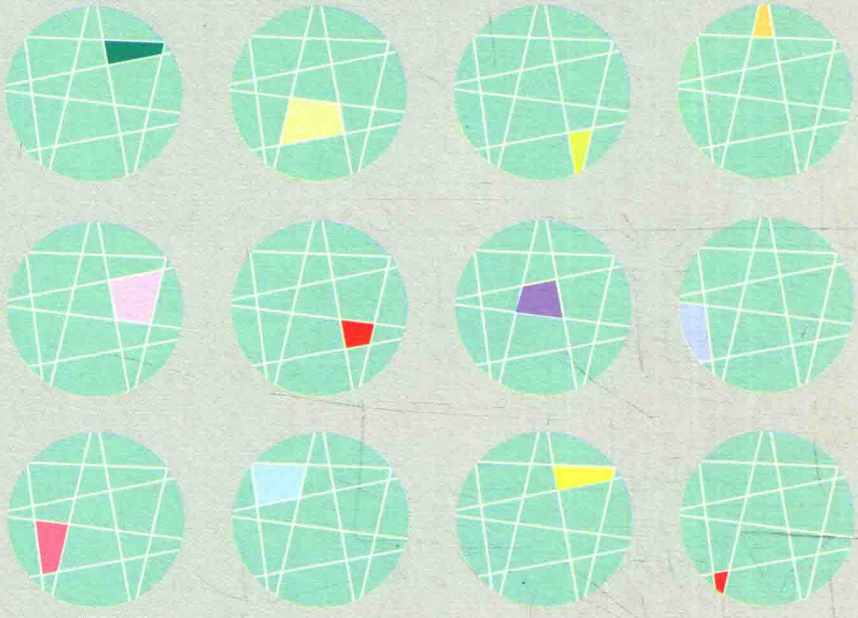
中国教育质量监测与评估丛书



基础教育质量监测
抽样设计与数据分析

张丹慧 张生 刘红云 著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

基础教育质量监测
抽样设计与数据分析

Sampling Design and Data Analysis of National Assessment of Educational Quality

张丹慧 张生 刘红云 著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

基础教育质量监测抽样设计与数据分析/张丹慧等著.
—北京：北京师范大学出版社，2015.8
（中国教育质量监测与评估丛书）
ISBN 978-7-303-19293-9

Ⅰ.①基… Ⅱ.①张… Ⅲ.①基础教育—教育质量—质量管理—研究 Ⅳ.①G632.0

中国版本图书馆 CIP 数据核字(2015)第 173427 号

营销中心电话 010-58805072 58807651
北师大出版社学术著作与大众读物分社 http://xueda.bnup.com

JICHU JIAOYU ZHILIANG JIANCE CHOUYANGSHEJI YU SHUJUFENXI

出版发行：北京师范大学出版社 www.bnup.com
北京市海淀区新街口外大街 19 号
邮政编码：100875

印　　刷：北京联兴盛业印刷股份有限公司
经　　销：全国新华书店
开　　本：730 mm×980 mm 1/16
印　　张：12.25
字　　数：200 千字
版　　次：2015 年 8 月第 1 版
印　　次：2015 年 8 月第 1 次印刷
定　　价：48.00 元

策划编辑：郭兴举 陈红艳　　责任编辑：戴 轶
美术编辑：袁 麟　　装帧设计：锋尚制版
责任校对：陈 民　　责任印制：马 洁

丛书编委会

序　言

随着我国九年义务教育的全面普及，全面提升质量成为义务教育改革发展的核心任务。促进教育质量提升的重要前提是全面、准确地了解和把握教育质量状况。然而，很长一段时间以来，我们既不能客观全面评价教育质量状况，也不能有效诊断教育存在的问题及其根源。为此，《国家中长期教育改革和发展规划纲要(2010－2020年)》提出，要建立教育质量监测、评估体系，定期发布测评结果。《中共中央关于全面深化改革若干重大问题的决定》指出，要“委托社会组织开展教育评估监测”。开展教育质量监测，对教育质量进行科学、全面、有效的评价，已成为实现我国基础教育科学发展、内涵发展的重大举措和战略任务。

