第 31 卷第 2 期
2014 年 2 月
统计研究
Statistical Research
Vol. 31, No. 2
Feb. 2014

大数据时代下数据分析理念的辨析*

朱建平　章贵军　刘晓葳

内容提要: 本文在剖析了国内外大数据研究和应用现状的基础上,提出了“大数据时代”的定义,并从统计学的角度界定了“大数据”概念。同时,根据大数据的特点,本文重新审视了在大数据时代统计研究工作过程及统计思维所面临的挑战,明确了统计工作和统计研究转变的基本思路。

关键词: 大数据时代; 大数据; 统计学; 数据分析
中图分类号: C81　　文献标识码: A　　文章编号: 1002 - 4565(2014) 02 - 0010 - 08

Clarity of a Philosophy of Data Analysis During the Age of Big Data

Zhu Jianping　Zhang Guijun　Liu Xiaowei

Abstract: This paper sets forth background of the age of big data and proposes the definition of big data based on background of the age of big data after explicitly analyzing some studies and applications of big data at home and abroad. Meanwhile, based on the characteristics of big data, this paper re-examines the challenges of statistical research and ideology will face during the age of big data. Furthermore, we point out the basic thinking of the transition of statistical work and statistical research.

Key words: The Age of Big Data; Big Data; Statistics; Data Analysis

一、引言

20 世纪 50 年代一场波澜壮阔的信息公开运动在美国拉开序幕,各种信息方便了人们的生活和工作,从而信息公开为数据的可获得性提供了依据; 20 世纪 60 年代计算机硬件技术的迅速发展,促使全世界数据处理和存储不仅越来越快、越来越方便,还越来越便宜,为数据积累提供了便利; 20 世纪 70 年代最小数据集的大规模出现,使得各行各业的最小数据集越来越多,为数据结构的多元化提供了条件; 20 世纪 80 年代前期,数据在不同信息管理系统之间的共享使数据接口的标准化越来越得到强调,为数据的共享和交流提供了捷径; 20 世纪 80 年代后期,互联网概念的兴起、“普适计算”(Ubiquitous Computing) 理论的实现以及传感器对信息自动采集、传递和计算成为现实,为数据爆炸式增长提供了平台; 20 世纪 90 年代,由于数据驱动,数据呈指数增长,美国企业界、学术界也不断对此现象及其意义进行探讨,为大数据概念的广泛传播提供了途径。进入 21 世纪以来,世界上许多国家开始关注大数据的发展和应用,在此期间大数据分析和应用的学者和专家发起了关于大数据研究和应用的深入探讨,例如 Vikor Mayer-Schönberger 和 Kenneth Cukier 所著的《大数据时代》等,对大数据促进人们生活、工作与思维的变革奠定了基础。

近年来,对大数据的研究和应用不仅引起了我国自然科学和人文社会科学界的广泛重视,也受到我国中央政府的高度关注 《“十二五”国家战略性新兴产业发展规划》明确提出支持海量数据存储、处理技术的研发与产业化 《物联网“十二五”发展规划》提出将信息处理技术列为四项关键技术创新工程之一,这些是大数据产业的重要组成部分。我国国家统计局统计科学研究所于 2012 年 8 月就召开了大数据应用研究座谈会,提出了在大数据时代运用现代信息技术建立统计云架构的研究目标。2012 年 11 月国家统计局总统计师鲜祖德在会见美

* 本文获国家社会科学基金项目重大项目(13&ZD148) 和国家社会科学基金项目(11BTJ001) 资助; 本文为“大数据背景下统计调查与数据分析”研讨会特邀报告。

国华裔大数据专家学者时,明确提出国家统计局十分重视大数据在统计中的应用,并成立了专门的课题组着手研究如何通过对大数据的处理推进统计方法制度改革,改进政府统计工作。10 月 28 日至 29 日,"第十七次全国统计科学讨论会"在浙江省杭州市召开,其主题是大数据背景下的统计。从目前来看,我国大数据的理论研究和应用研究刚刚起步,学术界、企业界及政府部门对该领域的重视程度前所未有。

毫无疑问,由于计算机处理技术发生着日新月异的变化,人们处理大规模复杂数据的能力日益增强,从大规模数据中提取有价值信息的能力日益提高,人们将会迅速进入大数据时代。数据时代,不仅会带来人类自然科学技术和人文社会科学的发展变革,还会给人们的生活和工作方式带来焕然一新的变化。

