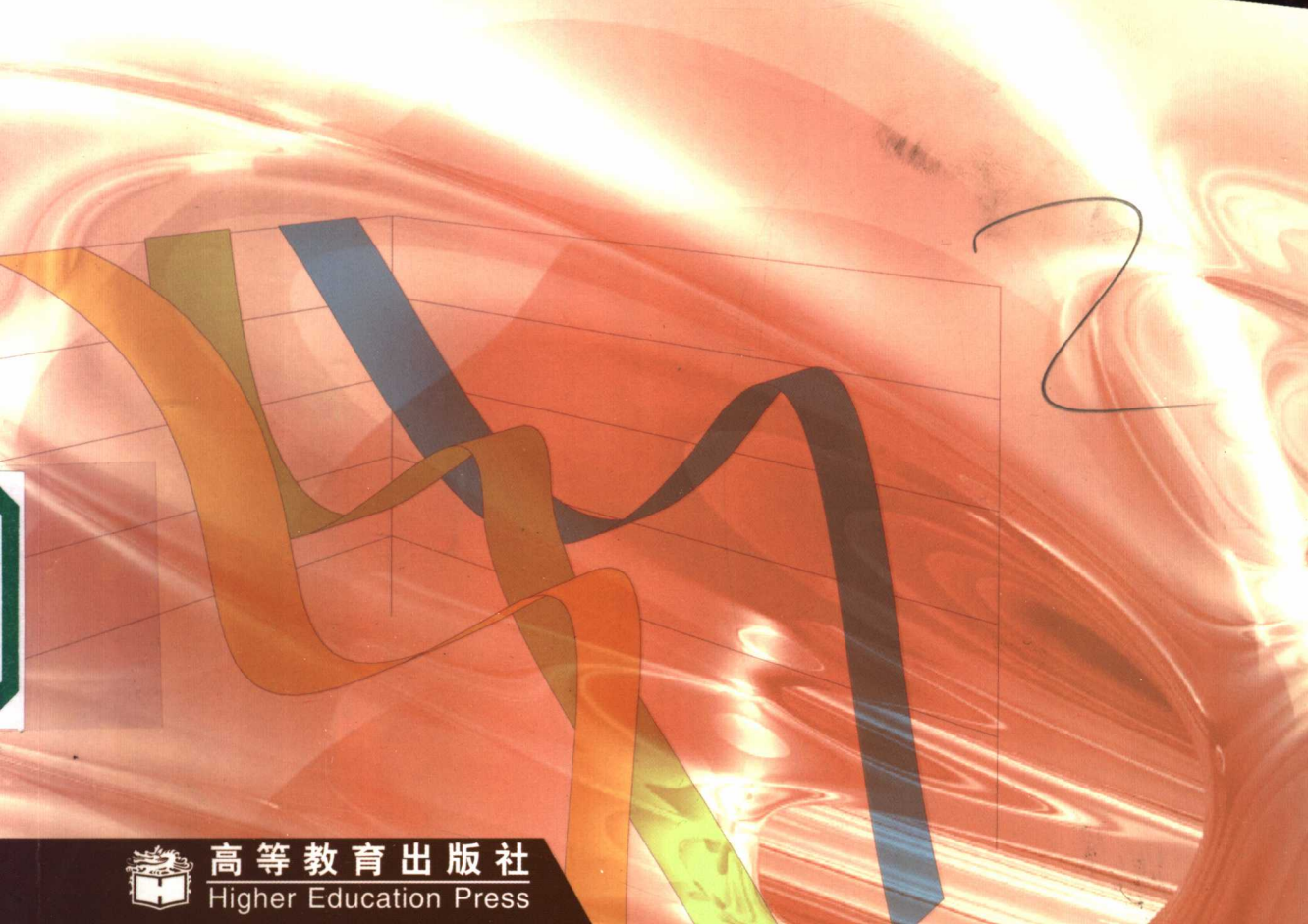
高等学校文科计算机课程系列教材

数据统计分析与实践
——SPSS for Windows

□ 袁克定 编著



高等教育出版社
Higher Education Press

高等学校文科计算机课程系列教材

数据统计分析与实践

——SPSS for Windows

袁克定 编著

高等教育出版社

内 容 提 要

本书是根据教育部高等教育司组织制订的高等学校文科类专业《大学计算机教学基本要求(2003年版)》,通过教育部高等学校文科计算机基础教学指导委员会组织编写的,是作者十多年来数据统计分析课程的教学和科研工作实践的结晶。

