Statistics 21世纪统计学系列教材

Statistics
An Introduction

统计学概论

贾俊平 编著

中国人民大学出版社

Statistics 21世纪统计学系列教材

Statistics
An Introduction

统计学概论

贾俊平　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计学概论/贾俊平编著. —北京：中国人民大学出版社，2011.9
21世纪统计学系列教材
ISBN 978-7-300-14227-2

Ⅰ. ①统… Ⅱ. ①贾… 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（2011）第173903号

21世纪统计学系列教材
统计学概论
贾俊平 编著
Tongjixue Gailun

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×260 mm　16开本　　**版　　次** 2011年9月第1版
印　　张 13.25插页1　　**印　　次** 2011年9月第1次印刷
字　　数 294 000　　**定　　价** 28.00元

前　言

什么样的教材能让学生更好地理解统计呢？根据笔者对统计的理解及多年的教学经验，尽可能少使用专业的统计术语、少纠缠复杂的公式、少用晦涩的词汇表述统计问题和结果，或许是个不错的选择。本书在写法上做了一些新的尝试：

- 力图把统计方法的思想用书中标题的形式表达出来，尽管这种表达不一定确切。
- 在书中内容的表述上，每种方法都尽量用实际问题引出，而不是从概念开始，尽量不使用专业的统计术语。
- 书中例题的解答直接使用计算机的输出结果，尽可能抛弃手工计算过程，书中例题的计算使用 SPSS 和 Excel 两个软件，但以 SPSS 为主，对软件操作的一些说明放在每章后的附录里。

作为一门应用性很强的学科，多数人的学习目的主要是应用。但初学者学习统计时面临的主要困惑是学完不会用。问题在于学习过程中多把注意力集中在公式和计算上，而忽视对统计思想的理解。学习统计关键在于理解。记住公式，不等于学会统计；学会计算，不等于会用统计。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。有些初学者对统计课程往往感到畏惧，被书中的公式吓倒。实际上，抛开公式照样可以学会统计。特别是在计算机应用已经普及的今天，所有的计算都可以由计算机来完成。只要清楚统计方法使用的前提，理解统计方法的实质，要应用统计并不难。

本书的初衷是作为统计学专业学生的入门课程教材，以替代过去的描述统计内容。作为本专业的学生，在最初接触统计时，应该让他们对统计有一个较全面的认识，了解一些统计思想，为后续的专业课学习奠定基础。当然，本书也可以作为非统计专业学生通开课的教材使用。由于书中的有些提法只是笔者的个人看法，不一定恰当，希望读者多提意见和建议，以便进一步修改和完善。

贾俊平

于中国人民大学

目　录

第 1 章
Chapter 1 统计学研究什么

说出哪些领域应用统计，这很困难，因为几乎所有的领域都用统计；说出哪些领域不用统计，同样也很困难，因为几乎找不到不用统计的领域。

不是每个人都能正确认识统计，也不乏有些人对统计抱有偏见。一提到统计就将其与统计工作联系起来，这样的人也不在少数。**统计学**（statistics）是什么？统计学研究什么？统计学能做什么？学习统计之前，有必要对这些问题有一个初步认识。

1.1 统计无处不在

1.1.1 每个人都离不开统计

了解一些统计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，你还要知道各球队的技术统计结果，等等。要正确阅读并理解下面的一些统计研究结论，就更需要具备一些统计知识：

- 吸烟对健康是有害的。
- 不结婚的男性会早逝 10 年。
- 身材高的父亲，其子女的身材也较高。
- 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。

- 每天服一片阿司匹林会减少心脏病再次发作的机会。
- 身体超重30%会使寿命减少1 300天。
- 每天摄取500毫克维生素C，生命可延长6年。
- 怕老婆的丈夫得心脏病的几率较大。
- 学生在听了莫扎特钢琴曲10分钟后的推理测试会比其听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的？你相信这些结论吗？学点儿统计知识你就会正确理解它们。

了解一些统计知识，对政策制定者或企业管理者来说同样很重要，这有助于他们作出正确决策，也能避免因不懂统计而闹出笑话。一个统计办公室的主管与一些统计学者开会，统计学者抱怨从其他部门收到的一些估计值没有给出标准误差（估计时的误差大小，表示估计的精度），这个主管马上问道："对误差也有标准吗？"健康部门的一位官员看到一个统计学者提供的报告，报告中提到去年由于某种疾病，平均1 000人中死亡人数为3.2人，这位官员对这个数字产生了兴趣。他问私人秘书，3.2个人是如何死法？他的秘书说："先生，当一个统计学家说死了3.2个人时，意味着3个人已经死了，两个人正要死。"

