中/青/文/库

大数据流下调查数据的统计分析

张喆 著

中国社会科学出版社

中/青/文/库

大数据流下调查数据的统计分析

张喆 著

中国社会科学出版社

图书在版编目（CIP）数据

大数据流下调查数据的统计分析／张喆著．—北京：中国社会科学出版社，2018.8

ISBN 978-7-5203-2809-8

Ⅰ.①大… Ⅱ.①张… Ⅲ.①数据处理—统计分析 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2018）第 161058 号

出 版 人 赵剑英
责任编辑 张 湉
责任校对 姜志菊
责任印制 李寡寡

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京明恒达印务有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2018 年 8 月第 1 版
印 次 2018 年 8 月第 1 次印刷

开 本 710×1000 1/16
印 张 7.5
字 数 128 千字
定 价 38.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

本书得到中国青年政治学院出版基金资助

《中青文库》编辑说明

《中青文库》，是由中国青年政治学院着力打造的学术著作出版品牌。

中国青年政治学院的前身是1948年9月成立的中国共产主义青年团中央团校（简称中央团校）。为加速团干部队伍革命化、年轻化、知识化、专业化建设，提高青少年工作水平，为党培养更多的后备干部和思想政治工作专门人才，在党中央的关怀和支持下，1985年9月，国家批准成立中国青年政治学院，同时继续保留中央团校的校名，承担普通高等教育与共青团干部教育培训的双重职能。学校自成立以来，坚持“实事求是，朝气蓬勃”的优良传统和作风，坚持“质量立校、特色兴校”的办学思想，不断开拓创新，教育质量和办学水平不断提高，为国家经济、社会发展和共青团事业培养了大批高素质人才。目前，学校是由教育部和共青团中央共建的高等学校，也是共青团中央直属的唯一一所普通高等学校。学校还是教育部批准的国家大学生文化素质教育基地、全国高校创业教育实践基地，是首批“青年马克思主义者培养工程”全国研究培训基地、首批全国注册志愿者培训示范基地，是中华全国青年联合会和国际劳工组织命名的大学生KAB创业教育基地，是民政部批准的首批社会工作人才培训基地，与中央编译局共建青年政治人才培养研究基地，与国家图书馆共建国家图书馆团中央分馆，与北京市共建社会工作人才发展研究院和青少年生命教育基地。2006年接受教育部本科教学工作水平评估，评估结论为“优秀”。2012年获批为首批卓越法律人才教育培养基地。2015年中宣部批准的共青团中央中国特色社会主义理论体系研究中心落户学校。学校已建立起包括本科教育、研究生教育、留学生教育、继续教育和团干部培训等在内的多形式、多

层次的教育格局。设有中国马克思主义学院、青少年工作系、社会工作学院、法学院、经济管理学院、新闻传播学院、公共管理系、中国语言文学系、外国语言文学系等9个教学院系，文化基础部、外语教学研究中心、计算机教学与应用中心、体育教学中心等4个教学中心（部），中央团校教育培训学院、继续教育学院、国际教育交流学院等3个教育培训机构。

学校现有专业以人文社会科学为主，涵盖哲学、经济学、法学、文学、管理学、教育学6个学科门类，拥有哲学、应用经济学、法学、社会学、马克思主义理论、新闻传播学等6个一级学科硕士授权点、1个二级学科授权点和3个类别的专业型硕士授权点。设有马克思主义哲学、马克思主义基本原理、外国哲学、思想政治教育、青年与国际政治、少年儿童与思想意识教育、刑法学、经济法学、诉讼法学、民商法学、国际法学、社会学、世界经济、金融学、数量经济学、新闻学、传播学、文化哲学、社会管理等19个学术型硕士学位专业，法律（法学）、法律（非法学）、教育管理、学科教学（思政）、社会工作等5个专业型硕士学位专业。设有思想政治教育、法学、社会工作、劳动与社会保障、社会学、经济学、财务管理、国际经济与贸易、新闻学、广播电视学、政治学与行政学、行政管理、汉语言文学和英语等14个学士学位专业，，其中思想政治教育、法学、社会工作、政治学与行政学为教育部特色专业；同时设有中国马克思主义研究中心、青少年研究院、共青团工作理论研究院、新农村发展研究院、中国志愿服务信息资料研究中心、青少年研究信息资料中心等科研机构。