本书从教育类专业读者的角度出发,结合作者对SPSS软件的教学和应用研究的经验,本着循序渐进的原则,在介绍数据统计分析工具的同时,将统计学的知识融入其中,详细介绍数据统计的新方法和新观点,且对应于每章均有综合的应用实例。其内容包括:现代教育研究方法概述、数据统计分析工具软件、数据的编码和编辑、数据整合、变量的描述统计分析、均值差异性的假设检验、样本分布的非参数检验、相关分析与回归分析、聚类分析与判别分析、因子分析等。

本书可作为高等学校教育、心理、经济等类各个专业的本科生和相关专业研究生的教材,亦可作为相关领域研究人员的参考书。

图书在版编目(CIP)数据

数据统计分析与实践/衷克定编著. —北京:高等教育出版社,2005.4
ISBN 7-04-016537-6

Ⅰ.数… Ⅱ.衷… Ⅲ.统计数据-统计分析(数学) Ⅳ.0212.1

中国版本图书馆CIP数据核字(2005)第027265号

策划编辑 陈红英 **责任编辑** 陈红英 **封面设计** 李卫青 **责任印制** 孔 源

出版发行	高等教育出版社	**购书热线**	010-58581118
社 址	北京市西城区德外大街4号	**免费咨询**	800-810-0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	北京蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京市卫顺印刷厂		
开 本	787×1092 1/16	**版 次**	2005年4月第1版
印 张	17	**印 次**	2005年4月第1次印刷
字 数	370 000	**定 价**	19.80元

物料号 16537-00

作者介绍

衷克定 北京师范大学教授，博士生导师。现任北京师范大学教育技术学院副院长，兼任教育部高等学校文科计算机基础教学指导委员会委员。长期从事计算机基础教育以及计算机在教育技术与教育管理方面的应用研究。参加与美国新闻总署合作及世界银行贷款项目多项，承担国家级重点科研项目一项，主持完成多项校级科研项目。在国家核心期刊发表论文十余篇，编写计算机类教材5部。多次荣获北京市优秀教学成果奖。

前　言

在信息社会中，人的信息素养是生存的基本条件之一。对学生信息的获取、表示、存储、处理和应用等能力的发展将是当前信息技术教育的核心内容。遵照教育部文科计算机基础课程教学指导委员会的《大学计算机基础教学基本要求》（以下简称《基本要求》），在开设必修的计算机公共基础课程后，应当增设更加结合所在专业需要的后续课程。本书就是结合教育、心理等类专业的有关知识和应用，对数据统计分析方法和手段进行学习的后续课程教材。

人类科学及其理论的发展与科学研究能力和水平的提高是密切相关的。尤其是当今交叉学科的发展极其迅速，教育、心理等学科原属于思辨的、定性的、纯理论的研究已经向实证的、定量的、数据分析与挖掘的新的研究模式转化。新的研究模式与学科方法论密切结合形成新的研究方法。可以说，科学严谨的研究方法不仅能够提高研究水平，对于科学理论本身也是有重大的意义的。

通过大量的国际交流我们看到，相当多的发达国家和地区的理论界在研究方法方面不仅基础深厚，而且在应用创新发展方面都保持了一定的领先地位。同发达国家和地区相比我们确实有一定的差距。我国高等学校的教育及其相关专业的本科生、研究生的理论研究水平普遍还存在一些方法论的问题。即使是国内知名大学的硕士研究生，在硕士论文的设计和实施方面仍然对科学研究存在理解上的偏差。例如，调查问卷设计时不考虑数据采集，使数据编码产生困难和数据信息丢失。数据统计分析手段停留在低层次、低水平，不能真正挖掘出有意义的结论和规律。同时我们还看到，当前在科学研究方法方面总体水平的差距的客观存在。缩小差距的途径则必须遵从科学规律，从思想体系方面理解，而不是采用形而上学式的照搬和模仿。肤浅的应用而忽略其深层的意义必将给科学研究带来损失。从科学方法论的高度理解研究方法的学习是极其必要的。

