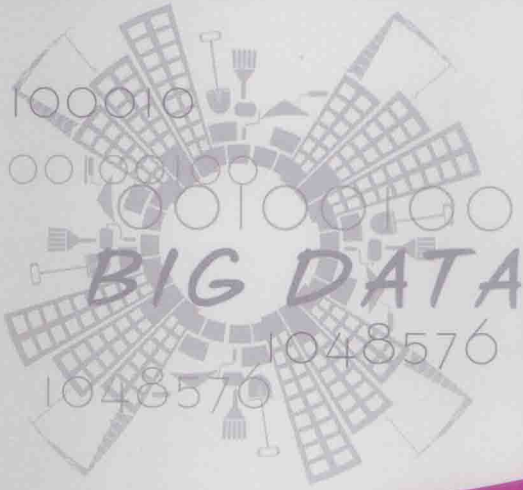
统计科普丛书

WUCHUBUZAI DE TONGJI

无处不在的统计（四）



中国统计学会 编

中国统计出版社
China Statistics Press

WUCHUBUZAI DE TONGJI

无处不在的统计（四）

中国统计学会 编

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

无处不在的统计. 4 / 中国统计学会编. — 北京 : 中国统计出版社, 2014.9

(统计科普丛书)

ISBN 978-7-5037-7201-6

Ⅰ. ①无… Ⅱ. ①中… Ⅲ. ①统计学－普及读物 Ⅳ. ①C8－49

中国版本图书馆 CIP 数据核字(2014)第 182612 号

无处不在的统计(四)

作　　者/中国统计学会
责任编辑/陈悟朝　徐　颖
封面设计/杨　超　李雪燕
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/100 千字
印　　张/9.5
版　　别/2014 年 9 月第 1 版
版　　次/2014 年 9 月第 1 次印刷
定　　价/28.00 元

《无处不在的统计(四)》
编委会

编者的话

2011年统计科普读物《无处不在的统计》问世，接下来连续两年相继出版《无处不在的统计（二）》和《无处不在的统计（三）》。该套丛书受到各方读者出乎意料的喜爱。现在这本《无处不在的统计（四）》将在2014年9月中国统计开放日如期呈现在读者面前。

本书是“统计科普丛书”的第四本，通过阅读，您会感觉与统计更贴近了。一是与统计工作重点更贴近。例如大家耳熟能详的大数据，通过书中作者的解读，可以对大数据在当前政府统计中的应用状况、为何大数据让统计“火”了起来有一个更清晰的了解。二是与统计实际工作更贴近。例如大家都在关注小康问题，何为小康？小康指数如何得来？我们距全面建成小康社会有多远？通过此书，您一定可以找到答案；再例如，2014年城镇住户调查和农村住户调查合并为城乡和住户调查，为什么合并，合并后有哪些内容？都可以在书中找到答案。三是与百姓生活更贴近。本书涵盖了大量统计学在实际生活中应用的例子，例如许多城市机动车限行，作

者运用频数统计法告诉您哪个尾号的机动车相对少堵车；再例如，用统计的视角解读降水概率，希望能为大家读懂天气做点什么。这些文章会让读者在阅读中品味到统计的魅力，感受到统计无处不在。

撰写科普文章，实际上是一件相当难的创作过程，既要让外行人看得懂、又要让内行人不笑话，把深奥而复杂的统计知识化为通俗易懂的科普文章，是作者专业和语言功力的体现。感谢四年来为我们撰稿的作者，有你们，我们这套科普丛书才能有一、二、三、四；有你们，读者才能够从中品位统计之魅力。

在本书编辑过程中我们尽力做到认真、严谨，但由于时间仓促、水平有限，难免存在疏漏和不足之处，欢迎广大专家和读者对本书的不足之处提出批评和建议。

目录

走近大数据

潘　璠/文

大数据作为一个词语和现象，在今天已经耳熟能详、脍炙人口。但是，倒退到两年前，我们对此又有多少关注和了解呢？数据无处不在，大数据的影响则正在与日俱增，走近大数据、认识大数据、应用大数据，对我们把握好这个世界的特点和规律、并科学地决策和抉择，具有重要的现实意义。

由两幅曲线图说起

曾经有两幅曲线图形象地反映了大数据受关注的程度，即谷歌和百度以“大数据”为关键词的搜索曲线图。前者在2011年到2012年的时段，后者在2012年到2013年的时段，都有一个明显、甚至可以说是陡然的上升趋势。无疑，上升的曲线表明，大数据的炙手可热其实只是近两年的事情。而作为最大的中文