统计学是一门古老的学科,已经有三百多年的历史,在自然科学和人文社会科学的发展中起到了举足轻重的作用; 统计学又是一门生命力及其旺盛的学科,他海纳百川又博采众长,随着各门具体学科的发展不断壮大。毫不例外,大数据时代的到来,给统计学科带来了发展壮大机会的同时,也使得统计学科面临着重大的挑战。怎样深刻地认识和把握这一发展契机,怎样更好地理解和应对这一重大挑战,这就迫使我们需要澄清大数据的概念、明确大数据的特征; 重新审视统计的工作过程、提出新的统计思想理念。

二、大数据概念的界定

目前,关于大数据的定义众说纷纭,对大数据的理解取决于定义者的态度和学科背景。比较有代表性的定义主要有以下几种。

维基百科给出的定义是,大数据指的是所涉及的资料规模巨大到无法透过目前主流软件工具,在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯[1]。

大数据科学家 John Rauser 提出一个简单的定义是,大数据指任何超过了一台计算机处理能力的数据[2]。

美国咨询公司麦肯锡的报告是这样定义的,大数据是指无法在一定时间内用传统数据库软件工具对其进行抓取、管理和处理的数据集合[3]。

Gartner 公司的 Merv Adrian(2011) 认为,大数据超出了常用硬件环境和软件工具在可接受的时间内为其用户收集、管理和处理数据的能力[4]。

IDC(International Data Corporation,2011) 对大数据概念的描述为: 大数据是一个看起来似乎来路不明的大的动态过程; 但是实际上,大数据并不是一个新生事物,虽然它确确实实正在走向主流并引起广泛的注意; 大数据并不是一个实体,而是一个横跨很多 IT 边界的动态活动[5]。

还有一些学者如格雷布林克(Grobelink. M) (2012) 、Forrester 的分析师布赖恩·霍普金斯(Brian Hopkins) 、鲍里斯·埃韦尔松(Boris Evelson) (2012) 和 Oracle(甲骨文) 的刘念真(2013) 等虽未给出大数据的具体定义,但是他们概括了大数据的特点[6][7][8]。格雷布林克(2012) 认为大数据具有三个特点,即多样性(Variety) 、大量性(Volume) 、高速性(Velocity) ,又称 3V 特点[6]。布赖恩·霍普金斯(Brian Hopkins) 、鲍里斯·埃韦尔松(Boris Evelson) (2012) 认为,除了格雷布林克给出的三个特性外,大数据还具有易变性(Variability) 的特点,即 4V 特点[7]。刘念真则认为大数据除了 Grobelink. M 给出的特点外,还具有真实性(Veracity) 和价值性(Value) ,即五 V 特点[8]。

上述关于大数据概念的表达方式虽然各不相同,但从各种专业的角度描述出了对大数据的理解。总的来说,我们可以从两个角度来理解大数据,如果把"大数据"看成是形容词,它描述的是大数据时代数据的特点; 如果把"大数据"看成是名词,它体现的是数据科学研究的对象。大数据是信息科技高速发展的产物,如果要全面深入理解大数据的概念,必须理解大数据产生的时代背景,然后根据大数据时代背景理解大数据概念。

(一) "大数据时代"背景介绍

格雷布林克(Grobelink. M) 在《纽约时报》2012 年 2 月的一篇专栏中称,"大数据时代"已经降临,在商业、经济及其他领域中,管理者决策越来越依靠数据分析,而不是依靠经验和直觉[6]。"大数据"概念之所以被炒得如火如荼,是因为大数据时代已经到来。

