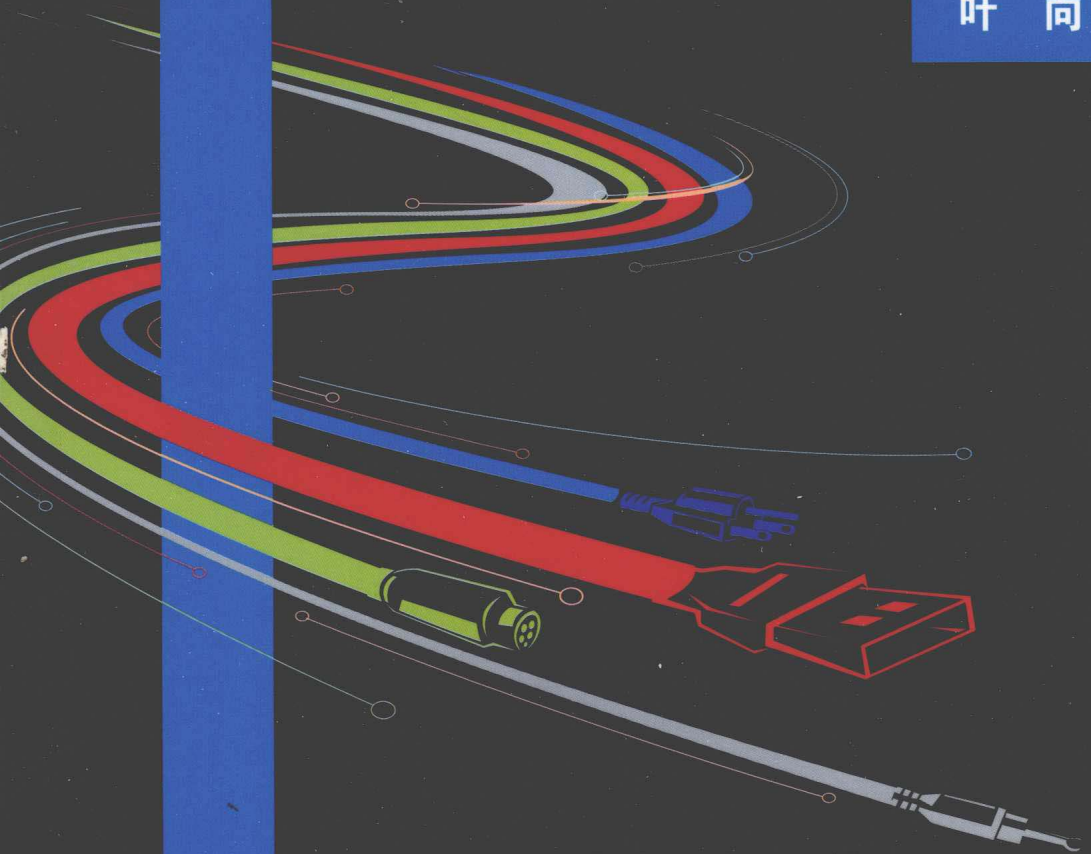
大学计算机基础与应用系列立体化教材

统计数据分析基础教程

——基于SPSS和Excel的调查数据分析

叶 向 编著



中国人民大学出版社

大学计算机基础与应用系列立体化教材

统计数据分析基础教程

——基于SPSS和Excel的调查数据分析

叶 向 编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计数据分析基础教程——基于 SPSS 和 Excel 的调查数据分析/叶向编著.
北京：中国人民大学出版社，2009
（大学计算机基础与应用系列立体化教材）
ISBN 978-7-300-11608-2

Ⅰ.①统…
Ⅱ.①叶…
Ⅲ.①统计数据-统计分析（数学）-高等学校-教材
Ⅳ.①O212

中国版本图书馆 CIP 数据核字（2009）第 243963 号

大学计算机基础与应用系列立体化教材
统计数据分析基础教程
——基于 SPSS 和 Excel 的调查数据分析
叶 向 编著
Tongji Shuju Fenxi Jichu Jiaocheng

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 北京鑫丰华彩印有限公司
规 格 185 mm×260 mm 16 开本 **版 次** 2010 年 2 月第 1 版
印 张 19 插页 1 **印 次** 2010 年 2 月第 1 次印刷
字 数 405 000 **定 价** 28.00 元