科学研究依赖于信息的支持，但信息本身的意义并不仅仅意味着数据量的庞大，而在于如何从大量的数据中提取出有参考价值的规律，这才是促进社会、经济、文化和科技发展的基本要素。这种规律的提取必须按照某种科学的模式，即一定的数学方法来完成，而模式和方法还需要一定的有效手段来支持。当前流行的比较有效的手段就是使用计算机来实现数据的统计分析，选择一个理想的数据统计分析平台显然是至关重要的。

美国 SPSS 公司推出的统计分析软件包 SPSS（Statistical Package for Social Science）是一个有十几年历史，并不断创新的优秀数据统计分析工具，其 DOS 版的 SPSS/PC 早在 20 世纪 80 年代中后期就成为在全球范围流行并受到学术界广泛认同的软件了。近年来，Windows 版的 SPSS for Windows 又充分占领了视窗环境的应用市场。它集“问卷设计、数据统计和精辟分析”于一体，其统计结果清晰科学、易学易用。国际学术界默认：凡是用 SPSS 统计分析

的结果，在国际学术交流中可以不必说明算法。其权威性和信誉度是世界公认的。因此，该软件系统为平台形成一门理论紧密联系实际的计算机教学后续课，是具有长远战略意义的。本书就是通过对该系统的教学，同广大教育及其相关领域的学习者共同探讨教育科学研究方法的知识与技能，并将本人在十余年教学和科研工作中总结的经验一并奉送给广大读者。

全书本着循序渐进的原则，从简到繁，由浅入深，不求全责备，而突出重点。这些特点完全是为了照顾到一般性的教学过程，使该课程的学习有一条清晰的线索和明确的路径，在达到一定水平后，学习者完全可以脱离教师的指导而自学。

本书是在北京师范大学教育学院、心理学院和经济学院等各个专业的本科生和全校各个专业方向的研究生的多年教学工作的基础上总结精选的，因而在许多方面都将突出教育学科及其相关专业的特征，甚至在一些例题中都引用了教学过程中比较优秀的研究课题的内容。

作　者

2005 年 1 月 1 日于北京师范大学

目　录

第1章　现代教育研究方法概述

教育科学研究的水平同其研究方法是密切联系在一起的。近年来，思辨与实证相结合、定性与定量相结合的研究方法已经成为教育科学及其相关学科的主导研究方法。在讨论和学习研究方法之前，有必要将研究领域的特点认真加以考虑，以便遵循教育及其相关学科的特有的研究规律，建立科学的研究方法体系。

1.1　现代教育科学研究方法的特殊性

现代教育科学的研究兼有社会科学与自然科学两重性质，与以往我们单纯研究自然科学问题相比有许多方面不同。

1. 被研究者与研究者（被试与主试）的特殊性

在教育教学活动中，被研究者是有意识、有情感、有心理活动的人，其活动具有一定的不确定性，即不一定完全按照自己的真实状况做出反应。这与做一次物理或化学实验有所不同。对被研究者的研究过程不能没有他们的主动配合。而研究者也同样具有自己的社会背景、价值观念等，同样在分析、判断方面存在一定的主观意志作用。因此在研究活动中调查和数据采集工作必须排除来自不同方面的干扰和误差。

2. 研究过程的特殊性

在教育教学活动中，被研究者与研究者将不可避免地出现交互。不论是访谈还是问卷指导语，都会直接影响到被研究者的心理变化。这种主、客体相互影响、相互作用的关系将在实验中产生无关因素，形成无关变量，造成对实验研究的干扰。这种主、客体相互干扰的现象称为“实验者效应”。尤其是在两者之间的社会、政治、价值观具有一定差异时，影响更为显著和突出。所以实验设计时不仅应当控制被研究者与研究者交互造成的不良影响，还应当在数据处理时控制无关变量的影响。

3. 研究方法的特殊性

绝大多数教育心理及其相关研究的研究对象都是人，研究方法仅仅能够达到对信息数据输入和输出的研究，即常说的黑箱方法。对人脑思维活动的生物和物理的研究还远未达到解释心理活动的水平。因此，对心理活动的研究往往是通过推测和判断，而对某活动的心理结构成分的不了解将不能描述心理变化的因果关系。在此应当强调，在相当多的教育心理等方面的研究中，相关性研究是无法得出因果关系的。