搜索引擎，百度曲线的变化，表明大数据火在中国，或只是近一年多的事情。两条曲线的对比表明，以“大数据”作为关键词进行搜索，其曲线的陡然上升，中文比英文慢了大约半年到一年的时间。但是，在中文搜索曲线陡然上升的同时，统计工作中有关大数据的应用研究也已经同步开始了。

2012 年 8 月 23 日，来自于新加坡的时任联合国统计司司长张保罗先生，在中国国家统计局作了一场题为《海量数据、社会调查和官方统计：改进数据来源》的报告。那是笔者第一次正式地、真正地意识到“大数据”是一个具有特殊意义的概念。后来，我们走访了商务部、国家测绘局等政府部门，到百度、淘宝、腾讯、华为、苏宁、1 号店、58 同城、京东商城、天脉聚源、擎天科技、天云融创、S. CN 鞋业等知名企业，与国内诸多业界知名专家、多位美籍华裔专家进行了探讨请教，聆听了大数据代表性论著《大数据》作者涂子沛先生、《大数据时代》作者舍恩伯格先生的讲座，一点点地解开了大数据的神秘面纱。

“大数据”作为一个词语，或许是一个发现，而不是一个发明。但是，“大数据”作为一种现象，或是许多发明的结果、聚集和延续，是科学技术、生产力水平发展到一定程度后的一种必然。银行系统有海量的储户个人信息及存储信息；商场超市有海量的商品信息及其价格信息；机场记录着许多乘客的出行情况；医院记录着许多病人的检查和治疗情况；门户网站每一条新闻下面的留言，汇集成对许多现象和问题的民意；实名注册微博中的喜怒哀乐，则都是情感和态度的表达；百度、谷歌引擎的每一次使用，都可以说明 IP 那端键盘操作者到底想要什么；透过大气层中弥漫着的无数手机短信，足以掌握无数手机

使用者“打死也不说”的秘密。从我们不变的属性到可变的态度，很多很多都已经在一不留神之间，汇入了浩瀚的大数据洪流之中。

在与多位华裔美国学者、专家、官员座谈时，他们说，以往历次技术革命，中国跟在别人后面亦步亦趋。而大数据时代来了，大家的起点都差不多。中国能不能在大数据应用方面与发达国家并驾齐驱、乃至作出一些贡献呢？这些在座的专家学者里面，就有著名的《大数据》一书的作者涂子沛先生。他曾是中国大陆基层政府的一名工作人员，后来去了美国，在最恰当的时候写出了这样一本书籍。座谈之后，我们也请他做了一场报告。而半年之后，再请他作报告就非常困难了。据《中国青年报》报道，涂子沛2013年6月进行第4次国内巡讲之旅时，连早餐时间都已经被占满了。6月16日，他在招商银行深圳总部大楼与刚卸任的原行长马蔚华共进早餐。从《大数据》一书出版后，每当这位在美国供职的中国程序员回到国内，都会受到热烈追捧。和马蔚华共进早餐的第二天，他在江苏常州就“教育与大数据”的话题进行了一场听众超过1000人的讲座。其间，他与国泰君安证券股份有限公司董事长万建华吃了一顿午饭。这时，要请他讲课，只能找他的秘书安排了。这个变化，恰恰是大数据现象快速发展变化的一个缩影。

不说不做也难，因为已经唯此为大

一年多以前，我们刚刚开始研究大数据及其对统计数据和统计工作带来的影响时，一位年轻同仁在笔者的博客上留下一段英文：“Big data is like teenage sex：everyone talks about it，

nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.”意思是“大数据就像是青少年性行为：每个人都在谈论它，没有人真的知道如何去做，每个人都认为别人在做，所以每个人都声称他们正在做它。”于是，笔者回复：“很经典，但是不说不做也难，因为已经唯此为大了。”

2013年10月，我们召开以大数据为主题的科学讨论会时，有一位代表说：“那么多专家讲了一天，连大数据的概念都还没搞清。”我说：“不对，大家都是从不同的角度对这个概念进行诠释，如同说到‘文化’这个概念，马上问一百个人，难道能够得出一个统一的答案吗？也许一百个人就有一百个答案。但这并不意味着大家对这个概念没有一种相差无几的共识。大数据也一样。”

对大数据现象和概念追根溯源，大致可以分为三个阶段。

一是20世纪80年代至90年代中期，是大数据认知的萌芽阶段。1980年，美国著名未来学家阿尔文·托夫勒在《第三次浪潮》一书中将大数据盛赞为“第三次浪潮的华彩乐章”。1996年，美通社（PR Newswire Inc.）在介绍高性能平行节点技术时也提到中央处理器集群以及大数据应用。这时提到的大数据，仅是字面意义，仅指数据量大，并不涉及类型、存储方式、处理技术等。