总序

随着计算机与互联网应用的普及、信息技术的发展及中小学对信息技术基础课程的普遍开设，针对大学计算机基础与应用教育的方向和重点，我们认为应该研究新的教育与教学模式，使得计算机基础与应用课程摆脱传统的“课堂上课＋课后上机”这种简单、低效的教学方式，逐步转向以实践性教学和互动式教学为手段，利用现代化的计算机实现辅助教学、管理与考核，同时提供包括教材、教辅、教案、习题、实验、网络资源在内的丰富的立体化教学资源和实时或在线答疑系统，使得学生乐于学习、易于学习、学有成效、学有所用，同时减轻教师备课、授课、布置作业与考核、阅卷的工作量，提高教学效率。这是我们建设这套“大学计算机基础与应用系列立体化教材”的初衷。

根据大学非计算机专业学生的社会需求和教育部对计算机基础与应用教育的指导意见，中国人民大学从2005年开始对计算机公共课进行大规模改革，包括增设课程、改革教学方式和考核方式、进行教材建设等多个方面的内容。在最新的《中国人民大学本科生计算机教学指导纲要（2008年版）》中，将与计算机教育有关的内容分为三个层次。第一层次为“计算机应用基础”课程，第二层次为“计算机应用类”课程（包含约10门课程），第三层次纳入专业基础课或专业课教学范畴，形成“1＋X＋Y”的计算机基础与应用教育格局。其中，第一层次的“计算机应用基础”课程和第二层次的“计算机应用类”课程，作为分类分层教学中的核心课程，走在教学改革的前列，同时结合中国人民大学计算机教学改革中开展的其他项目，已经形成了教材（部分课程）、教案、教学网站、教学系统、作业系统、考试系统、答疑系统等多层次、立体化的教学资源。同时，部分项目获得了学校、北京市、全国各级教学成果奖励和立项。

为了巩固我们的计算机基础与应用教学改革成果并使其进一步深化，我们认为有必要系统地建立一套更合理的教材，同时将前述各项立体化、多层次的教学资源整合到一起。为此，我们组织中国人民大学、中央财经大学、天津财经大学、河北大学、东华大学、华北电力大学等多所院校中从事计算机基础与应用课程教学的一线骨干教师，共同建设“大学计算机基础与应用系列立体化教材”项目。

本项目对中国人民大学及合作院校的计算机公共课教学改革和课程建设起着非常关键的作用，得到了各校领导和相关部门的大力支持。该项目将在原来的应用教学的基础上，更进一步地加强实践性教学、实验和考核环节，让学生真正地做到学以致用，与信息技术的发展同步成长。

本系列教材覆盖了“计算机应用基础”（第一层次）和“计算机应用类”（第二层次）的十余门课程，包括：

- 大学计算机应用基础

- Internet 应用教程
- 多媒体技术与应用
- 网站设计与开发
- 数据库技术与应用
- 管理信息系统
- Excel 在经济管理中的应用
- 统计数据分析基础教程
- 信息检索与应用
- C 程序设计教程
- 电子商务基础与应用

每门课程均编写了教材和配套的习题与实验指导。

随着信息化技术的发展，许多新的应用不断涌现，同时数字化的网络教学手段也在发展和成熟。我们将为此项目全面、系统地构建立体化的课程与教学资源体系，以方便学生学习、教师备课、师生交流。具体措施如下：

- 教材建设：在教材中减少纯概念性理论的内容，加强案例和实验指导的分量；增加关于最新的信息技术应用的内容并将其系统化，增加互联网和多媒体应用方面的内容；密切跟踪和反映信息技术的新应用，使学生学到的知识马上就可以使用，充分体现“应用”的特点。
- 教辅建设：针对教材内容，精心编制习题与实验指导。每门课程均安排大量针对性很强的实验，充分体现课程的实践性特点。
- 教学视频：针对主要教学要点，我们将逐步录制教学操作视频，使得学生的学习和复习更为方便。
- 电子教案：我们为教师提供电子教案，针对不同专业和不同的课时安排提出合理化的教学备课建议。
- 教学网站：纸质课本容量有限，更多更全面的教学内容可以从我们的教学网站上查阅。同时，新的知识、技巧和经验不断涌现，我们亦将它们及时地更新到教学网站上。
- 教学辅助系统：针对采用本教材的院校，我们开发了教学辅助系统。通过该系统，可以完成课程的教学、作业、实验、测试、答疑、考试等工作，极大地减轻教师的工作量，方便学生的学习和测试，同时网络的交流环境使师生交流答疑更为便利。（对本教学辅助系统有兴趣的院校，可联系 yxd@yxd.cn 了解详情。）
- 自学自测系统：针对个人读者，可以通过我们提供的自学自测系统来了解自己学习的情况，调整学习进度和重点。
- 在线交流与答疑系统：及时为学生答疑解惑，全方位地为学生（读者）服务。