为适应我国基础教育改革与发展的需要，北京师范大学联合国内外多家单位组建了中国基础教育质量监测协同创新中心。这是我国第一个、也是目前唯一一个通过教育部认定的教育学和心理学领域的国家协同创新中心。中心着眼于构建中国特色、国际先进的基础教育质量监测体系，围绕基础教育质量标准体系建设、监测体系构建与制度设计、大数据采集与分析平台建设、教育决策支持系统、教育质量提升工程等重要任务展开协同攻关，努力在科学研究、学科建设、人才培养、社会服务等方面取得突破与进展，为全面提升我国教育质量，促进亿万儿童与青少年的全面、个性发展提供有力支撑。

协同创新中心成立以来，集聚了国内外相关领域的优秀人才和资源，创新了合作研究工作机制，持续开展了相关领域的协同创新研究，取得了积极进展。中心先后研制了义务教育阶段语文、数学、科学、品德、体育、艺术等学科领域的监测标准和工具，并通过教育部审定。研制了

《国家义务教育质量监测方案》，经教育部审核通过并向社会正式发布。继2007～2014年连续承担国家义务教育质量试点监测任务后，2015年组织开展了我国首次国家义务教育质量监测。此外，积极探索国家监测结果的运用机制，开展基于证据的教育决策咨询服务，持续多年对所有样本省、县进行了“一对一”的监测结果反馈，为国家和地方教育决策、教育教学改进提供了客观依据以及针对性建议，取得了良好的效果。

除了满足国家基础教育质量监测的需求之外，协同创新中心还致力于服务地方开展基础教育质量监测的需要。中心开展了基础教育质量监测能力建设项目，先后组织专家通过各种方式对全国20多个省相关人员进行多轮次基础教育质量监测通识培训和专业技术培训，从一般性理论、理念到具体的技术、方法，为参训人员呈现了基础教育质量监测的“完整图像”，对转变相关人员观念、提升专业水平起到了良好的促进作用。同时，重点对部分队伍力量相对较强的省级监测机构进行针对性培训，全程指导这些机构独立开展本地基础教育质量监测，有力地提升了这些机构的专业能力。通过培训和指导，一些省级监测机构，如重庆市，已能独立、完整地开展市域内基础教育质量监测，为所在地区的教育决策、区域教育质量的提升与均衡发展提供了有力的支持。

随着各地基础教育质量监测机构的不断建立，对相关专业支持和指导的需求也不断加大，对协同创新中心也提出了新的要求。据中心初步统计，截至2014年底，全国共有22个省(区、市)成立了省级基础教育质量监测机构。这些机构大多依托省级督导评估中心、教研室、教科院、评估院、招生考试院等部门而成立，相关从业人员队伍普遍存在数量不足、年龄偏大、专业水平较低等问题，整体力量相对薄弱，难以有效开展本地区基础教育质量监测，迫切需要专业的支持和指导。在与协同创新中心交流过程中，地方监测机构有关从业人员纷纷表示，教育质量监测是一项专业性强、技术含量高的工作，希望中心能系统介绍开展监测相关的理论、技术和方法，分享自身开展这项工作的实践经验，以引领

和指导各地开展好这项工作。

为更好地满足各地对开展基础教育质量监测的不同需求，指导各地有效开展有关工作，协同创新中心围绕“什么是基础教育质量监测”“如何开展基础教育质量监测”两大主题组织编写了这套丛书。丛书包括《国际基础教育质量监测实践与经验》《基础教育质量监测工具研发》《基础教育质量监测抽样设计与数据分析》《基础教育质量监测报告撰写与结果应用》《大规模学业成就调查的开发：理论、方法与应用》《教育认知诊断评估理论与技术研究》六本。其中，《国际基础教育监测实践与经验》主要回答“什么是基础教育质量监测”的问题，《基础教育质量监测工具研发》、《基础教育质量监测抽样设计与数据分析》、《基础教育质量监测报告撰写与结果应用》则主要回答“如何开展基础教育质量监测”的问题。《大规模学业成就调查的开发：理论、方法与应用》与《教育认知诊断评估理论与技术研究》在理论与实践层面都进行了探讨。