二是20世纪90年代中期到21世纪前10年是大数据广受各界关注的阶段。数量经济学家迪博尔德（Diebold）2000年在《大数据，宏观经济度量与预测动态因素模型》一文中，讨论了如何使用大数据进行经济度量和预测。美国高德纳（Gartner Group）公司的分析师道格拉斯·兰尼（Douglas Laney）2001年

首次从大数据特征的角度对大数据进行了相对明确的定义，他强调大数据必须具备 3V 特征，即容量大（Volume）、多样化（Variety）和速度快（Velocity）。

三是 2010 年至今，是大数据战略应用被提上日程并迅速发展的阶段。2010 年，美国总统科学技术顾问委员会在呈给奥巴马总统和国会的报告《规划数字化的未来》中，要求联邦政府的每一个机构和部门，都需要制定一个应对大数据的战略。2011 年，麦肯锡公司发布报告《大数据：创新、竞争和生产力的下一个前沿》，提出了政府和企业决策者应对大数据发展的策略。2012 年 1 月，瑞士达沃斯世界经济论坛发布报告《大数据大影响》称，大数据已经成为一种新的经济资产类别，就像货币或黄金一样。2012 年 3 月 29 日，美国奥巴马政府颁布《大数据的研究和发展计划》，拟通过提高从大型复杂数据集中提取知识和观点的能力，进而加快美国科技进步的步伐。2012 年 5 月，联合国秘书长执行办公室发布报告《大数据促发展：挑战与机遇》，系统给出了在应用过程中正确运用大数据的策略建议。2012 年 6 月，经合组织 OECD 召开统计委员会第 9 届会议，发布《使用大数据作决策》研究报告，特别分析了大数据对官方统计带来的各种挑战。

2012 年，中国计算机协会决定成立“大数据专家委员会”，推动大数据的发展。2012 年 6 月，阿里巴巴集团宣布，将在集团层面设立首席数据官，负责全面推进“数据分享平台”战略。2012 年 7 月，“第二届大数据世界论坛”在北京召开。2012 年 7 月，首届中国大数据应用论坛在北京大学召开。2012 年 12 月，广东省宣布实施大数据战略，继而宣布成立大数据管理局。2013 年 7 月，“大数据时代统计学：机遇与挑战——中国统计

学高端论坛”在上海财经大学举办。全国统计学专家学者齐聚一堂，共同探讨在大数据时代统计学面临的机遇与挑战。2013年，第十七次全国统计科学讨论会在杭州举行，会议主题是“大数据背景下的统计”，这是国内第一次研究大数据与统计工作的科学研讨会。2013年11月19日，国家统计局与阿里巴巴、百度等11家企业签署大数据战略合作框架协议，共同在分享、开发、利用大数据方面进行合作，以推动大数据实现大价值，使之更好地服务于社会。这标志着在统计工作中应用大数据，已经从研究转向实操层面。

大数据大在哪里

有一个字节换算公式：

1KB（Kibibyte，千字节）＝1024B；

1MB（Mebibyte，兆字节，简称“兆”）＝1024KB；

1GB（Gigabyte，吉字节，又称“千兆”）＝1024MB；

1TB（Terabyte，万亿字节或太字节）＝1024GB；

1PB（Petabyte，千万亿字节或拍字节）＝1024TB；

1EB（Exabyte，百亿亿字节或艾字节）＝1024PB；

1ZB（Zettabyte，十万亿亿字节或泽字节）＝ 1024 EB；

1YB（Yottabyte，一亿亿亿字节或尧字节）＝ 1024 ZB；

1BB（Brontobyte，一千亿亿亿字节）＝ 1024 YB。

麦肯锡在全球研究所报告称，大数据是指大小超出传统数据库软件工具抓取、存储、管理和分析能力的数据群。1979年成立于美国马萨诸塞州霍普金市的EMC公司认为，大数据中的“大”是指大型数据集，一般在10TB规模左右；多用户把

多个数据集放在一起，形成 PB 级的数据量。维基百科（Wikipedia）的表述是，大数据是难以用现有数据库管理工具处理的兼具海量和复杂性特征的数据集成。涂子沛将大数据定义为那些大小已经超出传统意义上的尺度，一般的软件工具难以捕捉、存储、管理和分析的数据，大数据的数量级应该是“太字节”。工信部电信研究院 2014 年 5 月发布的大数据白皮书称，大数据是具有体量大、结构多样、时效强等特征的数据。我们的同仁在研究中提出，大数据是指采用多种数据收集方式，汇集不同数据源，通过采用现代信息技术和架构能够高速分析处理的、具有高度应用价值和决策支持功能的多种类型数据及其技术集成。