有点儿统计知识就不会闹出这样的笑话来。一位学者写到："假定你是市场部的新任经理，一次广告活动的统计结果摆到了你面前，声称某个结果是'统计显著'的。你如何解释这份报告而又不暴露你对该术语的无知？赶快学点统计，这对你和你的事业都非常有用。"①

1.1.2 几乎所有领域都要用统计

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。

① ［美］埃维森等著，吴喜之等译：《统计学——基本概念和方法》，北京，高等教育出版社，施普林格出版社，2000。

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、冰岛、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差。由此，Schmidt 推断所有各个不同海域内的鳗鱼是由海洋中某个公共场所繁殖的。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国家航空航天局（NASA）就如期发射还是推迟发射发生了争执。天气预报发射时的气温为 31°F。争执的结果采纳了 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°F 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应用。

1.1.3　统计的误用与滥用

大约在一个世纪以前，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。”历史学家 Andrew Lang 说，一些人使用统计“就像喝醉酒的人使用街灯柱——支撑的功能多于照明”。统计常常被人们有意无意地滥用，比如错误的统计定义、错误的图表展示、不合理的样本、数据的遗漏或逻辑错误等等。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：某产品的抽样合格率是 80%。乍看上去没什么问题，但实际上只

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？“调查结果表明……”调查了多少个？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。

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，即使做些统计分析，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；如果决策出了毛病，他们会找出一大堆开脱理由：市场难测，环境突变，竞争激烈，价格下跌，需求疲软，管理不善，成本上升，出口下降……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；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……这样的分析往往脱离了管理问题，对实际决策也未必有用。在工商管理中，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。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，不能歪曲统计，更不能把统计作为掩盖事实的陷阱。

曲解统计是一种常见的现象。在有些人的心目中，使用统计就是寻找支持：他们的心目中可能有了某种“结论性”的东西，或者说希望看到一种符合其需要的某种结论，而后去找些数据来支持他们的结论。如果数据分析的结果与其预期的结论一致，他们就会声称自己是用科学方法得到的结论；如果与预期的不一致，他们要么篡改数据，要么对统计弃而不用。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结论，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。

1.2 统计学研究数据

你问身边的人，GDP（国内生产总值）是什么？CPI（消费者价格指数）是什么？似乎都能说上几句。但要是仔细追问它们究竟代表了什么，就不是每个人能够说清楚的。统计也是一样。你要问一个人统计是什么，似乎没有人不知道，但多数人会将其与统计工作相联系。要问统计学是什么，就不是每个人都能够说明白的，要搞清楚统计学研究什么就更困难了。

1.2.1　有数据的地方就需要统计学

物理学研究的是像热、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。这些学科都有其特定的问题，而且有解决这些问题的各自方法，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。但统计学究竟研究什么？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。按萨维奇（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 这样看来统计似乎被边缘化了，实际上这也正说明统计在各学科领域的独特地位和作用，也表明统计作为一门独立学科的存在而具有的特点。

统计学研究的是数据，只要有数据的地方，就需要用统计方法进行分析。一堆数据不去分析，它也仅仅是数字而已，没什么价值。要分析这些数据，就一定要用到统计方法。没有数据，统计学就没有存在的必要了。

1.2.2　统计学提供研究数据的方法

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——做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，统计学是一门科学、一种工艺和一门艺术这三者的组合。

- **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

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

计结论，就要看统计表达这些结论的技巧和艺术性了。

统计学提供的是一套通用于所有学科领域的数据方法。它是为自然科学、社会科学的多个领域而发展起来的，它为多个学科提供了一种通用的数据分析方法。从某种意义上说，统计仅仅是一种数据分析的方法。与数学一样，统计学是一种工具，一种数据分析的工具。

统计研究数据所使用的方法通常分为**描述统计**（descriptive statistics）和**推断统计**（inferential statistics）两大类。描述统计研究的是数据收集、处理、汇总、图表描述、概括与分析等统计方法。推断统计是研究如何利用样本数据来推断总体特征的统计方法。如何划分其实并不重要，重要的是当你面对所研究的数据时，如何选择适当的统计方法进行分析，并对结果作出合理解释。本书对统计学方法体系的划分如图1—1所示。