具体而言，《国际基础教育质量监测实践与经验》主要包括基础教育质量监测概述、世界各国与国际组织开展的基础教育质量监测实践、中国开展基础教育质量监测的探索等内容。《基础教育质量监测工具研发》以监测工具开发的程序与规范为主线，详细介绍了学业成就测试工具、相关因素调查问卷开发的科学流程与具体要求。《基础教育质量监测抽样设计与数据分析》对抽样的基本概念、抽样方法、数据处理与分析、标准划定、测试题目质量检验分析方法等内容进行了介绍和说明。《基础教育质量监测报告撰写与结果使用》介绍了监测结果报告的种类、推动监测结果发挥最大使用效益的策略和方法等内容。《大规模学业成就调查的开发：理论、方法与应用》从介绍全球范围内大规模学业成就调查项目入手，详细阐述了当前大规模学业成就调查开发的理论基础、常用方法与结果应用。《教育认知诊断评估理论与技术研究》主要探讨了认知诊断测验编制原理、项目属性标定新方法、认知诊断新模型、认知诊断测验的信度和效度，以及计算机化自适应诊断测验等重要的前沿研究热点。

本套丛书既可作为相关专业与方向培训的教材，也可供有关从业人员自学之用。丛书从酝酿选题、内容编排到成书出版，经历了三四年的时间。期间，协同中心组织编著人员围绕有关内容进行了反复多次讨论，并在重庆、江西等省市围绕相关内容开展了系统的培训和试点，请参训人员提出了意见和建议，并进行了不断修改完善。有别于“急就章”式的著作，本套丛书凝聚了协同创新中心多年来的探索实践经验，当中的很多观点和内容均为中心这些年来不断思考、积淀的结果，是中心为广大正在从事或有志于从事基础教育质量监测工作的读者奉献的诚意之作。

本套丛书的出版得到了教育部基础二司和世界银行的有关项目资助，世界银行还为这套丛书的编写无偿提供了相关的材料，北京师范大学出版社，特别是策划编辑陈红艳女士为此倾注了大量心血，在此一并表示衷心的感谢。丛书虽经反复修改、不断完善，但由于教育质量监测相关的测量理论、技术、方法日新月异，疏漏和错误在所难免，恳请广大专家和读者批评指正。

中国基础教育质量监测协同创新中心

2015 年 8 月

目　录

第一章　教育质量监测中的抽样设计

人类社会在发展过程中离不开对数据、资料、信息的统计调查和分析，如常见的人口普查、工业生产普查等。然而，政府统计中的普查和定期报表只适用于对基本国情(国势、国力等)等的调查，而对于大量的社会现象不可能做全面调查。一方面因为比较费时、费力、费财，我们并不需要“为了知道牛肉的滋味而吞噬掉整头牛”(Samuelson)；另一方面因为某些现象根本无法做合理的普查(因为被调查对象可能不完全清晰)。因此，发展非全面调查非常必要，通过局部数据推断全局特征，这正是推论统计的一条主线。在教育研究中，为了全面了解国家或者区域的教育现状经常采用抽样调查的方法，比如由经济合作与发展组织开展的国际学生评估项目(Programme for International Student Assessment，PISA)就是在参加的国家(地区)采用抽样的方式进行测试以了解学生不同方面能力素养的现状。美国教育进展评价(National Assessment of Educational Progress，NAEP)也是通过抽样的方式评价学生学业成就的项目。在教育质量监测中最常用到的抽样方式就是多阶段分层整群抽样。

本章的重点不是全面介绍各种抽样的原理和方法，而是结合教育质量监测的特点，重点介绍抽样设计研究中涉及的基本概念、常用抽样方法的应用以及抽样误差的估计和控制。

第一节　教育质量监测中抽样设计的基本概念

教育研究中的抽样设计往往适用于通过调查部分(而不是全部)个体组成的样本信息来对总体特征做出概括化的推断。由科学的抽样过程抽取到的样本得来的信息与全部个体的调查相比有一些优势，如节省经费、

