于数据的一门学问。统计学提供的是一套有关数据收集、数据处理、数据分析的方法。概括地讲，**统计学**（statistics）是收集、处理、分析、解释数据并从数据中得出结论的科学。统计分析数据所用的方法大体上可分为**描述统计**（descriptive statistics）和**推断统计**（inferential statistics）两大类。

描述统计是研究数据收集、处理和描述的统计学方法。其内容包括如何取得研究所需要的数据，如何用图表形式对数据进行处理和展示，如何通过对数据的综合、概括与分析，得出所关心的数据的特征。

推断统计则是研究如何利用样本数据来推断总体特征的统计学方法，内容包括参数估计和假设检验两大类。参数估计是利用样本信息推断所关心的总体特征，假设检验则是利用样本信息判断对总体的某个假设是否成立。比如，从一批灯泡中随机抽取少数几个灯泡作为样本，测出它们的使用寿命，然后根据样本灯泡的平均使用寿命估计这批灯泡的平均使用寿命，或者是检验这批灯泡的使用寿命是否等于某个假定值，这就是推断统计要解决的问题。

### 2. 统计学研究什么

问问你身边的人，GDP（国内生产总值）是什么？CIP（消费者价格指数）是什么？他们似乎都能说上几句。但要是仔细追问它们究竟代表了什么，就不是每个人都能够说清楚的了。统计也是一样，要问一个人统计是什么，似乎没有人不知道，但你要问统计究竟是什么，就不是一两句话能够说明白的，要搞清楚统计研究什么就更困难了。

物理学研究的是热、光、电等这类自然现象的运动规律。化学家测定物质的组

成及化学元素之间的交互作用。生物学家研究植物和动物的生活。数学家则在给出的假定之下推演各种命题。这些学科都有它们自己的研究对象，而且有解决这些问题的各自的方法，各学科因此而成为一门单独的学科。

统计学是一门独立的学科，这似乎没人怀疑。但统计究竟研究什么？可能就有不同的看法。有人认为，统计学是一门独特的学问，没有任何固定的对象。乍听起来似乎难以理解，但仔细想想也许有道理。统计学研究的是来自各领域的数据，靠解决其他领域的问题而存在和发展。按 L. J. Savage 的说法："统计学基本上是寄生的。靠研究其他领域内的工作而生存。这不是对统计学的轻视，这是因为对很多寄主来说，如果没有寄生虫就会死。对有的动物来说，如果没有寄生虫就不能消化它们的食物。因此，人类奋斗的很多领域，如果没有统计学，虽然不会死亡，但一定会变得很弱。"① 看上去统计似乎被边缘化了，实际上这也正说明了统计在各学科领域的独特地位和作用，也表明了统计作为一门独立的学科所具有的特点。

按统计学家 C. R. Rao 的说法："今天，统计学已发展成为一门媒介科学，它研究的对象是其他学科的逻辑和方法论——作出决策的逻辑和试验这些决策的逻辑。统计学的未来依赖于向其他学习领域内的研究者正确传授统计学的观点；依赖于如何能够在其他知识领域内将其主要问题模式化。"② 因此，在他看来，统计学是科学、工艺和艺术这三者的组合。

统计学是一门科学。它提供一套方法和技术，这些方法和技术并不是一成不变的，使用者在给定的情况下必须根据所掌握的专门知识选择使用这些方法，而且，如果需要还要进行必要的修正。统计方法是通用的数据分析方法，这些方法不是为某个特定的问题领域而构造的。

统计学是一种工艺。如同工业生产过程中的质量控制程序一样，统计方法是在为保证产品达到所希望的质量和保持其稳定性的管理系统中建立起来的。统计方法也能用于控制、减少和考察不确定性。

统计学是一门艺术。它提供一种归纳推理的方法，推理就是一种艺术。既然是归纳推理，就不能保证结论百分之百正确，就不能没有争议。怎样让别人看懂并理解统计结论，就要看统计表达这些结论的技巧和艺术性了。

这些观点听起来有点儿"哲学"，但它能帮助我们理解统计学是什么，统计学研究什么。统计研究的是来自各领域的数据，提供的是一套通用于所有学科领域的获取数据、分析数据并从数据中得出结论的原则和方法。