如果说 19 世纪以蒸汽机为主导的产业革命时代终结了传统的手工劳动为主的生产方式,并从而推动了人类社会生产力的变革; 那么 20 世纪以计算

机为主导的技术革命则方便了人们的生活，并推动人类生活方式发生翻天覆地的变化。我们认为，随着计算机互联网、移动互联网、物联网、车联网的大众化和博客、论坛、微信等网络交流方式的日益红火，数据资料的增长正发生着“秒新分异”的变化，大数据时代已经到来毋庸置疑。据不完全统计统计，一天之中，互联网产生的全部数据可以刻满 1.68 亿张 DVD。国际数据公司(IDC)的研究结果表明，2008 年全球产生的数据量为 0.49ZB(1024EB = 1ZB，1024PB = 1EB，1024TB = 1PB，1024GB = 1TB)，2009 年的数据量为 0.8ZB，2010 年增长为 1.2ZB，2011 年的数量高达 1.82ZB，相当于全球每人产生 200GB 以上的数据，而到 2012 年为止，人类生产的所有印刷材料的数据量是 200PB，全人类历史上所有语言资料积累的数据量大约是 5EB[3]。哈佛大学社会学教授加里·金说“大数据这是一场革命，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，所有领域都将开始这种进程。”在大数据时代，因为等同于数据的知识随处可寻，对数据的处理和分析才显得难能可贵。因此，在大数据时代，能从纷繁芜杂的数据中提取有价值的知识才是创造价值的源泉。

我们可以这样来定义大数据时代，大数据时代是建立在通过互联网、物联网等现代网络渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和展示的信息时代。在这个时代，人们几乎能够从任何数据中获得可转换为推动人们生活方式变化的有价值的知识。大数据时代的基本特征主要体现在以下几个方面。

1. 社会性。在大数据时代，从社会角度看，世界范围的计算机联网使越来越多的领域以数据流通取代产品流通，将生产演变成服务，将工业劳动演变成信息劳动。信息劳动的产品不需要离开它的原始占有者就能够被买卖和交换，这类产品能够通过计算机网络大量复制和分配而不需要额外增加费用，其价值增加是通过知识而不是手工劳动来实现的；实现这一价值的主要工具就是计算机软件。

2. 广泛性。在大数据时代，随着互联网技术的迅速崛起与普及，计算机技术不仅促进自然科学和人文社会科学各个领域的发展，而且全面融入了人们的社会生活中，人们在不同领域采集到的数据量之大，达到了前所未有的程度。同时，数据的产生、存储和处理方式发生了革命性的变化，人们的工作和生活基本上都可以用数字化表示，在一定程度上改变了人们的工作和生活方式。

3. 公开性。大数据时代展示了从信息公开运动到数据技术演化的多维画卷。在大数据时代会有越来越多的数据被开放，被交叉使用。在这个过程中，虽然考虑对于用户隐私的保护，但是大数据必然产生于一个开放的，公共的网络环境之中。这种公开性和公共性的实现取决于若干个网络开放平台或云计算服务以及一系列受到法律支持或社会公认的数据标准和规范。

4. 动态性。人们借助计算机通过互联网进入大数据时代，充分体现了大数据是基于互联网的及时动态数据，而不是历史的或严格控制环境下产生的内容。由于数据资料可以随时随地产生，因此，不仅数据资料的收集具有动态性，而且数据存储技术、数据处理技术也随时更新，即处理数据的工具也具有动态性。

(二)“大数据”的定义

我们认为大数据定义之所以众说纷纭，主要是因为大数据所涉及的内容太“大”，大家看它的角度不一样，于是出现了仁者见仁，智者见智的局面。根据大数据的历史沿革和大数据所处的时代背景，我们就可以进一步充分了解大数据的内涵。

在大数据时代，数据引领人们生活，引导商业变革和技术创新。从大数据的时代背景来看，我们可以把大数据作为研究对象，从数据本身和处理数据的技术两个方面理解大数据，这样理解大数据就有狭义和广义之分：狭义的大数据是指数据的结构形式和规模，是从数据的字面意义理解；广义的大数据不仅包括数据的结构形式和数据的规模，还包括处理数据的技术。

狭义角度的大数据，是指计量起始单位至少是 PB、EB 或 ZB 的数据规模，其不仅包括结构化数据，还包括半结构化数据和非结构化数据。我们应该从横向和纵向两个维度解读大数据：横向是指数据的规模，从这个角度来讲，大数据等同于海量数据，指大数据包含的数据规模巨大；纵向是指数据的结构形式，从这个角度来说，大数据不仅包含结构化数据，更多的是指半结构化的数据和非结构化数据，指大数据包含的数据形式多样。大数据时代，由于有 90% 的信息和知识在“结构化”数据世界之外，因

