

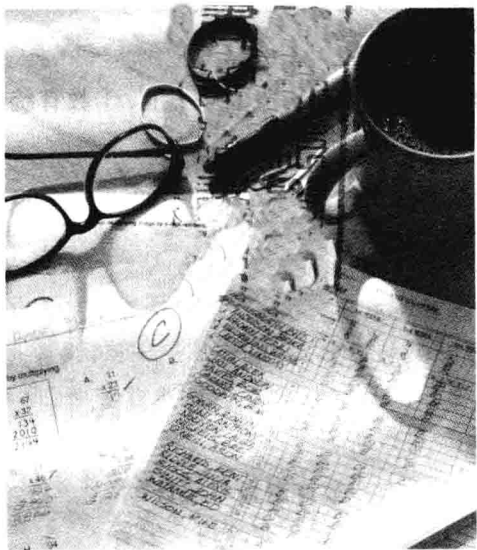


教育测量与统计

王 斌 龚玲梅【编著】

◆ 苏州大学出版社

虞山教育丛书



教育测量与统计

王斌 龚玲梅【编著】

◆ 苏州大学出版社

图书在版编目(CIP)数据

教育测量与统计/王斌,龚玲梅编著. —苏州:苏州
大学出版社,2003.9
(虞山教育丛书/何东亮主编)
ISBN 7-81090-155-9

I. 教… II. ①王…②龚… III. ①教育统计②教育-测量 IV. G40-051

中国版本图书馆 CIP 数据核字(2003)第 084934 号

教育测量与统计

王 斌 龚玲梅 编著

责任编辑 许周鹤

苏州大学出版社出版发行

(地址:苏州市干将东路200号 邮编:215021)

常熟高专印刷厂印装

(地址:常熟市元和路98号 邮编:215500)

开本 850×1168 1/32 印张 29.25(共六册) 字数 725 千

2003年9月第1版 2003年9月第1次印刷

ISBN 7-81090-155-9/G·64 定价:84.00元

(共六册)

苏州大学版图书若有印装错误,本社负责调换
苏州大学出版社营销部 电话:0512-67258802

《虞山文库》总序

许 霆

虞山，以“十里青山半入城”的姿态与文化历史名城常熟融合，对常熟文化的形成与发展影响巨大，并进而成为常熟的别名和常熟文化的标志。商末，周太王长子泰伯、次子仲雍让国避奔江南，建立“勾吴”，泰伯、仲雍相继成为首领。仲雍死后葬于常熟卧牛山，仲雍又名虞仲，山遂以虞为名。春秋时期的言偃生于常熟，北学中原，成为孔门七十二贤人中的“十哲之九”，晚年回归故土传道讲学，“道启东南”，“文开吴会”，死后葬于虞山东麓。仲雍和言偃，昭示了常熟文化源头的深邃和博大，标志着吴地早期文明曙光终于开启出一个区域文化的圣地。

常熟文化发展绵延不绝。南北朝昭明太子的“文选”，开始了常熟文化发展的自觉时代；自唐代陆器高中状元，常熟历史上出过8个状元483个进士；北宋时郑时性嗜书好藏书，开了明清时代常熟出版、藏书兴盛的先河，赵琦美与脉望馆、瞿氏与铁琴铜剑楼、毛晋与汲古阁都对中国文化史作出过重大贡献；元代的黄公望，以其绘画理论和创作开创了明清山水画的新纪元；明清之际以王翬为首的“虞山画派”、以钱谦益为代表的“虞山诗派”、严激的琴学理论和虞山琴派，还有虞山书派、虞山印派等，都达到全国一流水平，影响一时风气；近代以来，黄人的文学史论、曾朴的谴责小说等，表明常熟文化在求新变革时吐故纳新的活力。基于这种深厚的文化底