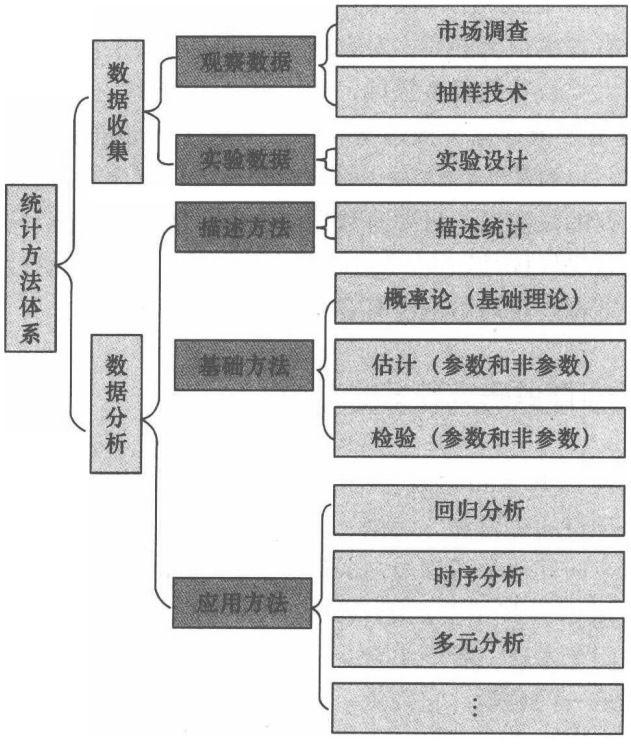

图 1—1 统计学的方法体系

1.2.3 统计方法不是万能的

无论是作为一个工商管理人员，还是一个研究人员，你都会面临大量的数据，也需要分析这些数据，通过分析找到隐藏在数据里面的有用信息。比如，你知道一个地区每个家庭的收入数据，难以给出这个地区收入状况的一个概括性认识；你知道20只灯泡的使用寿命，知道50件产品的合格率，这显然不够，因为你要知道的是这批灯泡的使用寿命，这批产品的合格率。要得到这样的结果，就需要用统计方法去分析。

但是，统计并不是万能的，它不能解决你面临的所有问题。吸烟能引起肺癌，这是一个统计结论，但吸烟为什么能引起肺癌，就不是统计所能回答的问题。统计能帮助你进行数据分析，并从分析中得出某种结论，但对统计结论的进一步解释，则需要专业知识。对于你所面临的数据，统计并不能告诉你应该用什么方法去分析，而是你自己选择你认为适合的方法。这就需要你除了具备统计知识外，还应具备所研究问题领域的专业知识。用灵活的头脑分析数据，是统计方法一贯强调的。

统计方法大多数都有一定的前提。有些人在使用统计方法时，往往忽视这些前提，特别是在社会科学领域，多数数据都是非实验性的，不一定能满足统计的假定。教条地使用统计方法，往往不能得到预期的结果。这不是统计方法的错，而是你的数据不满足统计方法使用的前提造成的。不好的数据，再好的统计方法也无济于事；再好的数据，错误地选择所用的方法，统计也不能得到你要的结论；如果你希望证实早已存在于心中的某种结论，再好的数据，再好的方法，统计也无能为力。

统计不能提供想要的一切技巧和方法。当你把它用于自己的研究领域时，统计能做的和你所需要的之间或许还有差距。针对所研究问题的特殊性，你需要灵活使用统计方法，而不是教条地照搬。必要时需要对统计方法做出修正，以适应所研究的问题。

统计是一种分析数据的工具。当你不需要这种工具时，它对你就是没有用的。当你需要它时，它也只能帮你做它能做的事情。你不能指望统计成为你解决问题的灵丹妙药。

1.3 怎样获得数据

统计学研究数据，就要先有数据。数据是什么？到哪儿去找数据？

1.3.1 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数量，今天与昨天不一样；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”、“上涨股票的数量”、“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”、“上涨股票的数量”、“生活费支出”、“投掷一枚骰子出现的点数”这些变量可以用阿拉伯数据来记录其观察结果，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）。由于分类数据在坐标轴上的位置是任意的，因此也称为无序分类数据。如果类别具有一定的顺序，这样的分类变量也称为**顺序变量**（rank variable），相应的观察结果就是**顺序数据**（rank data），也称为有序分类数据。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）。

1.3.2 怎样得到一个样本

从哪里取得所需的数据呢？对大多数人来说，研究社会科学问题，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，分布在各种报刊、杂志、图书、广播、电视传媒中的各种数据，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，然后从各类中抽样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。