节省人力和物力、减小调查实施过程的误差、高效等。

下面先简要介绍抽样调查涉及的一些基本概念。

一、总体、抽样框和样本代表性

(一)总体

总体是指研究对象的全体。在教育研究中，准确描述总体的组成元素非常重要，总体的定义即是研究关注的人群。为了对研究群体进行合适的描述，有必要按照总体的特征将其区分为期望的理想总体和实际研究的总体，即期望目标总体和抽样目标总体。最为理想的状况是，研究者完全控制了研究的环境，这两类总体包含的元素完全相同。然而实际情况中，两类总体存在一定的差异，例如：(1)不收敛。由于研究者不知道每一特殊群体的存在，在实际研究中忽略了总体中的某些元素。(2)资源的限制。研究者在其研究中有意识地排除了总体中的一些元素，因为这些元素在调查过程中会带来数据收集的困难或经费上的巨大投入等。

抽样目标总体的定义提供了抽样过程中一个可操作化的总体的定义，可以通过抽样框(样本将从这些元素中抽取)的定义来确定实际抽样的范围。

因此，对于总体而言，可以分为期望目标总体、抽样目标总体和被排除的总体(不包含在抽样目标总体中的元素称为被排除的总体)。实际抽样过程中需要对这三类总体进行明确的定义。

(二)抽样框

从抽样的目标总体中抽取样本需要首先建立抽样框。抽样框通常以总体元素列表的形式呈现，好的抽样框可以帮助研究者把握抽样的目标总体，而不用担心其中包含不合适的个体或者排除在外的个体对抽样过程造成污染。

一般来说，抽样框除了简单抽取元素的列表外，往往还包含很多其他的信息。如1991年关于30个国家阅读素养的调查(Ross，1991)，抽样框在罗列学校时根据的是一系列的分层变量：规模(学生人数)，地域(城市或农村)，学校性质，性别组成(男女是否混校)等。抽样框中包含的这些分层信息可以让研究者在数据分析中呈现不同层的数据结果，或

对不同层之间的差异进行分析。

(三)抽样代表性

“代表性”一词在抽样设计中很常用，但是在社会科学研究中常常被误用。如果样本数据特征的频率分布与对应的总体分布特征相似，那么称之为样本具有代表性。

要考察样本是否具有代表性并不容易，需要首先确定判断的标准，实际中往往通过比较抽取到的样本特征与总体特征是否一致来判断。被选择用于比较总体特征和样本特征的变量称为“标记变量”(Marker Variables)，这些变量通常是一些总体和样本中都包含的人口统计学变量。然而，实际中并没有客观的标准来判断哪个变量可以被称为标记变量。另外，也没有统一的用于判断样本频率分布与总体分布相似性程度的“基准点”，即没有统一的判断样本为代表性样本的基本标准。

在教育学研究中，最常用的标记变量有与学生相关的背景变量(如性别、年龄、社会经济地位等)，与学校相关的背景变量(如学校类型、学校位置和学校规模等)。例如，早期美国的许多抽样调查研究采用的标记变量主要有学生的性别、父母的受教育程度等。通过比较抽取到的样本中学生不同性别的比例与总体是否一致，样本中父母受教育程度的分布与总体是否一致来判断样本是否具有代表性。

二、概率抽样和非概率抽样

教育研究中的抽样往往有两个方面的目的：(1)通过样本所获得的统计量的信息(如样本平均数)来估计总体的参数(如总体平均数)。(2)通过样本信息来检验总体的某一假设，如通过样本得到的男生的学业成绩和女生的学业成绩，推断总体中男女生的学业成绩是否存在差异。要实现这两个目的，要求我们具备一些基本的推断统计的知识，如样本统计量的抽样分布和标准误。要考虑样本估计带来的误差，在抽样方式上往往需要采用概率抽样的方法。