蕴，常熟当代文明，更是显示了勃勃生机。

常熟高等专科学校就坐落在人文荟萃的虞山脚下，接受着常熟深厚博大的传统文化和生生不息的现代文明的滋养。学校在与地方经济和文化的互动发展中获得不竭的创造精神，塑造崭新的主体形象，确立自身的价值目标。学校有一批人文和理工学人，更是为常熟的传统文化甘泉所浸润，以虞山的人格精神塑造品行，用致远的人生追求敬业乐教。宋人朱熹在《丹阳公祠堂记》中说言偃为人，“必当敏于闻道而不滞于形器，岂所谓南方之学，得其精华者，乃自古而已然也耶”。明末龚立本纂修《常熟县志》15卷，其中《风俗志》说常熟士人“贫不负诺，富不易交，吐纳风流，意气横溢。表人胜士，千里命驾者比比，人物显晦殊途，或矜名节，或树勋庸，或敦学术”，这都揭示了常熟传统文化中独特的人格精神。这种精神是常熟文化生生不息的产物和动力，也是常熟文化走向现代文明的底蕴和财富。常熟高等专科学校的学人，在市场经济发展的大潮中，自觉地从立足的虞山福地的传统人格精神中汲取营养，坚持自强不息、敏捷好学、达美达诚的学风，在学术园地和育人园圃播种、耕耘和收获，形成了一批学术探索和教学研究成果，这是可喜可贺的。

常熟虞山，由于其深厚的文化积淀和不断的文化传承，已经成为一种文化创造的意象。正因为如此，我们愿意把这批初步的成果以“虞山文库”为名，汇集出版。我们无意创造学派，而意在宣示精神，表明当代学人对传承人文传统、创造现代文化使命的一种担当。我们衷心希望这项工作能够继续下去，能有更多的成果充实文库，承当起当代学人文化建设的重任。

2003年4月

目 录

第一章 绪 论	
第一节 教育测量与统计概述·····	(1)
第二节 教育测量与统计中的几个概念和符号·····	(6)
第二章 教育测量的理论及测验的编制	
第一节 教育测量的基本理论·····	(10)
第二节 教育测验的编制·····	(15)
第三节 有效测验的必要条件·····	(21)
第三章 学绩测验	
第一节 学绩测验概述·····	(28)
第二节 标准化学绩测验·····	(30)
第三节 教师自编课堂测验·····	(37)
第四章 数据的初步整理	
第一节 数据的来源、种类及其分类·····	(43)
第二节 统计表和统计图·····	(45)
第三节 频数分布表与频数分布图·····	(48)
第五章 描述统计	
第一节 集中量·····	(53)
第二节 差异量·····	(64)
第六章 抽样分布与假设检验	
第一节 抽样与抽样分布·····	(72)

第二节	假设检验	(81)
第七章	平均数的显著性检验	
第一节	总体平均数的显著性检验	(87)
第二节	平均数差异的显著性检验	(91)
第八章	方差分析	
第一节	方差分析的原理及步骤	(99)
第二节	完全随机设计的方差分析	(105)
第三节	随机区组设计的方差分析	(109)
第九章	χ^2 检验	
第一节	χ^2 检验概述	(114)
第二节	配合度检验	(115)
第三节	独立性检验	(119)
附 表		
附表 1(1)	正态分布表	(127)
附表 1(2)	正态分布表	(128)
附表 1(3)	正态分布表	(129)
附表 1(4)	正态分布表	(130)
附表 1(5)	正态分布表	(131)
附表 2	t 值表	(132)
附表 3(1)	F 值表	(134)
附表 3(2)	F 值表	(136)
附表 3(3)	F 值表	(138)
附表 3(4)	F 值表	(140)
附表 4	χ^2 值表	(142)
主要参考文献	(144)