在学校的跨越式发展中，科研工作一直作为体现学校质量和特色的重要内容而被予以高度重视。2002年，学校制定了教师学术著作出版基金资助条例，旨在鼓励教师的个性化研究与著述，更期之以兼具人文精神与思想智慧的精品的涌现。出版基金创设之初，有学术丛书和学术译丛两个系列，意在开掘本校资源与迻译域外菁华。随着年轻教师的增加和学校科研支持力度的加大，2007年又增设了博士论文文库系列，用以鼓励新人，成就学术。三个系列共同构成了对教师学术研究成果的多层次支持体系。

十几年来，学校共资助教师出版学术著作百余部，内容涉及哲学、

政治学、法学、社会学、经济学、文学艺术、历史学、管理学、新闻与传播等学科。学校资助出版的初具规模，激励了教师的科研热情，活跃了校内的学术气氛，也获得了很好的社会影响。在特色化办学愈益成为当下各高校发展之路的共识中，2010 年，校学术委员会将遴选出的一批学术著作，辑为《中青文库》，予以资助出版。《中青文库》第一批（15 本）、第二批（6 本）、第三批（6 本）、第四批（10 本）、第五批（13 本）、第六批（9 本）陆续出版后，有效展示了学校的科研水平和实力，在学术界和社会上产生了很好的反响。本辑作为第七批共推出 5 本著作，并希冀通过这项工作的陆续展开而更加突出学校特色，形成自身的学术风格与学术品牌。

在《中青文库》的编辑、审校过程中，中国社会科学出版社的编辑人员认真负责，用力颇勤，在此一并予以感谢！

目　录

图 目 录

表目录

第1章　绪论

1.1　选题的背景和意义

1.1.1　选题的背景

有人会问，当今社会的热点话题是什么？“大数据”这个词语肯定首当其冲。可以通过一组数据来看出我们所处时代的特性，国际数据公司（IDC）给出了一系列的描述：2008年全球产生数据的总量约为0.49ZB，2009年的数据总量约为0.8ZB，2010年的数据总量增长为1.2ZB，2011年的数量总量高达1.82ZB，相当于全球每人产生200GB以上的数据量。而到2012年为止，人类生产的所有印刷材料的数据总量约是200PB，人类历史上所有语言资料总和的数据量也大约只有5EB。可见面对以ZB计量的数据量冲击下，我们认识和分析问题的方法都应有一定的改变。

根据数据量级的发展，可以将数据的历史大致分为以下几个阶段.

（1）1970年—1980年：数据级由MegaByte（MB）增长为GigaByte（GB）阶段。

在这一阶段产生的商业数据量从MB逐步达到了GB的量级，此时数据的存储以及关系型数据的查询遇到了瓶颈，可以将其看作是“大数据”萌芽阶段带来的挑战。此时数据库计算机技术随之产生，它通过集成了硬件和软件的功能来解决问题，以付出较小的代价来获得处理性能的提升。

（2）1980年—1990年：数据级由GigaByte（GB）增长为TeraByte（TB）阶段。

此阶段数字技术的提升促使数据容量从GB增长到TB的级别，因此单个计算机系统的存储和处理能力已将无法满足数据的要求。在这种

情况下学界提出了数据并行化计算技术，其基本思想是分配数据和相关任务到独立的硬件平台上运行处理，从而增强计算机存储能力并提高处理速度。

（3）1990 年—2000 年：数据级由 TeraByte（TB）增长为 PetaByte（PB）阶段。

随着互联网的普及和发展，这一阶段可以看作是互联网时代的产物，其显著特点是处理的数据量巨大，能够达到 Petabyte 级别，同时出现了半结构化和无结构的数据类型。尽管并行数据处理方式能够胜任结构化数据处理问题，但在无结构化数据面前却起不到任何作用，为此 Google 提出了 GFS 文件系统和 MapReduce 模块处理系统来应对大数据带来的挑战。

