冯　缨　樊茗玥 著

网络调查数据质量控制的方法与对策研究

上海三联书店

网络调查数据质量控制的方法与对策研究

冯 缨 樊茗玥 著

SJPC
上海三联书店

图书在版编目(CIP)数据

网络调查数据质量控制的方法与对策研究/冯缨，樊茗玥著. —上海：上海三联书店，2013.6
ISBN 978-7-5426-4170-0

Ⅰ.①网… Ⅱ.①冯…②樊… Ⅲ.①计算机网络—应用—统计数据—质量管理—研究 Ⅳ.①0212.1-39

中国版本图书馆 CIP 数据核字(2013)第 077898 号

网络调查数据质量控制的方法与对策研究

著　　者 / 冯　缨　樊茗玥

责任编辑 / 冯　征
装帧设计 / 鲁继德
监　　制 / 李　敏
责任校对 / 张大伟

出版发行 / 上海三联书店
(201199)中国上海市都市路 4855 号 2 座 10 楼
网　　址 / www.sjpc1932.com
邮购电话 / 021-24175971
印　　刷 / 上海叶大印务发展有限公司

版　　次 / 2013 年 6 月第 1 版
印　　次 / 2013 年 6 月第 1 次印刷
开　　本 / 890×1240　1/32
字　　数 / 180 千字
印　　张 / 7.75
书　　号 / ISBN 978-7-5426-4170-0/C·470
定　　价 / 25.00 元

敬启读者，如发现本书有印装质量问题。请与印刷厂联系　021-66019858

前 言

网络调查是现代网络技术和传统调查技术相结合的产物。随着互联网的飞速发展和网络普及程度的不断提高，网络调查实施得越来越广泛。与传统调查相比，网络调查在组织实施、信息采集、信息处理、调查效果等方面具有鲜明的优势，但也正是由于网络的特性使得网络调查存在独特的缺陷，如网络的覆盖率、网络的低控制性、网络的开放性以及网络的安全性等问题。这些问题成为提升网络调查数据质量的障碍。

本书从数据质量控制机理深入研究网络调查的数据质量，以全面质量管理理论、优化控制理论和数据误差理论为基础，界定网络调查数据质量等相关概念；提出了网络调查数据质量控制体系；构建了以网络调查数据的内生质量、传递质量及控制质量特征因素为二级指标的网络调查数据质量特征因素体系；对网络调查数据质量特征因素进行分析；分析表明，网络调查数据质量获得性和有效性是网络调查数据质量的主要特征因素。因此，针对提出的问题，本书先后引入网络调查数据的有效聚类、信心—信任容忍以及水印与群签名等信息技术，同时引入加权调整法、二级抽样法、热卡插补法以及随机化回答模型等统计技术，从控制网络调查数据质量的微观层面，对由网络特性导致的网络调查数据质量问题进行研究，从数据质量管控的方法层面寻找有效控制网络调查数据质量的方法。在理论分析的基础上，经过强假设，构建仿真样本

集，设计网络调查无回答误差修正与计量误差修正仿真流程，利用S-Plus和SPSS统计软件设计并实施仿真程序，验证各误差修正技术的可行性和有效性。

研究认为，推动调查组织的全面参与、提升网民的总体素质、设计科学的调查方案、增强网络调查的可信任程度、加强网络调查的过程监控、采用必要的数据修正技术、纳入丰富的先验辅助信息、明确适合的网络调查范围并以混合方式辅助实施调查、采用多学科交叉的技术与方法等是网络调查数据质量控制的有效途径。

与国内外相关研究相比，本书的创新点主要体现在以下四个方面：

第一，科学界定了网络调查数据质量以及网络调查数据质量控制的概念，提出了网络调查数据质量控制理论，构建了网络调查数据质量特征因素指标体系，拓展了网络调查数据质量的研究领域。