此,人们通常认为大数据的分析对象为半结构化的数据和非结构化数据。

此外,大数据时代的战略意义不仅在于掌握庞大的数据信息,而且在于如何处理数据。这就需要从数据处理技术的角度理解大数据。

广义角度的大数据,不仅包含大数据结构形式和规模,还泛指大数据的处理技术。大数据的处理技术是指能够从不断更新增长、有价值信息转瞬即逝的大数据中抓取有价值信息的能力。在大数据时代,传统针对小数据处理的技术可能不再适用。这样,就产生了专门针对大数据的处理技术,大数据的处理技术也衍生为大数据的代名词。这就意味着,广义的大数据不仅包括数据的结构形式和规模,还包括处理数据的技术。此时,大数据不仅是指数据本身,还指处理数据的能力。

不管从广义的角度,还是从狭义的角度来看,大数据的核心是数据,而数据是统计研究的对象,从大数据中寻找有价值的信息关键在于对数据进行正确的统计分析。因此,鉴定"大数据"应该在现有数据处理技术水平的基础上引入统计学的思想。

从统计学科与计算机科学的性质出发,我们可以这样来定义"大数据": 大数据指那些超过传统数据系统处理能力、超越经典统计思想研究范围、不借用网络无法用主流软件工具及技术进行单机分析的复杂数据的集合,对于这一数据集合,在一定的条件下和合理的时间内,我们可以通过现代计算机技术和创新统计方法,有目的地进行设计、获取、管理、分析,揭示隐藏在其中的有价值的模式和知识。

根据大数据的概念和其时代属性,我们认为大数据的基本特征主要体现在以下四个方面:

1. 大量性。是指大数据的数据量巨大。在大数据时代,高度发达的网络技术和承载数据资料的个人电脑、手机、平板电脑等网络工具的普及,数据资料的来源范围在不断拓展,人类获得数据资料在不断更改数据的计量单位。数据的计量单位从从 PB 到 EB 到 ZB,反映了数据量增长质的飞跃。据统计,截止 2012 年底,全球智能手机用户 13 亿,仅智能手机每月产生的数据量就有 500MB,每个月移动数据流量有 1. 3EB 之巨。

2. 多样性。是指数据类型繁多,大数据不仅包括以文本资料为主的结构化数据,还包括网络日志、音频、视频、图片、地理位置等半结构或非结构化的数据资料。多样化的数据产生的原因主要有两个方面: 一是由于非结构化数据资料的广泛存在。二是挖掘价值信息的需要,传统的数据处理对象是结构式的,我们从数据的大小多少来感受对象的特征,但这远远不够具体。很多时候,我们希望了解得更多,除了了解对象的数量特征外,我们还希望了解对象的颜色、形状、位置、甚至是人物心理活动等等,这些是传统的数据很难描述的。为了满足人们对数据分析深层次的需要,由于大数据时代对音频、视频或图片等数据资料处理技术不再是难题,于是半结构化数据和非结构化数据也成为数据处理的对象。

3. 价值性。指大数据价值巨大,但价值密度低: 大数据中存在反映人们生产活动、商业活动和心理活动各方面极具价值的信息,但由于大数据规模巨大,数据在不断更新变化,这些有价值的信息可能转瞬即逝。一般来讲,价值密度的高低与数据规模的大小成反比。以视频数据为例,一部 1 小时的视频,在连续不间断的监控中,有用数据信息出现时间可能仅有 1 秒。这就表明,大数据不仅是禁止的,更是流动的。因此,在大数据时代,对数据的接收和处理思想都需要转变,如何通过强大的机器算法更迅速地完成数据的价值"提纯"成为目前大数据背景下亟待解决的难题。

4. 高速性。指数据处理时效性高,因为大数据有价值信息存在时间短,要求能迅速有效地提取大量复杂数据中的有价值信息。根据 IDC 的"数字宇宙"的报告,预计到 2020 年,全球数据使用量将达到 35. 2ZB。在如此海量的数据面前,处理数据的效率关乎智能型企业的生死存亡。

三、如何理解大数据和分析大数据

维克多(Vikor Mayer-Schǒnberger) 在其《大数据时代》一书中并未直接给出大数据的定义,他认为在大数据时代,传统的数据分析思想应做三大转变: 一是转变抽样思想,在大数据时代,样本就是总体,要分析与某事物相关的所有数据,而不是依靠少量数据样本; 二是转变数据测量的思想,要乐于接受数据的纷繁芜杂,不再追求精确的数据; 三是不再探求难以捉摸的因果关系,转而关注事物的相关关系[10]。毫无疑问,上述三个转变均与统计研究工作息息相关,从统计研究工作角度理解维克多的三个转变会更深刻、更全面。