（4）2000 年—至今：数据级由 PetaByte（PB）增长到 ExaByte（EB）和 ZetaByte（ZB）阶段。

根据目前数据量的激增趋势，互联网和物联网等公司存储和分析的数据量级已经从 PB 发展到 EB 甚至是 ZB。可见大数据的浪潮已经来临，此时各学科和各领域都在投入大量人力、财力和物力来应对大数据带来的挑战。

从大数据的发展历史可以总结概括出其所具有的特性。有关大数据的特性，最为经典的论述就是 4V 特性，它们分别为多样性（Variety）、大量性（Volume）、高速型（Velocity）和价值型（Value）。

多样性指的是数据类型比以往更多，不单单仅有数字这一种表现形式，还包含图片、音频、视频、地理信息等，复杂的数据类型给传统的数据分析带来了挑战。

大量性通俗易懂，就是数据量非常的庞大，以前数据级一般都是 TB，而现在常见的数据信息都可以到达 PB、EB 甚至是 ZB。

高速性是指大数据时代，数据的更新速度非常快，同时需要处理速度、分析速度也要随之提升。

价值性是指数据价值密度较以前相比会大幅度下降，由于数据量的增大，单位数据所包含的有用信息越来越少，如何从大量数据中准确地提取有用信息，是大数据时代面临的挑战。

由于大数据自身特性使得大数据影响的领域很广很深，在 Web of Science 数据库中，以 big data 为主题词或者标题进行检索，黄永勤

(2014) 最后获得了673个关键词，并绘制了关键词时序分布图，通过分析得出2000年开始就出现了与大数据相关的研究；2000—2005年间关键词相对较少，与大数据相关的领域主要涉及遗传算法、神经网络算法、数据挖掘和信息分类等热点关键词，可见此时更多的研究重点在计算机领域和算法领域，主要兴趣点是研究数据增大带来的问题和调整方法；2007年云计算成为当时的核心词汇，它的出现大大提升了数据存储和传输性能；2008年Nature杂志刊发Big Data专刊，使得大数据开始受到业界和学术界的广泛关注，最早的研究还主要是以宏观描述为主；之后随着计算机和统计科学水平的提升，大数据的研究更加实用，更加注重现实影响；2010—2012年MapReduce和Hadoop两种数据处理模式一直都是研究热点，现在对这一领域的研究更多关注的是如何改进和实际案例应用；2012—2013年随着数据流的激增，在非结构化数据的处理工程中出现了像语义网、可视化等方面的研究。

正如哈佛大学社会学教授加里所说："大数据是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都开始这种进程。"目前大数据这个词汇经常出现在各大券商的投资报告中，大多数物联网商家组建大数据研究平台，连各国政府网站也经常会出现有关大数据的新闻。

各国政府斥巨资研究大数据相关课题，为政府决策提供准确的数据支持。在美国，2012年奥巴马政府宣布投资2亿美元推出"大数据研究和发展计划"，目的是搜集当前海量的数据和提高分析信息的能力，这一举动可以看作是将大数据研究从民间商业行为发展到了政府发展战略层次的分水岭。可以看出对大数据的掌控能力已经成为一个政府执政能力，国际竞争力是否强大的标志之一。

从2013年11月开始，中国国家统计局与中国联通、阿里巴巴、百度和58同城等11家企业签订了战略合作协议，以期在大数据应用上大展拳脚，马建堂局长还在2014年第六届中国人民大学国际统计论坛致辞中指出"以更加开放姿态推动大数据共享共赢，大数据应用要共享开放，统一标准，市场推动。"

大数据的热潮还不止如此，国内外还举办了多场高水平的大数据学术会议其中科技部召开了两次香山会议，2013年国家自然科学基金委组织了双清论坛，并且设立了教育部重大项目。另外中国统计学会于

2013 年召开了第十七次统计科学讨论会来共议大数据背景下的统计，此次会议国内外学者都踊跃投稿，共商我国大数据未来发展方向。2014 年 11 月首届“大数据与应用统计国际会议”在中国人民大学成功举办，来自国内外的专家学者就当前大数据的前沿话题及最新研究成果展开了讨论。本次会议讨论的领域涉及基因学、天文学、宇宙学、经济金融和流行病学等，对于了解大数据统计分析理论、方法及应用研究的国际前沿和热点问题有着十分重要的意义。