第二，系统研究了提升网络调查数据质量的信息技术。从网页式调查数据获取角度分别引入了数据聚类与错误容忍信息技术，以期在提升网络调查数据质量过程中体现有效监控与筛选作用；在电子邮件式网络调查数据的获取中，引入了数字水印技术与群签名技术，能够有效地避免在电子邮件式网络调查中形成的复制攻击，以筛除复制数据，提升网络调查数据质量。

第三，系统研究了提升网络调查数据质量的统计技术。在分析网络调查数据质量误差原因的基础上，引入数据误差领域的研究成果，并适当改进之以符合网络调查特征，分别对网络调查的覆盖误差、抽样误差、无回答误差和计量误差进行修正，从技术层面控制网络调查工作中误差因素带来的数据质量问题，其中包括参与调查的“人”的问题，调查本身的设计问题以及数据搜集等问题。并提出可操作的数据质量保障方法。

第四，有效验证了网络调查误差修正的可行性。在合适的网络调查数据误差仿真样本集的基础上，设计网络调查无回答误差

修正与计量问题修正仿真步骤，设计并实施仿真程序，从误差角度实现控制网络调查数据质量的技术，开拓了现有理论和方法的应用领域，提高了研究的深度和精度。

本书以国家社会科学基金项目“提高网络调查数据质量的方法与对策研究(09CTJ006)”的研究成果为基础撰写而成。该项目开始着手研究以来，课题组进行了大量的文献检索和资料积累，在此基础上，进行了系统的分析研究，明确研究目标形成研究大纲。2012 年 1 月，经过多次修改、讨论，完成了总体研究报告，并提请项目验收。2012 年 9 月获得全国哲学社会科学规划办公室颁发的结项证书。本课题的研究成果受到了国家社科基金鉴定专家的精心评审，他们的反馈意见给予了作者不断进行后续研究的动力，也为本书的完成奠定了良好的基础。

本书的完成，得到了诸多领导、专家的支持和鼓励，也凝聚了学科团队的集体智慧，这包括：江苏大学应用统计研究所的赵喜仓教授、宗明刚副教授、吴继英讲师、夏纯中讲师等为本书提供了基础资料；江苏大学可信系统研究所的王良民教授、熊书明副教授、王新胜副教授作为社科基金项目的合作者，在信息技术方面给予了我们最大的帮助，并完成了本书第四章的撰写。感谢国家哲学社会科学规划办公室和匿名评审专家，感谢本书参考文献的作者们，感谢本书责任编辑对本书的出版给予的精心指导和帮助，感谢家人长期以来给予的支持和理解。

网络调查数据质量控制理论、方法及其对策研究是集管理学、统计学、信息学等多学科领域的交叉性研究课题，本书的探讨是初步的和探索性的，一定存在不少疏漏和不足，恳请领导、专家和读者批评指正。

著 者

2012 年 9 月于镇江

目 录

第一章
绪　论

1.1　本书研究背景及目的

1.1.1　研究背景

在众多的管理科学研究方法中，调查几乎是一种运用最多的从某一主题收集数据的方法。网络调查是一种新兴的调查产业，随着人类进入信息时代，互联网的飞速发展，网络普及程度的不断提高，网络调查逐渐繁荣和广泛。

网络调查又称在线调查、联机调查，是指在各种计算机上通过互联网以电子邮件或其他形式把传统的调查与分析方法在线化和智能化。它能通过网络，大量、同时且直接地把问卷送到受访者的个人电脑，受访者也可以从网页或电子信箱取阅问卷，接受资料、填写回复与传递资料全部都在网络上进行。它在网络上实施数据采集、传输、上报、交换等业务，对数据进行自动处理和汇总。这是一种针对具有高度信息收集能力的网络用户群体而产生的调查方式，是现代网络技术和传统调查技术相结合的产物。

在传统调查中，设计调查、实施调查以及整理调查是统计数据质量控制的三个阶段，同时也是社会调查不可缺少的环节。然而，纵使调查设计各步骤执行得精准与清晰，却始终绕不开实施调查