第一章 绪 论

第一节 教育测量与统计概述

随着现代技术的发展和进步,人们不但对物理作出了越来越精确的测量,而且也不断尝试对人的知识、思维、能力、学术水平、成就等心理特征进行测量。现代教育理论的发展,更加注意强调人的素质教育,强调发挥人的主观能动性,强调因材施教。要检验教育的效果,离不开对被教育者的评价。其中最为重要的一环就是采用教育测量和教育统计的方法对教育过程和结果进行衡量,从而为教育评价提供科学依据。

一、教育测量的性质与功能

美国心理与教育测量学家桑代克提出:“凡是存在必有数量,凡有数量就可测量。”人们比较容易理解的是物理测量。如日常生活中的度、量、衡都是司空见惯的;自然科学中的各种测试,由于测量技术和测量工具的进步和完善,可以进行直接的测量。

与物理测量相比较,教育测量的对象是复杂得多的人。正如世界上没有两片完全相同的树叶,世界上也没有两个完全相同的人。所谓千人千面,人的内在的心理特征也是千差万别的,人与人之间的个别差异,很难像物理差异那样区分明显。同时,教育测量很难排除一些无关因素的影响,诸如知识水平、教学条件、师资水平、情绪、健康状况、主试等多方面因素或多或少地影响到教育测量的结果,使之出现各种误差。教育测量比物理测量要模糊得多,

也困难得多。

什么是教育测量呢？简单地说，教育测量就是根据一定的法则用数字对教育效果或过程加以确定。进行教育测量必须要有相应的测量工具，测量工具的好坏，直接影响到测量的效果。教育测量的主要工具是测验，测验旨在于对教育效果进行科学的测量。教育测量的主要内容就是测验的编制和使用。学校、社会中测量的应用是十分广泛的，如高考、会考、招工考试、分班、入学、成人高考、公务员选拔等，不同的测量目的有不同的测量要求，不同的测量又有不同的编制要求和不同的分数评定体系和标准。

测验总是由一组题目构成，题目是测验的基本元素，好的测验必须是优良题目的集合。比如，一个用于选拔性目的的测验就应当把具有不同学业水平的考生区别开来，如果在某道题目上所有考生都得满分或不得分，这道题目就失去了区分不同学业水平考生的效用。选好题目是进行教育测量的一项重要工作。测验应有统一的标准和尺度，有较高的信度和效度。测验的结果一般都用分数或等级来表示，测验分数的评定、等级的划分及对各个测验分数的解释等问题，也是教育测量的重要内容。

当今世界是充满决策的世界，如果学校能对学生的心理和教育效果进行全面系统的测量，根据测量所得到的结果反馈于教学，那么，必然可在实际的教学、教育决策中发挥很大的功用。教育测量的功能，在学校教育中主要表现在：

1. 因材施教

教育的一条基本原则就是因材施教，要在教育过程中贯彻这一原则，教师必须了解自己的学生。教师了解学生的途径有两条，一是凭借主观经验，一是借助于测验的客观测量。而主观经验有时并不可靠，常受到教师各种因素的影响，如教师的情绪、好恶、成见等。因此，为了更准确、客观地了解自己的学生，使用测验对学生进行测量并依据测量的结果来了解学生是必不可少的。只有这

样,教师才能针对学生的具体情况作出相应的合理安排,依据学生能力和已有知识水平的个别差异作出适当的教学决策,比如说编班、分组、课后个别辅导等,才能做到有的放矢。

2. 选拔人才

教育测量是一种选拔人才的手段,随着社会化生产的发展,在人才选拔方面依靠个人经验已不能适应社会的需要,高效、准确的测量结果已成为人才选拔的重要依据,如高考、研究生考试、特色班、公务员考试等,这大大提高了人才选拔和职业能力匹配的效率。

3. 评价教学

测验的评价功能在教学评估中既可面向教师与教学方法,也可面向学生与学业成就。测验可以选拔与评定学校管理人员与教师,做到优胜劣汰,还可以评价和鉴定教材和教学方法,从而提高教学质量。根据单元考试、期中和期末的综合考试,结合学生平时的表现,可以对学生的学业成就作出评定,决定学生的等级成绩,可以使教师了解学生,也能使学生自我了解和自我评价。