相信本套教材和教学管理系统不仅对参与编写的院校的计算机基础与应用教学改革起到促进作用，而且对全国其他高校的计算机教学工作也具有参考和借鉴意义。

杨小平

2009 年 6 月

前言

在经济全球化进程不断加快、世界经济联系日趋紧密、市场竞争越来越激烈的今天，一个企业要想赢得市场，求得生存和发展，必须最大限度地减少决策失误的概率。为此，决策者仅凭个人的经验、知识和感觉是很难做到这一点的。在决策过程中，必须充分利用集体的经验、知识、智慧和科学的分析方法，对收集到的数据做出准确、及时的分析并制定正确的决策。掌握数据分析方法和实施工具，是现代管理人才必备的基本技能。

近年来在西方发达国家，信息技术人才和统计应用人才一直排名在就业需求榜的前列，具备计算机知识和统计知识的复合型人才在未来将具有巨大的发展前景和明显的从业优势。

本教材坚持以案例为依托，利用国内外普遍流行的SPSS统计分析软件以及最普遍流行的Excel软件来解决案例中的问题，使读者能更好地利用数据分析方法和实施工具解决实际问题，使数据分析方法在决策中能发挥重要作用，也使学生对统计应用更加感兴趣。

本教材的主要内容包括：问卷设计及数据收集、问卷数据录入与清理、问卷数据基本统计分析（单变量的频率分析、双变量的交叉表分析、多选变量的频率分析、描述统计分析）、假设检验、单因素方差分析以及相关与回归分析。

笔者从事统计应用教学与研究十多年，收集整理了丰富的案例和实践经验。本教材重点突出以下特点：

1. 把SPSS和Excel放在一起介绍。本书的写作基础是安装于Windows XP操作系统上的SPSS 13.0英文版和Excel 2003中文版。

2. 对统计数据分析方法的介绍，力求通俗易懂、简明扼要。

3. 应用实际案例，从实际问题出发，重点介绍软件是如何帮助解决实际问题的，并不强调对软件中每个细节的介绍。

4. 在介绍应用SPSS和Excel软件进行统计分析时，给出了详细的步骤。对输出结果也尽量以读者较容易接受的口语方式进行阐述，而不是用难懂的统计术语讲解。也就是说，过程和说明并重，告诉读者如何根据统计分析结果撰写调查报告。

5. 每章提供了实际案例及相关数据，供读者练习。

本教材的大部分内容，在中国人民大学计算机应用类课程《SPSS基础与应用》及其他课程中讲授过多次，受到普遍欢迎。其教学辅助资源（习题与实验指导、课件、

考题等）也在同步建设中。

经过整整8个月紧张的写作，终于在祖国60周年华诞的喜庆日子里完成了全部书稿。在本书出版之际，要感谢的人很多。首先要感谢信息学院信息系的陈禹教授和方美琪教授，让笔者有机会于1998年3～11月到香港理工大学计算机系进行有关数据仓库与数据挖掘方面的合作研究。1999年2月，陈禹老师让笔者跟李丘副教授一起给全校本科生讲授《统计分析软件SPSS》，从李丘老师那里学到很多，特别是他能够利用统计分析软件解决实际问题（从调查问卷设计、在Excel中录入数据、在SPSS中进行统计分析、在Excel中绘制图表到在Word中撰写调查报告的整个社会调查过程）的讲课方法，给笔者留下了深刻印象。让笔者明白了，学生需要我们能够教给他们解决实际问题的能力。2002年2～6月，笔者去听了李燕琛副教授退休前最后一次给全校硕士研究生开设的《统计分析软件SPSS》和《数据分析软件——Excel高级功能》两门课程。在李老师的课堂上，笔者开始感受到了Excel强大的数据处理与分析功能。2003年8～12月，劳动人事学院的潘锦棠教授给了我一次组织大型的全国性的调查问卷数据录入和统计分析的实践机会。2003年9月，陈禹教授让笔者给信息学院的硕士研究生开设方法课《现代统计方法》，于是笔者去听了统计学院吴喜之教授给全校博士生开设的方法课《统计模型及应用》。笔者非常喜欢吴喜之老师的这门课，从头到尾认认真真听了两遍。从吴老师的课中，笔者知道了统计课程还可以这么教："统计已经渗入到人们的社会、生活、工作等各个领域；以应用为目标学习统计，通过学习获得解决和处理问题的能力；要还统计应用以其本来面目，使得统计变成人人都能够基本上理解和掌握的有用工具"。2009年3月，笔者正式从信息系调入信息技术基础教研室，杨小平教授和尤晓东副教授让我负责《SPSS基础与应用》这门课，包括教学内容、教材编写、教学系统、课件、考题等。要感谢这些给我机会的老师们、使我重新感受到做学生的幸福的教授们以及让我"教学相长"的学生们。