---

①② C. R. Rao：《统计与真理——怎样运用偶然性》，北京，科学出版社，2004。

## 1.1.2 统计的应用

### 1. 统计的应用领域

说出哪些领域应用统计，这很困难，因为几乎所有的领域都应用统计；要说出哪些领域不使用统计，同样也很困难，因为几乎找不到一个不用统计的领域。可以说，统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。只要有数据的地方就会用到统计方法。这里，我们不想列举统计的应用领域，只想通过几个简单的例子说明统计的应用。

**【例 1—1】** 用统计识别作者。1787—1788 年，三位作者 Alexander Hamilton，John Jay 和 James Madison 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。

**【例 1—2】** 用简单的描述统计量得到一个重要发现。R. A. Fisher 在 1952 年的一篇文章中举了一个例子，说明如何由基本的描述统计量的知识引出一个重要的发现。20 世纪早期，哥本哈根卡尔堡实验室的 J. Schmidt 发现不同地区所捕获的同种鱼类的脊椎骨和鳃线的数量有很大不同；甚至在同一海湾内不同地点所捕获的同种鱼类，也发现这样的倾向。然而，鳗鱼的脊椎骨的数量变化不大。Schmidt 从欧洲各地、亚速尔群岛以及尼罗河等几乎分离的海域里所捕获的鳗鱼的样本中，计算发现了几乎一样的均值和标准偏差值。由此，Schmidt 推断所有不同海域内的鳗鱼是由海洋中某公共场所繁殖的。后来名为“Dana”的科学考察船在一次远征中发现了这个场所。

**【例 1—3】** “挑战者号”航天飞机失事预测。1986 年 1 月 28 日清晨，载有 7 名宇航员的“挑战者号”进入发射状态。就在发射前，有冰片牢附在机壳上。几分钟后，正当电视新闻报道它已进入轨道时，航天飞机在毁灭性的爆炸声中化成碎片，机上的宇航员片骨未存。推动航天飞机进入太空的两个固体燃料发动机是由 Thiokol 公司制造的。失事前一天晚上，Thiokol 公司的经理们和美国宇航局（NASA）就如期发射还是推迟发射产生了争执。天气预报发射时的气温为 31°F。争执的结果是采纳了 Thiokol 公司经理们的建议：按计划发射航天飞机。因为他们觉得没有确凿证据表明低温会对固体燃料火箭推进器的性能产生影响。在此次失事前，该航天飞机 24 次发射成功。将航天飞机送入太空的两个固体燃料推进器由 6 只 O 型项圈密封。在几次飞行中，曾发生过 O 型项圈被腐蚀或气体泄漏事故。这

样的事故是极其危险的。前24次发射中有一次发动机遭到了永久性破坏。根据23次飞行中发生腐蚀或泄漏事故的次数（因变量 $y$）及火箭连接处的温度（自变量 $x$）数据，进行线性回归得到的回归方程为 $\hat{y}=3.698-0.047\,54x$。当温度为31℉时，O型项圈发生事故的预计次数为2.225次。结果显示连接处的温度与O型项圈事故之间有一定的相关性。如果当时那些经理们看到了回归的预测结果，也许推迟发射会成为其谨慎的选择。

前两个是统计得以应用并取得成效的例子，后一个是统计结果未被采纳而酿成惨剧的例子。不管怎样，它们都表明统计在许多领域都有广泛应用。

**2. 统计的误用与滥用**

大约一个世纪以前，政治家 Benjamin Disraeli 曾有一个著名的论断："有三类谎言：谎言、糟透的谎言和统计。"统计常常被人们有意或无意地滥用，比如，错误的统计定义、错误的图表展示、一个不合理的样本、数据的遗漏或逻辑错误等。这些误用有些是常识性的，有些是技术性的，有些则是故意的。作为从数据中寻找事实的统计，却被有些人变成了歪曲事实的工具。你也许常常看到这样的产品质检报告：某某产品的抽样合格率是80%。乍看上去没什么问题，但如果事实上只抽查了5件产品，有4件合格，这样的合格率能说明什么问题呢？在马路上随便采访几个人，他们的看法能代表大多数人的观点吗？调查了多少个人？是随机调查的吗？样本是怎样选取的？这看上去是在用事实说话，实际上成了统计陷阱。在有些人的心目中，数据分析就是寻找支持。他们的心目中可能有了某种"结论"性的东西，或者说他们希望看到一种符合他们需要的某种结论，而后去找些统计数据来支持他们的结论，这恰恰歪曲了数据分析的本质。数据分析的真正目的是从数据中找出规律，从数据中寻找启发，而不是寻找支持。真正的数据分析事先是没有结论的，通过对数据的分析才得出结论。