过程中与过程后的数据质量问题，如数据收集成本高、数据反馈不及时、数据获得有偏差、调查组织庞大不易管理、调查过程监测不到位、分析结果误差较大等。但是，网络调查相对传统调查在调查过程中具有不可比拟的优势：它不受时空限制，具有开放性、自由性、平等性、广泛性和直接性，同时具有调查组织简单、便捷，调查费用低等特点。从统计调查数据质量的管理角度看，网络调查在组织、管理、方法和技术方法方面较传统调查改进很多。

但是，由于网络自身的开放性、不安全性以及局域性等特点，在数据误差的预防、检验、控制方面带来了很多新问题，如难以回答具有普适代表性的问题、难以预测与消化过量的数据、难以辨识数据的真伪、难以监测调查过程、难以保证数据安全、难以处理不一致的数据形式等。这些问题均使得调查者无法保证网络调查的数据质量。

胡帆(2011)认为，从组织、管理、方法、技术等方面，对数据质量进行预防、检验、控制和校正，全面提升统计数据质量，才能保证统计数据达到应有的质量标准。显然，在必须面对的一个新调查方式的研究领域中，无论网络调查方法未来的走向以及对社会调查的贡献如何，我们都有必要对这一新生且影响不断扩大的调查方法从提升数据质量的角度对调查过程中与过程后产生的数据质量问题进行研究，从而更好地发挥统计调查的真正作用。

1.1.2 研究目的

本书旨在从提升数据质量的多种角度深入研究网络调查的数据质量，以全面质量管理理论、优化控制理论和数据误差理论为着落点，通过分析网络调查数据质量的主要特征因素，基于网络调查数据质量控制体系，分别对网络调查的事前监控与事后修正过程中产生的数据获取问题和数据误差问题进行全面剖析，并试图因此寻找有效提升数据质量的方法。值得强调的是，在本书的技术

分析和仿真分析中，笔者将引入国际上先进的模型和技术，在加以改进后，对控制网络调查数据质量的微观层面进行深入研究，力争在数据质量管控的方法层面有所突破，这对改进当前市场调查工作体系、提高网络统计数据质量、提供可信的数据分析结果及信息服务都具有重要的理论及实践意义。

1.2 国内外研究现状

1.2.1 网络调查的研究现状

网络调查作为一种新兴的统计调查方法，在一些特定的领域已经得到了应用。最早的网络调查可以追溯到 1994 年——佐治亚理工学院的 GVU Center 进行的关于互联网的使用情况、用户情况及人口统计状况的调查。但是在 1995 年之前，国外使用网络问卷调查法的机构和研究并不多见。但到了 20 世纪 90 年代中期之后，使用网络问卷调查法的专业调查机构开始变得越来越多。1997 年，“欧洲民意和市场调查协会”(ESOMR)关于市场调查行业短期发展趋势的调查结果显示，在未来 5 年内，影响市场调查行业的 6 个关键因素之中，首要一条就是“对调查技术的需要将更迫切”，这些调查技术包括：互联网(在线)调查、自动数据收集、数据库管理、市场建模、创造性(交互式)的广告测试等。在 1998 年，英国“全国统计局政府办公室社会调查处”(ONS)所实施的一项针对英国 202 所商业调查机构的调查结果显示，自 20 世纪 80 年代以来，英国专业调查机构使用各种基于计算机的调查技术的比例呈逐年上升趋势，尤其是 1992 年之后，网络调查法(英国称之为 Computer-assisted Web Interviewing，CAWI)开始被应用于数据收集。至 1996 年，已经有 19.2%的调查机构开始使用网络调查，其中比较有名的是美国的 InterSurvey 公司[1]，利用该公司创建的