科学的教育测量要求每一个教育工作者都应该掌握教育测量的基本原理和方法,但由于实际的原因,教育测量对许多教师来说还是一门不熟悉的科学。要促进教育事业的发展,科学合理地甄选人才,做到因材施教,教育工作者必须了解和熟悉教育测量。

二、教育统计的性质和内容

统计学是研究统计原理和方法的科学。具体地讲,它是研究如何搜集、整理、分析反映事物总体信息的数字资料,并以此为依据,对总体特征进行推断的原理和方法。

统计学分为两大类。一类是数理统计学。它主要是以概率论为基础,对统计数据数量关系的模式加以解释,对统计原理和方法给予数学上的证明。它是数学的一个分支。另一类是应用统计学。它是数理统计原理和方法在各个领域中的实际应用,如工业

统计学、医学统计学、商业统计学等。数理统计学与应用统计学有着密切的关系。数理统计学是应用统计学的理论基础,应用统计学是数理统计学的实践和应用。应用统计学为数理统计学提出了实践中需要解决的新问题,从而促进数理统计学的内容进一步丰富和发展。

教育统计学是运用数理统计的原理和方法,研究教育问题的一门应用科学。它的主要任务是研究如何搜集、整理、分析由教育调查、教育测量和教育实验所获得的数字资料,并以此为依据,进行科学推断,揭示教育现象所蕴含的客观规律。但是从研究内容来说,教育调查和教育实验课题的提出,内容的界定,对象范围的确定,假设的建立,结论的得出以及分析,却不是教育统计学的研究任务,因为这些问题还需要依靠与研究内容有关的教育专业知识来解决。而教育统计学只能提供各种统计方法的应用条件和统计结果的解释。至于统计原理和方法的数学证明及公式推导,不是它的主要任务。

教育统计学研究的内容,从具体应用的角度来分,可以分成描述统计、推断统计和试验设计三部分。

1. 描述统计

对已获得的数据进行整理、概括,显现其分布特征的统计方法,称为描述统计。通过教育调查和教育实验获得了大量的数据,用分组、编表、绘图等统计方法对之进行归纳、整理,以直观形象的形式反映其分布特征;通过计算各种特征量,来反映它们分布上的数字特征。例如,计算集中量(如算术平均数、中位数、众数、加权算术平均数、调和平均数等)来反映它们的集中趋势;计算差异量(如全距、四分位差、百分位差、平均差、方差和标准差、差异系数等)来反映它们的离散程度;计算相关量(如积差相关系数、等级相关系数、点二列相关系数、二列相关系数等)来反映一个事物两种特性之间变化的一致性程度。这些均属于描述统计范围。其目的

在于将大量零散的、杂乱无序的数字资料进行整理、归纳、减缩、概括,使事物的全貌及其分布特征清晰、明确地显现出来。

2. 推断统计

根据样本提供的信息,运用概率的理论进行分析、论证,在一定可靠程度上,对总体分布特征进行估计、推测,这种统计方法称为推断统计。推断统计的内容包括总体参数估计和假设检验两部分。例如,对总体参数值(如总体平均数、总体标准差、总体相关系数)的估计;对总体参数或总体参数之差(如总体平均数之差、总体方差之差等)的假设检验,都属于推断统计的范围。其目的在于根据样本的已知情况,在一定概率的意义上估计、推断总体的性质,并标明可能发生的误差。

3. 试验设计

试验者为了揭示试验中自变量与因变量的关系,在实验之前所制定的试验计划,称为试验设计。其中包括选择怎样的抽样方式;如何计算样本容量;确定怎样的实验对照形式;如何实现实验组和对照组的等组化;如何安排实验因素和如何控制无关因素;用什么方法处理及分析试验结果;等等。