从存在形态看，大数据分为可以用二维表反映的结构化数据和不能以二维表反映的非结构化数据，如音频、视频、图片等；从数据来源看，大数据可分为行政记录数据、商业记录数据、互联网及搜索引擎数据三大类。大数据的特征，从最初的 3V 已经被归纳为 6V 加 1C，即数据体量大（Volume）、类型多样化（Variety）、处理速度快（Velocity）、应用价值大（Value）、数据获取与发送的方式自由灵活（Vender）、准确性（Veracity）及处理和分析难度非常大（Complexity）。

从“喝醉的海盗”到斯诺登的“泄密”

舍恩伯格先生告诉我们这样一个故事：史黛西·施奈德（Stacy Snyder）梦想成为一名教师。2006 年春天，她完成了自己的学业，并对未来充满期待。但她心仪的学校明确拒绝她，理由是她的行为与一名教师不相称，因为她的个人网页上有一

张取名“喝醉的海盗”的照片。照片里的她头戴一顶海盗帽子，举着塑料杯轻轻啜饮着。学校里的一位教师发现了这张照片，并上报给校方，校方认为网上的这张照片是不符合教师这个职业的，因为学生可能会因看到教师喝酒的照片而受到不良影响。于是，史黛西打算将这张照片从她的个人网页上删除，但她的个人网页已经被搜索引擎编录，照片也已经被网络爬虫（Web Crawler）程序存档。

而斯诺登披露的棱镜门事件更加清晰地揭示，当我们个人的行动乃至心动都融入浩瀚的大数据洪流、成为其中的一个细小浪花之后，一切已经皆在掌握之中。当人们揣着手机从一个地区到达另一个地区，马上会接到来自运营商的欢迎短信。而每一部手机都是名副其实的跟踪器和定位仪，可以精确锁定到某一座建筑物内。如果调查者和运营商联手，时间分配利用调查不用填写问卷，就可以既精准又及时地掌握所需要的信息。当人们通过博客、微博、微信、飞信表达自己情感上的喜怒哀乐时，通过网上留言、跟帖表达自己对不同事物的态度和意愿时，通过输入关键词搜索自己需要和感兴趣的结果时，不仅留下了不可磨灭的痕迹，也为后台的汇总与分析提供了最具基础性的依据。当人们坐在电脑前轻点鼠标、轻松完成一次又一次购物的时候，不仅切实感受到网购商品的物美价廉，还能享受到送货上门的方便和快捷。但与此同时，每一个网购者也不得不把自己的银行账号及其他相关信息，交给一双或几双看不见的手去掌握、控制和操作……

尽管如此，人类社会毕竟还是要向前发展的，我们不可能再退回到没有网络、没有手机的年代。其实，即使倒退半个世纪，隐私安全问题也依然存在。因为我们毕竟要到银行存款，

到医院看病，通过邮局寄信，通过单位电话或公用电话沟通。而对很多自动生成的大数据信息进行搜集处理，可以生成很多有价值的统计信息。如通过对网上交易情况的处理，可以得出很多价格和交易方面的信息；对大量临床电子病历的处理，可以进行流行病学分析，并进行医学研究；对工资收入信息的汇总分析，可以为收入分配制度的调整提供有价值的依据；……因此，法律既要保障这些合法的开发利用，也要明确指出在此过程中，个人和企业信息既不得向国家统计部门以外的任何第三方提供，也不得用来对个人和企业进行处罚，更不得对社会发布。法律要细化处罚条款。一旦发生上述情况，不仅必须承担法律责任，而且要付出一生付不起的代价。

电影搜索曲线与电影票房曲线高度相似

美国麻省理工学院承担的“十亿价格项目”（Billion Price Project），基于学术研究方法对全世界海量网上零售价格进行价格指数计算。每天实时收取50万条互联网上的商品信息，是美国政府统计收集的5倍。由于“每日网上价格指数”每天更新，研究人员和政策制定者在官方统计数据发布之前就能够依此判断价格涨跌趋势。该指数并非用于预测官方公布的通货膨胀率，而是为判断通货膨胀趋势提供实时信息。2008年9月美国雷曼倒闭时，“每日网上价格指数”很快显示出价格下降的趋势，而官方统计的CPI直到11月才显示出下降趋势。

联合国全球脉冲计划与Crimson Hexagon分析公司合作，分析了美国和印度尼西亚1400万Twitter用户中与食物、燃料和住房相关的数据，以更好地理解人们关注点。分析者以“负