互联网络固定样本，可提供网络市场调查服务（Robert，2006）。在美国，2003 年，Pioneer 市场研究中心发表的调查数据显示，在美国的专业调查机构中，利用互联网来收集调查数据的比例在逐年上升。调查结果显示，约四分之三（72.5%）的受访者表示，其所在的调查机构正在使用互联网来收集数据。其中 39.2%的受访对象表示，网络调查法目前是所在调查机构唯一的数据收集方式；33.3%的受访对象表示，除了网络调查法以外，他们还同时使用其他调查方法，如计算机辅助电话调查（CATI）等。研究资料还表明，在问到各种调查研究方法的未来发展趋势时，38.4%的受访者认为，CATI 将是未来数年中美国调查机构在收集数据时所使用的主要方法；25.5%的受访对象认为，各种基于互联网的调查方式（如网络调查法和网络访谈法）将是调查研究机构未来数年中最主要的数据收集方式。另外，认为混合调查法、印刷问卷调查法和交互语音调查法将是未来数年中调查机构所使用的主要数据收集方法的比例分别是：9.1%、7.0%和 4.5%。在网络调查领域颇有成果的美国学者 Mick P. Couper（密西根大学社会研究所的教授，美国 GPS（General Population Survey）的调查方法专家组成员），自 20 世纪 90 年代起，开始关注电子问卷的设计与应用，并于 1997 年主编出版了第一本有关电子问卷设计与应用的著作《Computer Assisted Survey Information Collection》，同时，还发表了大量的有关电子问卷设计的论文。从文献资料来看，国外在网络调查领域所开展的研究中，实证性的研究占了很大的比重，且多为调查性研究。应用统计方法对所得到的数据进行定量分析，从数据质量角度来定量地比较传统问卷调查和网络调查的差异，成为目前国外研究的一个热点。

在国内，1999 年 10 月 16 日，北京零点专业市场调查公司与爱特信搜狐网络公司正式携手，创立了搜狐—零点网上调查公司，共同拓展网上调查业务。这标志着中国调查业步入“网络时代”。近年来随着互联网在我国应用的迅速发展，网络调查也逐步地开

始发展起来。

在网络调查的实际应用方面，以中智库玛(www. comr. com. cn)为代表的我国网络调查应用在很多领域已经取得了实质性的进展，主要成功的案例有 TCL 集团员工满意度调查等。专业的网络调查主要以北京大学教育技术系网络调查研究中心与唯思瑞(Wise Real)公司合作开发的国内第一个专业网络问卷调查系统(www. websurvey. cn)为代表，该系统主要面向教育领域，已经成功实施了“北京大学本科教学总体状况调查”、“2005 年中国高校信息化调查”等，从调查过程与结果来看，效果显著。此外，研究领域的张凯昀等(2006)设计的一种新型的网络调查问卷生成平台利用了信息技术中的本体的思想[2]；邢苗条等(2005)采用了大量的信息技术实现了一种基于 Web 的网络调查统计信息系统的系统框架[3]；张涛(2009)利用 ASP 技术建立了网络调查投票系统[4]；张清等(2009)利用网络调查来分析旅游电子商务的满意度[5]；姜晓洁(2009)通过分析 JSP 技术的特点并结合 JavaBeans 和 JDBC 动态访问 Web 数据库的方法、模型和关键技术，提出了一个基于 JSP 技术的通用网络调查系统的设计方案，提高网络调查系统的性能[6]；张玲等(2009)利用网络调查法，调研了我国 32 所高校图书馆开展信息素质教育的现状、信息素质教育在各馆主页中的组织及呈现方式[7]；朱庆华等(2009)基于网络调查方法获得的数据，较为全面地分析了我国信息通信产业的现状，评价了我国信息通信政策的实施效果[8]；朱明芳(2005)研究了电子邮件形式的网络调查在旅游企业中的应用[9]；刘晨(2004)利用网络调查的方法，对中美高校图书馆网络信息服务的现状进行比较分析，以寻求适合我国高校图书馆网络信息服务发展的有效途径，提升网络信息服务品质等[10]。

但是根据方佳明(2006)、李锐(2005)等对我国网络调查研究的论文进行整理和分析看出，自 1994 年以来，我国学者对网络调查的研究主要集中在以下几个方面：(1)网络调查的优势和缺陷，