此外，统计也往往被作为两个极端使用：一个极端是不懂或不太懂统计的人认为统计没什么用，他们因为不懂统计而瞧不起统计，他们不用或几乎不用统计方法分析数据，即使作些统计分析，也往往是表面上的。走入这一极端的人，他们的决策依据就是自己的大脑：一些杂乱无章的信息组合出的某种直觉。如果他们的决策是正确的，更增加了他们的自信，更加感到不用统计也挺好；如果他们的决策出了毛病，便会找出一大堆理由：市场难测，环境突变，竞争激烈，需求疲软，价格下跌，管理不善，成本上升，出口下降……总之，决策失误的理由肯定与统计无关。另一个极端是把简单问题复杂化，特别是在管理领域，一些管理者把本来可以用简单方法解决的问题故意复杂化，他们不用简单的分析方法，而是用复杂的分析方法；他们为证明管理的科学性，建立一个别人看不懂的模型，编一大堆程序，输出了一大堆数字和符号；他们得出用统计语言陈述的结论，提出一些似是而非的建

议；等等。这样的分析往往是脱离了管理问题，对实际决策也未必有用。在统计应用中，这两个极端都是不可取的。管理决策中不用统计几乎不可想象；把简单问题复杂化对管理决策也未必有用。从统计的实际应用来看，简单的方法不一定没用，复杂的方法也不一定有用。正如有的学者所言，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。不能把统计神秘化，更不能歪曲统计，把统计作为掩盖实事的陷阱。

## 1.2 怎样获得统计数据

### 1.2.1 变量与数据

观察一个企业的销售额，你会发现这个月和上个月有所不同；观察股票市场上涨股票的家数，今天与昨天数量不一样；观察一个班学生的生活费支出，一个人和另一个人不一样；投掷一枚骰子观察其出现的点数，这次投掷的结果和下一次也不一样。这里的“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等就是**变量**（variable），它们的特点是从一次观察到下一次观察会出现不同结果。把观察到的结果记录下来就是**数据**（data）。

“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”这些变量可以用数字记录其观察结果，这样的变量称为**定量变量**（quantitative variable）或**数值变量**（metric variable）。定量变量的观察结果称为定量数据或**数值型数据**（metric data）。但你要观察人的性别、企业所属的行业、学生所在的学院等，这些变量的观察结果就不是数字，而是表现为不同的类别。比如“性别”表现为“男”或“女”，“企业所属的行业”表现为“制造业”、“零售业”、“旅游业”等，“学生所在的学院”则可能是“商学院”、“法学院”等，这些表现为不同类别的变量称为**分类变量**（categorical variable）。分类变量的观察结果就是**分类数据**（categorical data）。如果类别具有一定的顺序，这样的变量也称为**顺序变量**（rank variable）或有序分类变量，相应的观察结果就是**顺序数据**（rank data）或有序分类数据。比如考试成绩按等级分为优、良、中、及格、不及格，一个人对事物的态度分为赞成、中立、反对。这里的“考试成绩等级”、“态度”等就是顺序变量。分类变量和顺序变量也统称为**定性变量**（qualitative variable）。

### 1.2.2 数据的来源

从哪里取得所需的数据呢？对大多数人来说，可以使用已有的数据。比如，公

开出版或公开报道的数据，像统计部门公开出版的各种统计年鉴；分布在各种报纸、杂志、图书、广播、电视媒体中的各种数据；其他管理部门已有的数据；等等。也可以在网络上获取所需的数据，比如，各种金融产品的交易数据，官方统计网站的各种宏观经济数据等。

已有的数据不能满足需要时，可以亲自去调查。比如，你想了解全校学生的生活费支出状况，可以从中抽出一个样本获得样本数据。这里“全校所有学生”是你所关心的**总体**（population），它是包含所研究的全部个体（数据）的集合。从全校学生中抽取200人进行调查，这就是一个**样本**（sample），它是从总体中抽取的一部分元素的集合。构成样本的元素的数目称为**样本量**（sample size）。