这里还要特别感谢策划本书的中国人民大学出版社的潘旭燕老师，她非常热心，工作认真负责，一直鼓励我将有自己特色的统计应用教学方法写出来，也告诉我很多把书写好的方法。在编写过程中参考了大量的国内外有关文献书籍，它们对本书的成文起了重要作用。在此对一切给予支持和帮助的家人、朋友、同事、同学、有关人员以及参考文献书籍的作者一并表示衷心感谢。

为了使广大读者更好地掌握本教材的内容，加深理解并增强处理问题的能力，我们将本书所有例题和习题的数据文件放在中国人民大学出版社的网站（www.crup.com.cn）的资源中心处，读者可以登录该网站免费下载；为支持教师的教学，本书的作者还将把多年教学中积累的教学课件奉献给老师们。需要的老师，请与本书作者或中国人民大学出版社编辑部联系，电子邮箱：yexiang@ruc.edu.cn 或 panxuyan@263.net。

为方便教师教学和学生自学，我们还将出版本教材的配套辅导书《统计数据分析基础教程习题与实验指导》。

鉴于编著者的水平和经验有限，书中错误和不妥之处在所难免，恳请各位专家和

广大读者给予指正并提出宝贵意见，同时欢迎同行进行交流。编著者联系邮箱是：yexiang@ruc.edu.cn。

叶向
于中国人民大学信息学院
2009年10月

目录 CONTENTS

第 1 章

概　述

本章将介绍统计数据分析中经常用到的一些基本概念，包括什么是统计、统计与计算机的关系、统计软件、变量、数据的收集等。

本章附录将介绍在社会调查方面做得非常出色的“数字 100”市场研究公司，旨在让读者从现实生活中了解调查、在线调查、样本库在线调查、调查报告等内容。

1.1 什么是统计

你想过下面的问题吗?①

(1) 当你买了一台电视，被告知三年内可以免费保修时，你想过厂家凭什么这样说吗？说多了，厂家会损失；说少了，会失去竞争，也是损失。到底这个保修期是怎样决定的呢？

(2) 在同一年级中，同样统计学的课程可能由一些不同教师讲授。教师讲课方式当然不一样，考试题目也不一定相同。那么如何比较不同班级的统计学成绩呢？

(3) 大学排名是一个非常敏感的问题。不同的机构得出不同的结果，各自都说自己是客观、公正和有道理的。到底如何理解这些不同的结果呢？

(4) 如何通过大众调查来得到性别、年龄、职业、收入等各种因素与公众对某件事物（比如商品或政策）的态度的关系呢？

(5) 如何才能够客观地得知某个电视节目的收视率，以确定广告的价格是否合理呢？

其实，这些都是统计应用的例子。这样的例子太多了。因为统计学可以应用于几

① 参见吴喜之编著：《统计学：从数据到结论》(第二版)，1 页，北京，中国统计出版社，2006。

乎所有的领域，包括社会学、新闻调查、精算、农业、动物学、人类学、考古学、审计学、人口统计学、牙医学、生态学、计量经济学、教育学、选举预测和策划、工程、流行病学、金融、水产渔业研究、遗传学、地理学、地质学、历史研究、人类遗传学、水文学、工业、法律、语言学、文学、劳动力计划、管理科学、市场营销学、医学诊断、气象学、军事科学、眼科学、制药学、物理学、政治学、心理学、心理物理学、质量控制、宗教研究、分类学、气象改善、博彩等。当然，大家用不着也不可能理解所有的统计应用，只要能够解决自己身边的统计问题就足够了。