(一) 转变抽样调查工作思想

传统的统计学观点认为数据处理特点是通过局部样本进行统计推断,从而了解总体的规律性[11]。囿于数据收集和处理能力的限制,因此,传统的统计研究工作总是希望通过尽可能少的数据来了解总体。在这种背景下,于是,产生了各式各样的抽样调查技术。尽管如此,由于各种抽样调查工作是在事先设定目的前提下展开工作,不管多完美的抽样技术,抽到的只是总体中的一部分,样本都只是对总体片面的、部分的反映。传统的统计学观点是建立在数据收集和处理能力受到限制的基础上的,在大数据时代数据资料收集和数据处理能力对统计分析工作的影响越来越小。大数据时代,我们面对的数据样本就是过去资料的总和,样本就是总体,通过对所有与事物相关的数据进行分析,既有利于了解总体,又有利于了解局部。总的来讲,传统的统计抽样调查方法有以下几个方面的不足可以在大数据时代得到改进。

1. 抽样框不稳定,随机取样困难。传统的抽样调查方案在实施时经常碰到导致抽样框不稳定的问题: 一方面,随着网络信息技术的迅速发展,人们获得信息的途径越来越便捷,人们更换工作、外出学习和旅游的机会和次数也越来越多,这导致人口流动性加快,于是表现在对某小区居民收入水平调查过程中,经常会出现户主更换或空房的情况; 另一方面,是经营状况不稳定,有些经营者抓住市场机会使企业规模日益壮大,有些经营者经营不力导致企业破产倒闭,这就出现了在对企业经营状况调查中,抽样框中有的企业实际找不到,实际有的企业抽样框中没有的情况。

2. 事先设定调查目的,会限制调查的内容和范围。传统抽样调查工作往往是先确定调查目的,然后再根据目的和经费确定调查的方法和样本量的大小。这样做的问题是受调查目的限制,调查范围有限,即调查会有侧重点,从而不能全面反映总体。

3. 样本量有限,抽样结果经不起细分。传统抽样调查是在特定目的和一定经费控制下进行的,往往调查样本量有限,如果进一步对细分内容调查,往往由于样本量太小而不具代表性。随机采样结果经不起细分,一旦细分,随机采样结果的错误率就会大大增加[10]。如以对某地企业调查情况为例,在完成调查工作后想具体了解当地小型服装企业生产经营状况,可能抽到的样本中满足条件的企业凤毛麟角或根本没有这样的企业。在大数据时代,对数据处理的技术不再是问题,我们可以对任何规模的数据进行分析处理,可以做到既全面把握总体,又能了解局部情况。

4. 纠偏成本高,可塑性弱。正如前文所述,传统统计抽样过程中,抽样框不稳定的情况经常存在,一旦抽样框出现偏误,调查结果可能与历史结果或预计结果大相径庭; 另外,如果想了解与事先调查目的不一致的方面,或者想了解目标总体的细分结果,在传统的抽样调查思路中,解决问题的方法一般是重新设计调查方案,一切重来。在大数据时代,信息瞬息万变,待重新调整调查方案,得到的调查结果可能已经没有价值。

(二) 转变对数据精确性的要求

传统的统计研究工作要求获得的数据一般具有完整性、精确性(或准确性)、可比性与一致性等性质。在数据结构单一、数据规模小的小数据时代,由于收集的数据资料有限以及数据处理技术落后,分析数据的目的是希望尽可能用有限的数据全面准确地反映总体。那么,在小数据时代对数据精确性要求相对于其他要求是最严格的。在大数据时代,由于数据来源广泛和数据处理技术的不断进步,数据的不精确性是允许的,我们应该接受纷繁芜杂的各类数据,不应一味追求数据的精确性,以免因小失大。

1. 大数据时代,数据规模大,数据不精确性在所难免,盲目追求数据的精确性不可取。在小数据时代,无论是测量数据还是调查数据,都可能因为人为因素或自然不可控因素导致搜集到的这些数据是不精确的; 在大数据时代,数据来源渠道多,数据量多,我们在获得关于反映总体精确数据信息的同时,不可避免地会获得不精确性数据。另外,我们必须看到不精确数据的有益方面,不精确数据并不一定妨碍我们认识总体,有可能帮助我们从另一个方向更好地认识总体。