怎样获得一个样本呢？要在全校学生中抽取200人组成一个样本，如果全校学生中每一个学生被抽中与否完全是随机的，而且每个学生被抽中的概率是已知的，这样的抽样方法称为概率抽样。概率抽样方法有简单随机抽样、分层抽样、系统抽样、整群抽样等。

**简单随机抽样**（simple random sampling）是从含有 $N$ 个元素的总体中，抽取 $n$ 个元素组成一个样本，使得总体中的每一个元素都有相同的机会（概率）被抽中。采用简单随机抽样时，如果抽取一个个体记录下数据后，再把这个个体放回到原来的总体中参加下一次抽选，叫做**重复抽样**（sampling with replacement）；如果抽中的个体不再放回，再从所剩下的个体中抽取第二个元素，直到抽取 $n$ 个个体为止，这样的抽样方法叫做**不重复抽样**（sampling without replacement）。由简单随机抽样得到的样本称为**简单随机样本**（simple random sample）。

**分层抽样**（stratified sampling）也称分类抽样，它是在抽样之前先将总体的元素划分为若干层（类），然后从各个层中抽取一定数量的元素组成一个样本。比如，要研究学生的生活费支出，可先将学生按地区进行分类，然后从各类中抽取一定数量的学生组成一个样本。分层抽样的优点是可以使样本分布在各个层内，从而使样本在总体中的分布比较均匀。

**系统抽样**（systematic sampling）也称等距抽样，它是先将总体各元素按某种顺序排列，并按某种规则确定一个随机起点，然后，每隔一定的间隔抽取一个元素，直至抽取 $n$ 个元素组成一个样本。比如，要从全校学生中抽取一个样本，可以找到全校学生的花名册，按花名册中的学生顺序，用随机数找到一个随机起点，然后依次抽取就得到一个样本。

**整群抽样**（cluster sampling）是先将总体划分成若干群，然后以群作为抽样单元从中抽取部分群组成一个样本，再对抽中的每个群中包含的所有元素进行观察。比如，可以把每一个学生宿舍看作一个群，在全校学生宿舍中抽取一定数量的宿舍，然后对抽中的宿舍中每一个学生进行调查。整群抽样的误差相对要大一些。

## 主要术语

- **统计学**（statistics）：收集、处理、分析、解释数据并从数据中得出结论的科学。
- **描述统计**（descriptive statistics）：研究数据收集、处理和描述的统计学方法。
- **推断统计**（inferential statistics）：研究如何利用样本数据来推断总体特征的统计学方法。
- **变量**（variable）：每次观察会得到不同结果的某种特征。
- **分类变量**（categorical variable）：观测结果表现为某种类别的变量。
- **顺序变量**（rank variable）：又称有序分类变量，观测结果表现为某种有序类别的变量。
- **数值型变量**（metric variable）：又称定量变量，观测结果表现为数字的变量。
- **分类数据**（categorical data）：只能归于某一类别的非数字型数据。
- **顺序数据**（rank data）：只能归于某一有序类别的非数字型数据。
- **数值型数据**（metric data）：按数字尺度测量的观察值。
- **总体**（population）：包含所研究的全部个体（数据）的集合。
- **样本**（sample）：从总体中抽取的一部分元素的集合。
- **样本量**（sample size）：构成样本的元素的数目。
- **简单随机抽样**（simple random sampling）：从含有 $N$ 个元素的总体中，抽取 $n$ 个元素组成一个样本，使得总体中的每一个元素都有相同的机会（概率）被抽中。
- **分层抽样**（stratified sampling）：也称分类抽样，在抽样之前先将总体的元素划分为若干层（类），然后从各个层抽取一定数量的元素组成一个样本。
- **系统抽样**（systematic sampling）：也称等距抽样，先将总体各元素按某种顺序排列，并按某种规则确定一个随机起点，然后每隔一定的间隔抽取一个元素，直至抽取 $n$ 个元素组成一个样本。
- **整群抽样**（cluster sampling）：先将总体划分成若干群，然后以群作为抽样单元从中抽取部分群组成一个样本，再对抽中的每个群中包含的所有元素进行观察。

## 思考与练习

**思考题**

1.1　请举出统计应用的几个例子。