在解决上面所提到的5个问题时，所需使用的大多数统计分析方法将会在本书后面章节中陆续介绍。当然我们的例子并不一定就刚好是上面问题中的具体例子，但至少所使用的分析方法是类似的。

上面的例子并没有明确说出什么是统计。其实很简单，上面的所有例子都要通过各种直接或间接的手段来收集数据（Data），都要利用一些方法来整理和分析数据，最后通过分析得到结论。一句话，统计学（Statistics）是用以收集数据、分析数据并进而由数据得出结论的一组概念、原则和方法。因而有学者也将统计学统称为统计方法（Statistical Method）。比如要得到某电视节目的收视率，可能首先要在该节目播出时，利用电话对看电视的人进行采访，同时问他们在观看什么节目。在得到了被采访的看电视的总人数和其中观看该节目的人数之后，就有可能得到这部分观众中观看该节目的比例，即大致的收视率了。之后还要经过统计分析，评估这个收视率的可信度和代表性等。显然，这是一个收集数据，然后通过分析数据得到结论的简单例子。

1.2 统计、计算机与统计软件

现代生活越来越离不开计算机了。最早使用计算机的统计当然更离不开计算机了。事实上，最初的计算机仅仅是为科学计算而设计和制造的。大型计算机的最早一批用户就包含统计。现在，统计仍然是进行数字计算最多的用户。当然计算机现在早已脱离了仅有数字计算功能的单一模式，而成为百姓生活的一部分。计算机的使用，也从过去必须学会计算机语言发展到只需要“傻瓜式”地点击鼠标；结果也从单纯的数字输出发展到包括漂亮的表格和图形在内的各种形式。①

统计软件的发展，也使得统计从统计学家的圈内游戏变成了大众的游戏。只要输入你的数据，点几下鼠标，做一些选项，马上就得到令人惊叹的漂亮结果了。人们可能会问，是否傻瓜式统计软件的使用可以代替统计课程？当然不是。数据的整理和识别，方法的选用，计算机输出结果的理解都不像使用傻瓜相机那样简单可靠。有些诸如法律和医学方面的软件都有不少警告，不时提醒你去咨询专家。但统计软件则不那么负责。只要数据格式无误、选项不矛盾而且不用零作为除数就一定给你结果，而且几乎没有任何警告。另外，统计软件输出的结果太多。即使是同样的方法，不同软件

① 参见吴喜之编著：《统计学：从数据到结论》（第二版），8～9页，北京，中国统计出版社，2006。

输出的内容也不一样。有时同样的内容名称也不一样。这就使得使用者大伤脑筋。即使是统计学家也不一定能解释所有的输出。因此，就应该特别留神，明白自己是在干什么，不要在得到一堆毫无意义的垃圾之后还沾沾自喜。

统计软件的种类很多。有些功能齐全，有些价格便宜，有些容易操作，有些需要更多的实践才能掌握。还有些是专门的软件，只处理某一类统计问题。面对太多的选择往往给决策带来困难。这里介绍最常见的几种。

1. SPSS

这是一个很受欢迎的统计软件。它操作容易，输出漂亮，功能齐全，价格合理。它也有自己的程序语言，但基本上已经“傻瓜化”。对于非专业统计工作者它是很好的选择。

2. Excel

它严格说来并不是统计软件，但作为数据表格软件，必然有一定统计计算功能。而且凡是有 Microsoft Office 的计算机，基本上都装有 Excel。但要注意，有时在装 Office 时没有装“数据分析”的功能，那就必须装了才行。当然，画图功能是已经具备了的。对于简单分析，Excel 还算方便，但随着问题的深入，Excel 就不那么“傻瓜”了，需要使用宏命令来编程，这时就没有相应的简单选项了。多数专门一些的统计推断问题还需要其他专门的统计软件来处理。

3. SAS

这是功能非常齐全的软件。尽管价格相当不菲，但许多公司，特别是美国制药公司都在使用，这多半因为其功能众多和某些美国政府机构一些人的偏爱。尽管现在已经尽量“傻瓜化”，但仍然需要一定的训练才可以进入。也可以对它编程，但对于基本统计课程则不那么方便。

当然，还有很多其他的软件，如 S-plus、R 软件、Minitab、Matlab 等，没有必要一一罗列。其实，聪明的读者只要学会使用一种“傻瓜式”软件，使用其他的软件也不会困难；最多看看帮助和说明即可。如果只有英文帮助，那还可以顺便提高英文阅读能力。学习软件的最好方式是需要时在使用中学。