出于上述三个方面的特殊性，现代教育研究方法有其自身的研究规律，需要在不断的研

究中加以总结和归纳。

1.2 数据采集的基本要求

1.2.1 关于数据的管理规范

数据统计分析研究首先将面对的就是数据。SPSS 12.0 for Windows 的数据管理是有比较严格的规范的，由于是原文的系统，许多术语需要统一。下面列举一些常用术语称谓，如图 1.1 所示。

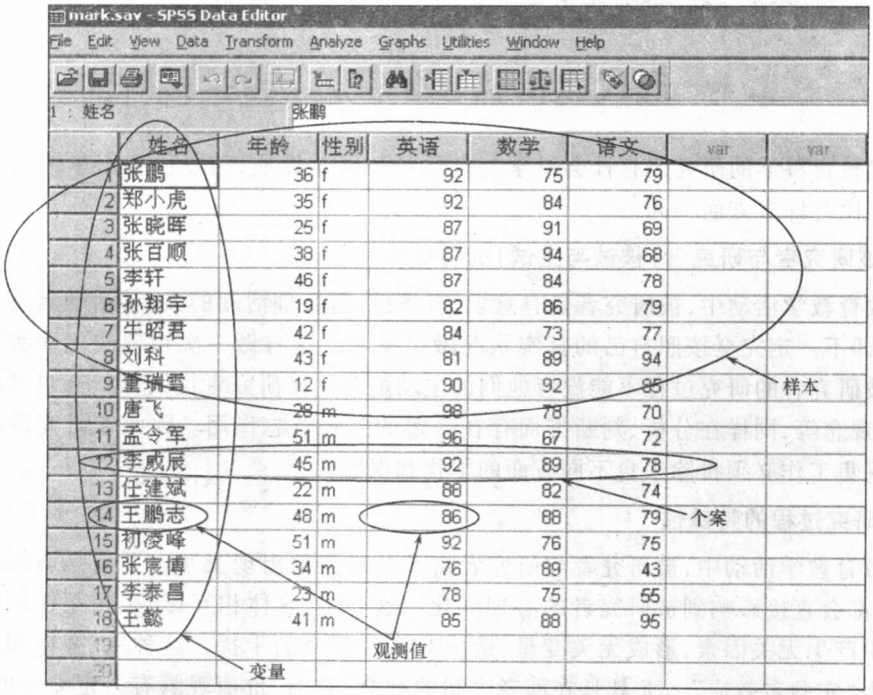


图 1.1 数据的管理规范

1. 变量与观测值(Variable & Observed Value)

变量是描述研究对象的定量化属性的，如编号、性别、身高、成绩等，在数据表格中表现为“一列”。

观测值是变量中的一个数值，是描述某研究对象的特定属性值的，如“34”号、“男”生、“1.74”米身高、“93”分等，表现为一个单元格中的数值。

2. 个案与样本(Case & Sample)

个案是所研究个体对象的全部属性，如某人、某学校、某地区的所有信息，在数据表格中

表现为“一行”。

样本表示的是具有某共同属性的群体的全部信息。例如,男生、女生,学校甲、学校乙,班级1、班级2等的全部信息。样本含多个个案,在数据表格中表现为“n行”。

3. 样本因素或分组变量(Sample Factor or Grouping Variable)

样本因素是用于区分不同样本属性的变量。样本因素通常可以区分为不同水平。例如,性别的不同水平是“男”和“女”,班级的不同水平是“1班”、“2班”、“3班”和“4班”。按照不同的因素水平可以将个案区分为不同的群体(也就是样本),故也称为“分组变量”或自变量。

1.2.2 数据的问卷采集

数据采集的方法通常有观察法、访谈法、问卷调查法等。各种方法最终都将落实在数据表格中,形成数据文件,经过编码的数据将形成统计分析的基础。下面将重点以调查问卷法为讨论对象。