概率抽样又称随机抽样，以概率理论为依据，通过随机化的操作程序取得样本，能避免抽样过程中人为因素的影响，保证样本的客观性。

虽然随机样本一般不会与总体完全一致，但它所依据的是大数定律，而且能计算和控制抽样误差，从而正确地说明样本的统计值在多大程度上适合于总体。另外，根据样本调查的结果可以从数量上推断总体，也可以在一定程度上说明总体的性质和特征。

与之相对的非概率抽样又称为不等概率抽样或非随机抽样，就是调查者根据自己的方便或主观判断抽取样本的方法。它不是严格按随机抽样原则来抽取样本，所以失去了大数定律的存在基础，也就无法确定抽样误差，无法正确地说明样本的统计值在多大程度上适合于总体。虽然根据样本调查的结果也可以在一定程度上说明总体的性质特征，但不能从数量上推断总体。由于非概率抽样的方法不能客观地处理总体参数估计或者假设检验的问题，所以如果研究的目的是要对总体参数做出推断，这一抽样方式就不合适了。

(一)非概率抽样的类型

1. 主观抽样

主观抽样假设研究者能从适当的目标总体中选择出有代表性的“典型样本”(Typical Sample)。基于这一假设选出来的样本依赖于研究者对典型样本组成主观解释的准确性。采用主观抽样很难获得有意义的结果，因为没有哪两个专家对典型样本构成的解释完全一致，因此，由于缺乏外部评判标准，很难判断基于某个典型样本得到的结果比另一个典型样本的结果更准确。

2. 方便抽样

方便抽样从字面上理解就是基于研究者易得到和方便的考虑，从目标总体中抽取样本的方法。方便样本又称为“偶然样本”(Accidental Samples)，由于总体中的元素是否被抽取通常取决于研究者收集数据时的时间、空间等因素，方便抽样的最基本假设是目标总体中的成员具有同质性。与主观抽样得到的样本相似，研究者无法对不同研究者获得的方便样本的精度进行比较。对这一方法最主要的批判是总体中方便得到的元素与不方便得到的元素之间往往存在很大的区别，因此基于方便样本所得到的结论往往是有偏的。

3. 定额抽样

定额抽样是最常见的非概率抽样的方法，由于从不同目标总体层中抽取的元素的数量与层的大小成比例，因此该方法有时会被误认为是“代表性抽样”。定额抽样对每一层中样本元素的抽取有很大的局限性，在每一层中元素的抽取几乎不加任何控制，例如在层中元素的抽取采取方便取样或主观抽样的方法。因此，这一方法得到的样本在每一层中都不可能具有代表性，即使每一层的样本量与层的大小成比例，也无法保证样本的总体代表性。

(二)概率抽样的类型

1. 简单随机抽样

简单随机抽样是随机抽样的一种，指直接从总体(而不是层之类的子总体)中抽取个体(而不是群之类的大单元)，而且抽取过程不带任何非随机性色彩。简单随机抽样是最基本的抽样方法，它是其他概率抽样的核心基础。另外，简单随机抽样相对比较容易实施。

一般地，设一个总体含有 N 个个体，从中逐个不放回地抽取 n 个个体作为样本($n \leqslant N$)，如果每次抽取时总体内的每个个体被抽到的机会都相等，就把这种抽样方法叫作简单随机抽样，这样抽取的样本叫作简单随机样本。

简单随机抽样首先将总体的 N 个单元从 $1 \sim N$ 编号，每个单元对应一个号码，如果抽到该号码，则对应的那个单元入样。要选出 n 个单元入样，方法很多，常用的方法有抽签法、随机数字表法等。

2. 分层随机抽样

分层随机抽样中分层的实质是根据辅助信息将总体划分为若干个子总体，或称为层，然后分别从每个层中抽取样本。分层之所以有好处，是由于各层中的样本量是由抽样者所控制，而不是由抽样过程随机决定的。当总体各单元差异比较大时，对参数估计误差比较大。将总体分层，使得同一层中各单位差异小，从每一层中抽取构成样本，这样样本就有了代表性，提高了估计的精度。另外，这种抽样方法还可以同时对子总体进行参数估计，便于依托各级管理机构进行组织和实施。