1.3 为何要使用 Excel 来学习统计

前面介绍的 SPSS、SAS 等统计分析软件，在市面上的普及率很低，在个人计算机上，占有率甚至连千分之一都不到。[①] 其原因如下：

1. 价格昂贵

价格高达上万到几十万。通常，教育单位、国家机关或大型研究单位才有能力购买；一般学生、个人或中小企业都因负担不起而不可能购买。

2. 学习困难

这几个软件，若仅是要进行操作而取得分析结果，在有人指导下或有相关优秀书

① 参见杨世莹编著：《Excel 数据统计与分析范例应用》，2 页，北京，中国青年出版社，2006。

籍来参考的情况下并不困难。但因学习的人不多，受过完整训练的师资本来就不多，市面上可用的书籍也不是很多。所以，对大部分人来说，学习起来还是很困难的。况且很多单位所使用的还是英文版，更加大了学习的难度。

事实上，SAS 软件较适合学习统计理论时使用，无论操作或编写程序均较为困难。而 SPSS 软件则较适合分析市场调查的数据，通常不需要编写程序，即可进行操作。

3. 报表难懂

这几个高级的统计软件，所分析出来的统计结果相当多，很多是读者学统计时从来就没看过的统计量，根本就不知道其作用。而事实上，一般统计应用所使用的统计结果并不很复杂，主要目的是要让大部分人能看得懂。计算并列出这些无法拿来应用的统计结果，不只是浪费资源，更是对学习者信心的一大打击。

如果统计结果只有少数几个人能看懂，其作用将大打折扣。例如，市场调查的分析结果，如果只有少数几个研究人员能看懂，老板或主管又怎能有信心根据一个完全不懂的结果来做决策？又如政策支持度的结果，只要简单的几个百分比大概也就够了，列出一大串的相关统计量，不仅管理者看不懂，各报刊杂志的记者及主编也看不懂，刊登出来后，读者也看不懂，如何引起共鸣？政策又将如何修正或推行？

如果用这类高级统计分析软件来学习统计，因普及率过低，将来离开学校后很容易面临没有适当软件可用的窘境，纵有一身绝技，也难以发挥。

由于微软的 Office 已相当普及，并且广泛地应用于工商企业及个人使用领域。要想在一台个人计算机上找到 Excel，要比找到 SPSS 或 SAS 软件容易得多，而且 Excel 具有易学易懂的特性。所以本书决定除了使用原有的 SPSS 统计软件作为工具外，还以 Excel 为工具来帮助读者学习统计[①]。虽然 Excel 并没有被归类为统计软件，并且其与统计有关的函数和数据分析工具的功能是绝对无法与 SPSS 或 SAS 统计软件相提并论的，但对绝大多数人而言已经足够了。

生活在“信息时代”中的人比以前任何时候都更频繁地与数据打交道，Excel 就是为现代人进行数据处理而定制的一个工具。无论是在科学研究、医疗教育、商业活动还是家庭生活中，Excel 都能满足大多数人的数据处理需求。Excel 拥有强大的计算、分析、传递和共享功能，可以帮助用户将繁杂的数据转化为有用的信息。伟人说“实践出真知”，在 Excel 中，不但实践出真知，而且实践出技巧。

1.4 变量及其分类

1. 变量定义

变量（Variable）是用来描述总体中成员的某一特性。

在搜集数据的过程中，需要搜集各类的变量。例如，性别、年龄、职业、教育程

① 这本书是编著者为“计算机应用类”课程《SPSS 基础与应用》编写的，原来的老师都用 SPSS 授课，但编著者根据多年来的教学经验，认为 Excel 也很适合这门课程。

度、收入等人口统计变量。又如，为了预测明年的销售量，所搜集到的数据如广告费、人事费、销售人员数等，也都是一种变量。

在现实生活或自然界中的一些现象，通常都不是单一变量可以描述得很清楚的。例如，要描述某一个人，仅使用性别变量，说他（或她）是男性（或是女性），肯定是无法说明白的。但随着变量（例如年龄、肤色、头发、身高、体重、种族等）的增加，可以逐渐描述得更清楚一些。

2. 变量类型

（1）定性变量

定性变量（Qualitative Variable）也称离散变量（Discrete Variable）或分类变量（Categorical Variable）。例如，使用的手机品牌、就读的班级、宗教信仰、参加的社团、喜好的运动、最常饮用的饮料类别、最喜欢的歌手、最喜欢的影星、民族、党派，均属定性变量。