2. 大数据时代,数据不精确性不仅不会破坏总体信息,还有利于了解总体。大数据时代,越来越多的数据提供越来越多的信息,也会让人们越来越了解总体的真实情况。例如,假设某人的身高是 1 米 8,在小数据时代,由于各种原因仅能测量两次,一次是 1 米 8,一次是 1 米 6,那么很可能认为该人身体

身高为两次测量的平均值,即 1 米 7; 在大数据时代,这个人的身高测了 10 万次,其中有 10 次是 1 米 6,其他情况测得数据均为 1 米 8,那么很可能认为这个人的身高就是 1 米 8(1 米 6 作为异常值剔除) 。似乎,大数据时代,越来越多的数据在帮助我们了解总体时有点大数定律的感觉,大数定律告诉我们,随着样本数量的增加,样本平均数越来越接近总体; 但大数据告诉我们的总体信息要比大数定理更真实,大数据时代,由于样本就是总体,大数据告诉我们总体的真实情况。

3. 大数据时代,允许不精确性是针对大数据,而不是统一标准。大数据的不精确性是偶然产生的,而不是为了不精确性而制造不精确。并且,在专门性的分析领域,仍需千方百计防止不精确性发生。譬如,为了精细管理公司业务,对公司财务分析就应该越精确越好。

(三) 转变数据关系分析的重点

传统统计分析工作一般在处理数据时,会预先假定事物之间存在某种因果关系,然后在此因果关系假定的基础上构建模型并验证预先假定的因果关系。在大数据时代,由于数据规模巨大、数据结构复杂以及数据变量错综复杂,预设因果关系以及分析因果关系相对复杂。于是,在大数据时代,分析数据不再探求难以琢磨的因果关系,转而关注事物的相关关系。需要注意的是,大数据时代事物之间大数据的相关分析与传统统计学相关分析并不完全相同,主要表现在以下几个方面。

1. 分析思路不同。用传统统计方法分析问题时,往往是先假设某种关系存在,然后根据假设有针对性地计算变量之间的相关关系,这是一个“先假设,后关系”的分析思路,传统的关系计算思路适用于小数据。在大数据时代,不仅数据量庞大,变量数目往往也难以计数,“先假设,后关系”的思路不切实际。大数据关系分析往往是直接计算现象之间的相依性,是既关联又关系。另外,与传统统计分析不同的是,在小数据时代,数据量小且变量数目少,构造回归方程和估计回归方程比较容易。于是,人们在分析现象之间的相关关系时,往往会建立回归方程探求现象之间的因果关系。

2. 关系形式不同。在小数据时代,由于计算机存储和计算能力不足,大部分相关关系仅限于寻求线性关系[7]。大数据时代,现象的关系很复杂,不仅可能是线性关系,更可能是非线性函数关系。更一般的情况是,可能知道现象之间相依的程度,但并不清楚关系的形式。目前,针对结构化的海量数据,不管函数关系如何,Reshef(2011) 认为,最大信息相关系数(the maximal information coefficient , MIC) 均可度量变量之间的相关程度[12]。但有些情况可能连函数关系都没有,譬如半结构化数据变量和非结构化数据变量之间可能存在某种关联关系,但没法知道变量之间关系的形式,因此,度量相关程度的方法还有待完善。

3. 关系目的不同。传统统计研究变量之间的相关关系往往具有两个目的: 一是为了弄清楚变量之间的亲疏程度; 再则是为了探求变量之间有无因果关系,是否可以建立回归方程,然后在回归方程的基础上对因变量进行预测。一个普遍的逻辑思路并且在计算上可行的是,变量间的相关关系是一种最普遍的关系,因果关系是特殊的相关关系,相关关系往往能取代因果关系,即有因果关系必有相关关系,但有相关关系不一定能找到因果关系。所以传统的统计学往往在相关关系基础上寻找因果关系。在大数据时代,统计研究的目的就是寻找变量或现象之间的相关关系,然后根据变量或现象之间的相关关系进行由此及彼、由表及里的关联预测。大数据时代一般不做原因分析,一方面是因为数据结构和数据关系错综复杂,很难在变量间建立函数关系并在此基础上探讨因果关系,寻找因果关系的时间成本高昂; 另一方面是大数据具有价值密度低、数据处理快的特点,大数据处理的是流式数据,由于数据规模的不断变化,变量间的因果关系具有时效性,往往存在“此一时,彼一时”的情况,探寻因果关系往往有点得不偿失。