层的样本量常常被确定为与各层的规模成比例，换言之，即在各层中使用相同的抽样比。这种分配样本的方法被称为比例分配，这一情况下层内每个个体被抽到的概率是相同的，数据不用加权就能够得到总体参数的估计。不过总体样本在各层之间的分配并不一定限于按比例分配，也可以进行非比例分配。非比例分配常用于为了保证层内参数估计的精度，以比较层与层之间的差异。非比例抽样中层内个体被抽到的概率是不等的，因此需要对数据进行加权，这一加权因子主要是为了避免抽取概率的不同而导致的对总体参数估计的偏差。

3. 整群抽样

整群抽样又称为聚类抽样，是将总体中各单位归并成若干个互不交叉、互不重复的集合(称为群)，然后以群为抽样单位抽取样本的方法。应用整群抽样时，要求群具有较好的代表性，群内各单位的差异要小，而群体间个体的差异要大。采用整群抽样往往由于群与群之间的差异较大，因而抽样误差要大于简单随机抽样的误差。然而在实际应用中，这种方法具有实施便利、节约经费的优点。

对于整群抽样，实施过程中首先要确定分群的标准，将总体分成若干个群；其次根据实际情况和抽样精度的要求，确定抽取的群数；最后采用简单随机抽样的方法确定被抽到的群。

第二节　抽样的精度

在抽样过程中，往往需要对抽样的误差或者精度进行控制。本节主要介绍与抽样精度有关的一些概念，教育研究中常用的两阶段整群抽样方法以及抽样设计。

一、参数估计的精度

样本估计值与总体参数之间差异的大小可以用来评价概率抽样的精度。但是在实际中，由于总体参数往往是不知道的，因此对于一次抽样结果的精度没有办法通过描述估计值和真实值的差异计算得到。然而，

我们可以借助采用相同抽样设计的所有可能得到的样本估计值来评价样本估计的精度。

(一)均方误差

假设样本容量为 n 的概率样本，计算样本的均值来估计总体的均值。如果从总体中可以独立抽取无穷多组容量为 n 的样本，对于每一个样本计算其平均值，那么这些样本均值的平均值就是期望值。样本均值对总体均值估计的精度就可以通过均方误差来描述。对于样本均值，均方误差定义为：

$$MSE(\overline{X})=E(\overline{X}-\mu)^2=E[\overline{X}-E(\overline{X})]^2+[E(\overline{X})-\mu]^2$$
$$=\overline{X}\text{ 的方差}+(\text{偏差})^2$$

其中：$\overline{X}$ 为样本平均数；$E(\overline{X})$ 为样本平均数的期望值；μ 为总体均值。根据中心极限定理，样本均值的期望值与总体参数的偏差趋近于零，所以样本均值估计的精度可以用样本均值的方差来估计。

对于无限样本，一次随机抽样的结果，样本均值估计的精度为：

$$Var(\overline{X})=\frac{S^2}{n}$$

对于总样本量为 N 的情况，一次随机抽样的结果，样本均值估计的精度为：

$$Var(\overline{X})=\frac{N-n}{N}\times\frac{S^2}{n}$$

根据中心极限定理，大多数统计量在大样本的情况下都会近似服从正态分布。由正态分布的知识可以知道，总体参数 68%的置信区间为样本均值±1×均值的标准误，95%的置信区间为样本均值±2×均值的标准误。

(二)抽样精度的比较

概率样本精度的比较通常采用一定样本量下，样本估计量方差的大小比较来判断。Kish(1965)提出一个概率样本精度的高低可以通过与简单随机抽样设计比较来判断，即采用简单随机抽样设计作为其他一些复杂概率抽样精度的比较标准。Kish(1965)提出设计效率(Design Effect，deff)的概念来描述样本量相同时复杂抽样设计样本均值方差与简单随机抽样设计样本均值方差的比值，即：