性别为男或女，只是描述性别的现象。将男性标示为1或将女性标示为2，仅是为了方便计算机处理，并没有任何大小或倍数的关系。直觉上读者可能认为2>1，2为1的两倍。但若转为口语，将变为：女大于男，女为男的两倍。任谁都不可能同意。而且若其均值为1.69，也不具任何意义。最多也只能知道此次调查的女性样本比男性样本多些而已。

（2）定序变量

定序变量也称有序变量。例如，成绩：优［5］、良［4］、中［3］、及格［2］、不及格［1］；产品质量：特等品［3］、一等品［2］、二等品［1］；文化程度：小学［1］、中学［2］、大学［3］、研究生［4］；职称：教授［4］、副教授［3］、讲师［2］、助教［1］；评价：非常重要［5］、重要［4］、普通［3］、不重要［2］、非常不重要［1］。

定序变量，更多的时候是将其看作定性（分类）变量的一种，可进行频率分析和交叉分析。

定序变量，只有大小先后的关系，无倍数关系。例如“非常重要”用5表示，“非常不重要”用1表示，只能说5比1重要而已，无法说“非常重要”是“非常不重要”的5倍。但为了方便，研究上，有时也将其视为连续的数字数据，而直接求其均值、标准差等统计量。

（3）定量变量

定量变量（Quantitative Variable）也称数量变量或连续变量（Continuous Variable）。例如，成绩、年龄、收入、国民生产总值、体重、身高、智力、温度等均属定量变量。

定量变量（连续变量）有大小和倍数的关系，例如：3 000>1 000，3 000是1 000的3倍。

在实际应用中，变量类型一般只分为定性变量和定量变量两大类。

1.5 数据的收集[①]

1.5.1 怎样得到数据

每天翻开报纸或打开电视，就可以看到各种数据，比如高速公路通车里程、股票行情、外汇牌价、房价、流行病的有关数据。当然还有国家统计局定期发布的各种国家经济数据、海关发布的进出口贸易数据等。从这些数据中，各有关方面可以提取对自己有用的信息。显然，这些间接得到的数据都是二手数据。

获得第一手数据并不像得到二手数据那么轻松。某些企业每年至少要花三四千万元来收集和分析数据。他们调查其产品目前在市场中的状况和地位，并确定其竞争对手的态势。他们调查不同地区、不同阶层的民众对其产品的认知程度和购买意愿，以改进产品或推出新品种以争取新顾客。他们还收集各地方的经济、交通等信息，以决定如何保住现有市场和开发新市场。市场信息数据对企业是至关重要的，他们很舍得在这方面花钱。因为这是企业生存所必需的，绝不是可有可无的。更多的例子可参见本章附录。

上面所说的数据是在自然的未被控制的条件下观测到的，称为观测数据（Observational Data）。而对于有些问题，比如在不同的医疗手段下某疾病的治疗结果有什么不同，在不同的肥料和土壤条件下某农作物的产量有没有区别，用什么成分可以提高某物质变成超导体的温度等。这种在人工干预和操作情况下收集的数据就称为试验数据（Experimental Data）。

1.5.2 个体、总体和样本

要想了解北京市民对建设北京交通设施是以包括轨道运输在内的公共交通工具为主还是以小汽车为主的观点，需要进行调查。调查对象是所有北京市民，调查目的是希望知道市民中对这个问题的不同看法各自占有的比例。显然，不可能去问所有的北京市民，而只能够问一部分，并且根据这一部分的观点来理解整个北京市民的总体观点。在这个例子中，单个北京市民称为调查的对象（Object）；而他们的观点称为（这个调查问题的）个体（Element，Individual，Unit）；称所有北京市民对这个问题的观点为一个总体（Population），总体是包含所有要研究的个体的集合；而调查时问到的那部分市民的观点（也就是部分个体）称为该总体的一个样本（Sample），是总体中选出的一部分。当然，也有可能试图调查所有的人，那叫普查（Census），比如人口普查。有人喜欢把作为调查对象的北京市民称为个体，但每个市民还有其他诸如身高、体重、教育程度等无数特征，这些都不是我们调查的目标。因此，为了强调我们调查的目的，市民的观点才应称为个体。

在抽取样本时，如果总体中的每一个个体都有同等机会被选到样本中，这种抽样

① 参见吴喜之编著：《统计学：从数据到结论》（第二版），13～18页。