四、大数据对统计学科和统计研究工作的影响

对于统计学科的发展而言,大数据时代带来的不仅是变革,更多的是统计学发展壮大的机会。大数据将使传统统计学作为研究具体问题的方法科学发生改变,改变统计研究的工作程序,改变统计学研究具体科学的深度和广度。然而,大数据并不会改变传统统计学的性质。因此,对统计学而言,大数据带来的是挑战和机遇,同时也将壮大统计学的生命力。

(一) 大数据拓展了统计学的研究对象

大数据对每个领域都会造成影响,统计学也不例外。统计学的研究对象是指统计研究所要认识的客体,统计学的研究对象是客观事物的数量特征和数量关系,数量性是统计学研究对象的基本特点。但传统的统计学认为数据是来自试验或调查的数值,同时又认为并不是任何一种数量都可以作为统计对象。在大数据时代,不仅任何一种以结构数据度量的数量可以作为统计研究对象,而且不能用数量关系衡量的如文本、图片、视频、声音、动画、地理位置等半结构或非结构数据都可以作为统计研究的对象。从某种意义上来说,大数据拓展了统计研究的对象,也扩展了统计研究工作的范畴。

(二) 大数据影响统计计算的规范

传统统计学根据一定的数据计算规范,如用平均数、方差、相对数等反映客观事物量的特征、量的界限、量的关系等等,并且可以根据具体计算规范计算具体数值。然而,由于半结构化数据和非结构化数据并不能根据计算规范计算平均数、方差、相对数等数值。显然,在大数据时代直接利用计算规范计算平均数、方差、相对数等指标将遇到挑战。

(三) 大数据影响统计研究工作过程

统计学是关于数据搜集、整理、归纳和分析的方法论科学,这些工作构成了统计学科学体系的核心内容。根据统计学的核心内容,统计研究的全过程包括统计设计、收集数据、整理与分析和统计资料的积累、开发与应用等四个基本环节。在大数据时代,网络资料异常丰富,数据不再是通过试验或调查抽样的方式获得的,统计工作面对的数据就是总体数据,即样本就是总体。在这种情况下,传统的数据收集方法不再可行,针对大数据的数据收集往往通过传感器自动采集数据,数据资料不再需要设计和人工收集。大数据时代,统计研究的过程,只包括数据整理与分析和数据的积累、开发与应用两个基本环节。

1. 数据整理与分析。

统计数据的整理一般指对统计数据进行汇总,包括确定总体的处理方法和确定汇总哪些指标两个方面,具体而言,有统计资料的审核、资料的分组和汇总、编制统计表或绘制统计图、统计数据资料的积累、保管和公布等四个步骤。在针对大数据的整理过程中,由于数据资料巨大、数据类型复杂以及要求数据处理速度快等特点,对数据的分组和汇总、编制统计表或绘制统计图常常无法实施,统计资料的整理往往只有资料的审核和资料的储存两个环节。但大数据的审核和储存不同于传统统计意义上的资料审核和资料保存。

(1) 数据的审核。传统的数据审核是为了检查原始数据的完整性与准确性,而大数据的审核往往是在兼顾数据处理速度和预测的准确性前提下,确定要处理的数据规模,即确定数据量的级别。Pat Helland 认为处理海量数据不可避免地导致部分信息的损失[13]。另外,大数据本身是杂乱无章的,是有噪音的、混杂的、内部相关的和不稳定的,尽管如此,有噪音的数据也因为其能发现隐藏的关系模式和知识而比小样本更有价值[13]。因此,反映研究对象的数据可能是正确的,也有可能是错误的,但不管哪一种,都是大数据的一部分,只要是法规条件允许下,所有数据都是有价值的,一般不作删除或替换。

(2) 数据的储存。传统的数据保存是将经过审核、分组汇总和编制统计图表的统计资料作为重要的资料积累和保管起来。大数据的储存一般是为了控制存储成本,按照法规计划制定存储数据的规模。