(2)网络调查的方式(方法),(3)网络调查的改善和提高,(4)网络调查的行业应用。从研究本身来看,一方面,我国对网络调查的研究还不够热烈,并且研究者比较分散,还没有形成一个持久且稳定的研究队伍;另一方面,研究的深度不够,几乎全部是描述性研究,是对已有的观点和看法的综合和整理,很少有实践数据来支撑[11][12]。曾五一等(2002)认为这些网络调查的应用才刚刚开始起步,都仅仅局限于某些特定的领域,网络调查的许多重要的基本问题也没有获得彻底解决,要普及网络调查,充分发挥网络调查的优点,尚需要进一步的努力[13]。我国学者立足我国国情,对网络调查的特点尤其是其局限进行了深入的研究。徐浪等(2006)研究了网络调查的属性,认为网络调查具有典型的统计调查的特点,但是网络调查以网络为媒介的"人—机"交互模式取代了传统调查以问卷、电话为中介的"人—人"交互模式,从而形成了网络调查方法的特殊性,一方面具有时效强,费用低廉,调查区域广,调查形式多样,便于管理等独特优势,使得收集大容量的数据变得更为方便;另一方面,网络调查本身还存在如样本代表性小、答卷信度低等不足,而且传统的统计调查抽样方法及调查技巧无法解决此类问题[14]。曾五一(2007)和曾鸿(2004)分别就网络安全和个人隐私保护问题进行了分析,指出信息技术本身的缺陷也带来了网络调查研究的新问题[15][16]。

1.2.2 网络调查数据质量的研究现状

数据质量问题包括数据的准确性和可靠性,是网络调查讨论的重点,是网络调查的生命。但由于网络调查是新兴研究领域,国内外学者对网络调查数据质量的研究无论从定量还是定性角度却不多。耿修林等(2002)指出在我国社会经济统计方面存在数据质量不高,降低了统计信息使用价值的问题,其主要的特征因素有报喜不报忧、虚报瞒报、人为干扰;工作疏忽、组织措施不得力导致的

漏报和不报；统计数据的管理重视不够，资料的积累和保存问题；数出多门造成的统计指标差别等。虽然网络直报系统在一定程度上可以克服上述部分问题，但是对于蓄意的人为干扰，网络调查更加无能为力[17]。曾五一(2007)指出在网络调查中甚至存在利益相关人员企图利用网络来干扰网络调查数据、影响调查数据质量的问题。依据这些数据质量不高的调查数据获得的统计结论，常常会发生严重的偏差[15]。杜婷等(2004)开始关注网络调查数据的数据质量，指出了一些解决问题的网络调查技巧，后又对非抽样误差进行了研究，这对实施网络调查过程中如何关注数据质量问题具有重要意义[18]。蒲国华等(2003)在对未来网络调查的设想中明确指出了网络调查中数据误差问题、数据安全性问题、数据规范性问题和信息服务问题是最可能导致数据质量低下的根本问题[19]。方佳明等(2006)针对国内外网络调查失败率高的问题，提出了网络调查适用性的20个因素，最终分析得出在影响网络调查适用性上起主要作用的9个因素，分别为问卷调查的规模、有效电子邮件地址的可得性、被调查者的属性、可得到的资金数量、调查时间的紧迫性、被调查者的地理分布、被调查者对调查项目内容的兴趣度、被调查者对调查主体的认知程度、对被调查者的定向性要求[20]。刘全等(2007)认为，网络调查数据质量的特征因素既有调查方案设计和技术支持的因素，也有具体实施中出现的各种情况，并设计了以样本、问卷、填报、整理、分析、方案为一级指标的网络调查数据质量特征因素[21]。国外关于网络调查数据质量的研究开展得比较早，Don(1978，2007)最初讨论了计算机辅助调查的一些数据质量问题，随后的专著中系统阐述了网络调查数据质量问题[22][23]；Gunar(1990)编著的《Data Quality Control — Theory and Pragmatics》在第11章由Johnny blair撰文讨论了如何提高样本稀少时网络调查的数据质量[24]；Hanscom(2002)给出了网络调查与传统调查在回答率上等数据质量指标上的定量比较[25]。