这三部分内容,不是截然分开,而是相互联系的。描述统计是推断统计的基础,推断统计可以通过样本信息估计、推测总体,从已知情况推测、估计未知情况。良好的试验设计才能使我们获得真实的有价值的数据,对这样的数据进行统计处理才能得出正确的结论。而良好的试验设计又必须以统计原理为依据,符合统计方法的要求,才能对实验结果进行统计处理。由于试验设计是以多元统计为基础的比较复杂的问题,为了简化难度,在这里我们就不再进行过多的探讨。

第二节 教育测量与统计中的几个概念和符号

一、测量和测验

测量就是用一定规则给事物属性指派数字或符号的过程。测量包括三个元素：事物属性、法则、数字或符号。事物属性是测量的对象或目标。对教育测量而言，所测的是个体的外显行为或外在表现特征，比如说数学和语文成绩。但我们真正感兴趣的却是隐含于所测得的外显行为之中的个体潜在特质水平，如数学思维能力等。法则是所依据的规则和方法，是测量的关键。使用好的法则可以得到可靠的测量，使用差的法则就会得到不可靠甚至是错误的结果。法则的好坏取决于它是否符合客观事物属性的规律以及是否易于制定和操作。数字或符号是代表某一事物或事物某一属性的量。由于数字具有区分性、等级性、等距性、代数运算的封闭性等特点，所以通过测量所得的数，不仅可以表示事物属性的类别、大小、多少，而且还可以在一定的条件下由数的运算而对事物的属性进行推测。

测验实质上是行为样本的客观的和标准化的测量。这一定义包含了三个基本要素，即行为样本、标准化和客观的评价指标。行为样本是选取有代表性的行为来考察个体在相应行为领域的行为特征。当个体在某一测验中的反应很恰当地反映出测验所要测的东西时，该测验就为我们提供了有用的信息。标准化是指测验在编制、实施、记分、分数解释方面依据一套系统的程序。只有这样，测验才有统一的标准，使不同人的测验结果具有可比性。同时，可以减少无关因素对测验结果的影响，从而使之更为准确、可靠。客观的评价指标要求测验题目能够在一定程度上反映所要测量的内容，要求测验有较高的信度、效度、难度和区分度。

二、随机变量

为了解释随机变量的概念,先介绍随机现象和随机事件。

具有以下三个特征的现象,称为随机现象。第一,一次试验有多种可能结果,其所有可能结果是已知的;第二,试验之前不能预料哪一种结果会出现;第三,在相同的条件下可以重复试验。例如,抛一枚硬币,有两种可能结果,不是正面朝上,就是反面朝上。究竟哪面朝上,事先不能预料。相同的条件下可以重复抛多次,这种现象是随机现象。随机现象的每一种结果叫做一个随机事件。这些随机事件在一次试验中,可能出现,也可能不出现,而在大量重复试验中,它们的发生却具有一定的规律性。假如,硬币的正面朝上称为随机事件 A ,反面朝上成为随机事件 B 。在抛一次硬币时,事件 A 可能发生,也可能不发生,但如果重复抛许多次,事件 A 的发生就会具有某种规律性,即它出现的概率接近 $1/2$ 。我们把能表示随机现象各种结果的变量称为随机变量。统计处理的变量都是些随机变量。例如,学生的身高、体重、性别、智商、考试成绩、人数、年龄、工资等。本书一般用大写的 X 或 Y 表示随机变量。为了表示区分不同试验或不同测量方法得到的随机变量,有时用 X_1, X_2, \dots, X_i 或 X_n 表示一系列随机变量,而用 Y_n 表示另一列随机变量,或简写为 X, Y 表示。每个随机事件往往表现为一种数值。对于不是以数值表示的随机事件,可以将之数量化。例如,可将高考录取和未录取分别用 1 和 0 表示,将品德评定的优、良、中、差等级分别用 4、3、2、1 表示。