大数据是一把双刃剑，给各个学科的发展既带来了机遇同时也带来了挑战。而本文正是基于大数据的背景，试图从统计学角度分析大数据带来的机遇和挑战。机遇显而易见，数据量的增加，使得我们分析时候选用的分析样本更多样和充足，而且随着数据流源源不断的产生，我们会拥有一个动态的训练集，这样通过随时更新的数据来调整模型参数可以得到时效性强、准确性高的估计结果。

但是我们不得不认清一个事实，由于数据本身的复杂性，统计分析变成了一项系统性工程，截至目前国内外对大数据的分析都处于起步探索阶段，相当部分的研究都是在宏观层面进行分析，而对大数据带来的影响，量化影响效果和修正模型估计等问题还没有体系化，本文从这一点入手，试图建立适合大数据背景下统计模型并对其影响效果进行分析。

1.1.2 选题的意义

从大数据的背景可以看出我们每天遇到的数据量相当惊人，这些数据像河流一样喷涌向前，因此我们可以将大数据形象的描述成数据流，这在互联网、物联网和社交平台等媒介上显而易见。

举两个例子来说明大数据正融入我们的生活：

第一个是柴静自费拍摄的专题片《穹顶之下》，在优酷视频平台上点击率已经超过 2 亿，可见其受欢迎程度。她用近一年的时间搜集与雾霾有关的数据，包括时间上和地区上的维度，深度解析了雾霾的生成原因和解决之道。片中争论的焦点在于大量燃烧煤炭和石油导致雾霾日益严重进而引起环境污染，最终增加了人们患癌症可能性的逻辑关系是否成立，柴静通过大数据分析验证了其正确性。与此同时有网友指出柴静在大数据分析中的一些问题，例如霾与 PM2.5 的数据不能等同，利用

采样仪和采样膜搜集数据会产生问题等，并得出了雾霾不是癌症的罪魁祸首的结论，同时煤炭和石油厂商也通过数据和案例验证了燃烧煤炭和使用石油不会引起雾霾的恶化，而且网上还有报道称炒菜做饭是引起雾霾的主要原因。

可见利用来源不同的数据，通过大数据分析竟然得到了不同的结论。但真相只有一个，我们需要利用科学客观的方法来分析大数据，而不能被大数据摆布，迷失在大数据之中。这就需要我们找到大数据对分析结果的影响机制，同时完善大数据分析体系从而得到准确的结果，给人们以正确的数据支持，这也正是本文的意义所在。

第二个例子是娱乐时代的电影宣传，传统的电影宣传一般是采用开发布会、宣传片等形式。但是在新时代下，电影的宣传也越来越有大数据的特色，最新的电影《港囧》就采用了类似于演讲的形式，利用票房、观影笑声数、泪点数等数据形象地展现了这部电影的魅力。

可见在大数据时代，我们身边事物的外在表现形式都在悄然发生着变化，我们应该认识到大数据是时代的产物，在分析影响机制时应尽可能量化大数据自身特征给传统统计方法带来的冲击，以便为今后研究指明方向。根据大数据自身的特点以及统计学本身的研究方法，将大数据时代遇到的挑战归纳如下：

第一，复杂数据结构的挑战。

大数据不仅包含结构数据，还包括半结构、甚至非结构化数据，例如声音、图片和视频等资料。如何对它们进行分析给统计工作者带来了困难；数据自身的维度不再是简单的低维数据，而更多体现出高维或者是超高维数据形态，这给传统的统计分析带来不便。

第二，“垃圾信息”的处理。

大数据包含着大量的虚假信息，有些是随机产生而有些是非随机的，统计分析的目的在于利用样本信息对总体进行推断。而此时我们观测到的信息是不真实的，如何过滤“垃圾信息”而使用真实信息分析，也给统计学家们提出了挑战。

第三，抽样方式的转变。

尽管数据量越来越大，但是抽样调查在大数据背景下还是必不可少的。传统的抽样方法在大数据背景下有时显的“力不从心”，因此构造出新的抽样方式和调查数据分析方法给抽样领域的专家带来了挑战。