调查问卷是数据采集的常用方法之一。特点是简便、明确,标准化程度高,数据易处理。一份好的问卷经过长期使用、检验和修订,产生常模后将成为具有普遍意义的量表。

问卷制作的技术性很强。尽管许多学生在自己知识的基础上已经可以制作简单的问卷了,但不是所有的问卷或其中的所有数据使用和处理起来都很理想,尤其是要进一步使用数据统计分析工具进行处理时,对数据采集和调查问卷就有更高的要求了。在此,着重对需要进行计算机统计分析系统工具处理的调查问卷的设计制作加以论述和分析。

样本(Sample)的属性是由变量来描述的,而变量的编码将是数据处理的基础工作。编码将是数据统计分析的首要技术。但根据以往的经验,许多问题就发生在调查问卷的数据编码上。经常发生的问题或错误可以归纳为问题的非标准化使编码困难、问题测度低不利于深入分析、多选题的编码问题、测量虚伪的问题回答、增加题量以提高测度五个方面。

1. 问题的非标准化使编码困难

数据采集需要一定的可编码性。对于单项选择题应当设计互斥的选项,选项必须覆盖全部可能情况,主观与客观题分开等。某些问卷可能出现问题的非标准化,如表1.1所示。

表1.1 问卷调查题举例一

问题:学校开设的第二课堂你参加了哪一个?
① 数学兴趣小组
② 外语兴趣小组
③ 理化生兴趣小组
④ 美术兴趣小组
⑤ 舞蹈队
⑥ 合唱队
⑦ 计算机兴趣小组
⑧ _____(在上述之外的请自行填写)

表1.1出现了不能直接进行编码的选项。因为前7个选项不能覆盖全部可能情况,而最后一个的内容与前7个的数据显然不同,它是主观内容;前7个是客观内容,所以这种选择题应当更改。或者穷举所有课程类别供选择,或者将前7类之外的全部定义为一类,作为一个选项,而不要由被试者自行填写。

2. 问题测度低,不利于深入分析

数据的调研是为了揭示事物的内部属性,应当尽量提高数据测量的测度,即后面将介绍的数据测度(Measurement)。有时研究者为了让被研究者回答问题简单而设计了一些测度较低的完全标准化问题,而使得到的数据不能更深入地定量化研究事物的属性,这就对进一步研究造成了困难,如表1.2所示的情况。

表1.2 问卷调查题举例二

问题:你所在的院(系)有多少台可供本科生上课使用的计算机?
① 500台以上
② 400台~500台
③ 300台~400台
④ 200台~300台
⑤ 100台~200台
⑥ 100台以下

事实上,院系有多少台计算机,尤其是有什么范围数量的计算机并不是研究的目标,而研究者是希望调查本科生人均计算机数量。而上述的问题是不可能得到该结论的。故该问题实际可以变成:

问题:你所在的院(系)有_____台可供本科生上课使用的计算机。

由被研究者填写具体数量,即便有一定的误差也不会超过100台,这样只要进一步了解了学生人数就可以得到人均计算机数了。

3. 多选题的编码问题

在调查问卷中有些复杂问题不可避免地会出现多选问题。多选问题有些是重要性等同的,有些是重要性有优先级的。这类问题尽管被研究者回答时没有什么困难,但编码时就非常麻烦了,如表1.3所示。

表1.3 问卷调查题举例三

问题:你认为学生应当参加哪些学校管理?
① 宿舍管理
② 食堂管理
③ 教学管理

（续表）

④ 学生就业管理
⑤ 学籍管理
⑥ 参与学校领导决策
⑦ 学校财政管理
⑧ 党团组织管理

倘若问题变为：

问题：你认为学生应当参加下列哪些学校管理？按照重要性的先后顺序回答最重要的3个。

该问题的回答虽然会比前面的简单一些，但仍然有很多种可能。

类似问题在实际调查问卷中是很常见的，但数据处理就比较难了。根据一些问卷调查和数据处理的经验，可供参考的解决方法有两种。

（1）按选项分解

对于多选问题，由于没有限制被研究者必须选择几个，所以结果是很多的。如上述的问题，被研究者最多可能选择8个，最少可能一个也不选。如果用1、2、3、4、5、6、7、8分别表示选择情况，输入的字符串将无法进行处理。