三、总体和样本

总体是我们所研究的具有某种共同特性的对象的全体。总体中的每个单元成员称为个体。从总体当中抽取一部分个体,称为总体的一个样本。当对总体某种特性进行研究时,由于某种原因,不可能将总体中的每一个个体一一进行观测,往往需要抽取一部分个体,作为样本进行观察、分析,然后根据样本所获得的信息,在

一定可靠程度上推断总体。

当总体所包含的个体数目有限时,这一总体称为有限总体。而总体所包含的个体数目无限时,称为无限总体。例如,我们研究某区高三英语毕业考试成绩,这是有限总体。当我们研究8岁女童的身高,以古代人、现代人、中国人、外国人作为测查对象时,则这里的8岁女童的全体可以看作是无限总体。有限总体内所包含的个体数目,一般用 N 表示。在实际研究工作中,总体选择有限的,还是无限的,以及对于有限总体来说,总体内应当包含多少个体,这都应依据研究的问题所欲推断的范围而定。例如,从某区随机抽取7所学校,对学生的成绩进行调查,此时,这7所学校的学生可以作为该区的样本,也可以作为这7所学校的总体。总体和样本是相对的。

样本中所包含的个体数目称为样本的容量,一般用 n 表示。样本中个体数目大于30的称为大样本,等于或小于30的称为小样本。在对数据进行统计处理时,大样本和小样本所用的统计方法不同。

四、统计量和参数

样本上的数字特征是统计量。也就是说,根据教育调查或实验获得的数据所计算出来的能够描述这组数据各种特征的数量是统计量。例如,描述一组数据集中趋势的一种统计指标称为平均数(用 \bar{X} 或 \bar{Y} 表示);描述一组数据分散程度的一种统计指标称为标准差(用符号 S 或 SD 表示);描述某一事物两种特征之间关系的统计指标称为相关系数(用符号 r 表示);等等。这些都是统计量。

总体上的各种数字特征是参数。也即反映总体上各种特征的数量是参数。例如,反映总体集中趋势的一种统计指标称为总体平均数(用 μ 表示);反映总体内个体间分散程度的一种统计指标称为总体标准差(用 σ 表示);反映某一事物的两种特征之间在总

体内变化关系的一种统计指标称为总体相关系数(用 ρ 表示);等等。这些都是总体参数。

在进行统计推断时,就是根据样本统计量来推断总体相应的参数。如根据样本的平均数推断总体的平均数;根据样本的标准差推断总体的标准差;根据样本的相关系数推断总体的相关系数,或者根据样本的某种统计量指标的差数,推断总体相应指标差数的参数。

第二章 教育测量的理论及测验的编制

教育测量学作为教育科学的分支,具有很强的应用价值。但是,理论与实践是相辅相成的,实践推动了理论的发展,理论又指导了实践。在实际工作中应用教育测量,必须了解教育测量的各种理论,每一种理论都有自己的特点和适用范围,都有自己的长处和不足。

教育测量的基本工具是测验。通过测验,才能将研究对象的各种特征有效地反映出来。测验的编制是一项重要的工作,良好的测验必须客观、公正、有效,并要进行严格的实施。只有这样,才能保证测量的准确性。

第一节 教育测量的基本理论

早在 20 世纪 40 年代前后,欧美等国家的一些测验统计理论专家开展了测验的统计数学模型的研究,提出了测验信度、效度、项目的难度与区分度等指标及其经典的统计分析方法,为测验研究提供了理论模型及统计分析方法,进一步丰富了教育测量学科内容,并在 20 世纪 50 年代前后形成了经典测验理论体系。从 20 世纪 60 年代以来,除了经典测验理论(CTT)进一步拓展外,还创立了多种现代测验理论,其中项目反应理论(IRT)和概化(GT)是近段时期世界上最有影响的两种理论。这三大理论相互补充,共同支撑起教育测量的理论大厦。在这里,我们对经典测量理论和