通常的解决办法是：将该变量分成8个分变量，每个分变量对应一个选项。选择该项的用“1”表示，不选择该项的用“0”表示。这样处理的结果对于频度分析、分布检验等都十分方便。

（2）按优先级分解

对于有优先级的多选问题，虽然限制了被研究者选择的个数，但可能的结果仍然是很多的。该类问题如果也用1、2、3、4、5、6、7、8分别表示每个选项，输入的字符串也将无法进行处理。

解决的办法是：将该变量的8个选项对应1～8共8个数值，用3个分变量分别对应三种优先级（优先级最高的、次高的和最低的）。被研究者在不同的级别分别选择了某选项，都可以对应不同的数值。这样处理的结果在三种不同的优先级中对于不同的选项进行频度分析、分布情况分析等也是十分方便的。

4. 测量虚伪的问题

在问卷调查中应当适当设置一些测量虚伪问题，以便甄别被研究者在回答问题时是否有不看问题，一律答A或E（五段量表），或处于某种心理而一律肯定或一律否定。

问卷设计时适当地将某些选择题的选项进行倒序处理，即将A、B、C、D和E的顺序进行颠倒处理，以便抵消部分有倾向性的选择答案。

5. 增加题量以提高测度

对于标准化问题，被研究者的回答选择只是选择A、B、C、D和E。这种数据的测度是比较低的（定类变量或定序变量），只能在编码时编为相应的1～5。一个方面的问题，往往仅

根据一道这种标准化试题不仅测度低,而且误差大。应当对一个方面的问题酌情设计多个标准化试题,最后将多个试题的结果求和或求平均值,必要的时候还可以对不同的选择题赋予不同的权重。这样对不同的被研究者的观点、意志、态度,就可以得到量化的区分了,从而提高了问题的测度。

1.3 事物属性的定性与定量描述

数据变量的测度是描述事物属性程度的量化标准。测度高表示对事物量化测量程度高。对测度的衡量通常分为四种类型。

1. 定类变量(Nominal Variable)

又称为标称变量。其观测值既无大小之分,又无等级或次序之分,仅仅是一种标称或类别。如性别、部门单位或国家地区等。例如,该变量的观测值表达可以是男、女,或用"0"、"1"表示的男、女。其观测值彼此之间的次序是没有特定意义的,颠倒次序来定义也可以。该类变量既可以用数值型变量表示,又可以用字符串型变量表示,但无论如何表示都不能进行加、减、乘、除等数学运算。这类变量是测度最低的变量。

2. 定序变量(Ordinal Variable)

又称为顺序变量。其观测值尽管大小没有特定意义,但属于顺序计量类型,适合于按照顺序排列的变量。如名次、级别、职务、评价水平等变量。例如,其表达可以是用数值1、2、3、4分别表达的"教授"、"副教授"、"讲师"、"助教"等。其观测值彼此之间的次序是有一定意义的,打乱定义将产生错误。该类变量既可以用数值型变量表示,又可以用字符串型变量表示,但同样也是不能进行加、减、乘、除等数学运算的。这类变量的测度比标称变量高。

3. 定距变量(Interval Variable)

又称为区间变量。其观测值具有等级和次序之分。即观测值的大小和次序具有可比性,可以反映观测值之间的大小差异,但该类变量的观测值是在特定区间上有意义,超出该区间将没有意义。例如,学生的身高有一定的可比性,但身高为"零"没有实际意义,太高也是没有意义的,该类变量只可以用数值型变量表示。这种变量的测度比定序变量和定类变量都高。

4. 定比变量(Scale Variable)

又称为比例变量。按照一定间隔、比例来计量数据的变量类型。如长度、质量、重量等变量,其观测值"零"也是有定义的。观测值之间可以进行加、减、乘、除的四则运算,只适合于数值型变量。这种变量的测度是最高的。

上述四种测度的变量分别表示了不同的测量等级,等级低的应用范围受局限,等级高的应用范围广泛。测度低对事物属性的描述就是定性的,即只能研究到事物的某些属性的存在情况,不能研究到事物属性的量化程度。测度高对事物属性的描述既可以是定性的,也可以是定量的。由于定类变量和定序变量的测度低,属于定性描述的变量。定距变量和定比