2. 数据的积累、开发与应用。

(1) 数据的积累。传统的统计工作根据事先确定的研究目的对数据进行分类、汇总,然后保存数据,便于日后分析和查询。对大数据而言,有价值的信息往往是在对数据进行处理之后发现的,并不是在事先目的前提下处理数据发现的。Viktor 认为大数据的混乱应该是一种标准途径,而不应该竭力避免[10]。大数据的复杂性是客观存在的,在大数据积累的过程中,不要轻易地做出简单的处理,一方面是因为大数据规模庞大、结构复杂,很难对其进行简单的分类整理;另一方面是对大数据的简单整理,如排序、分类、删除,可能造成新的混乱,破坏了原有数据的真实性并因而损失原有数据中有价值的信息。

(2) 数据的开发。传统数据由于样本量小、解决问题目的性强,数据价值往往存在时效性特点,即数据价值会随着使用次数的增加或时间流逝而降低。而大数据具有流动性,会随着时间的日积月累而不断“壮大”,往往具有不断推陈出新、重塑价值的可能,数据价值具有“再生性”。在大数据时代,数据就像一个神奇的钻石矿,其价值被挖掘之后还能源源不断产生新的价值。可以说,在大数据时代,

数据不但不会贬值、过时,而且还会不断增值,为了更全面、深入地了解研究对象,往往需要对数据进行整合,即将部分数据合并,整合的数据因为对研究对象反映更全面,常常会发现新问题,创造新价值。从这个角度来说,整合的数据价值往往大于部分价值。因此,分析研究大数据,应怀有谦卑的心理,不用担心数据量的庞大,并且要有整合大数据的勇气。

(3) 数据的应用。传统数据应用的目的通常是为了解释现象和预测未来,即探寻相关关系和因果关系,然后在相关关系和因果关系的基础上进行预测。在大数据时代,建立在相关关系方法基础上的预测是大数据的核心。由于大数据具有价值性特点,这就表明在大数据时代商业竞争的环境里,要求对大数据的处理迅速及时。这里需要提及的是,由于数据量庞大,结构复杂,在数据的应用过程中,对数据结果解释,可视化就显得尤为重要,Agrawal D. 等认为大数据时代,数据分析结果可视化很有必要,有助于解释分析结果[14]。美国计算机学会的数字图书馆中第一篇使用"大数据"的文章是迈克尔·考克斯和大卫·埃尔斯沃思在第八届美国电气和电子工程师协会(IEEE)关于可视化的会议论文集中发表的《为外存模型可视化而应用控制程序请求页面调度》,他们在该文的篇首提到"可视化对计算机系统提出了一个有趣的挑战:通常情况下数据集相当大,耗尽了主存储器、本地磁盘、甚至是远程磁盘的存储容量"。虽然如此,但我们依然要关注数据的可视化,因为它是连接数据和心灵最便捷的桥梁。

五、小结

大数据从狭义的角度来讲,不仅是指数据规模巨大,还指数据结构复杂;从广义角度来讲,大数据还指处理大规模复杂数据的技术。由于在大数据时代数据意味着信息,所有有价值的信息都源自对数据的处理。大数据时代,数据对个人或家庭而言意味着良机,对厂商而言数据意味着商机,对国家而言数据意味着发展契机。对统计工作者而言,这种改变不仅意味着拓宽了统计研究的范畴、丰富了统计研究的内容、增强了统计学的生命力,还意味着统计工作及统计研究的四个转变。

1. 转变统计研究过程。传统的统计研究过程包括统计设计、收集数据、整理与分析和统计资料的积累、开发与应用等四个基本环节。大数据时代,由于数据规模巨大、数据结构复杂等特点,以及整理数据可能损坏原有数据中有价值信息,针对大数据的统计研究过程仅包括数据整理与分析和数据的积累、开发与应用两个基本环节。进一步的分析表明,大数据整理与分析过程仅指数据储存工作。总的说来,大数据统计研究过程包括数据储存和数据的积累、开发与应用两个环节。

2. 转变统计研究方法。传统的统计研究方法,如建立回归方程、估计模型参数、检验参数估计结果等因为大数据的特点而无法实施,对大数据的统计分析是以相关关系为基础展开的。但针对大数据的相关关系分析不同于传统的相关关系的分析,传统的相关分析基本是线性相关分析,大数据研究的相关关系分析的不仅是线性相关,更多的是非线性相关以及不明确函数形式的线性关系。

3. 转变统计研究目的。传统统计研究的目的主要是为了探寻现象(或变量)间的相关关系、因果关系以及建立在相关关系或因果关系基础上的预测分析。大数据由于数据规模巨大和数据结构复杂以及要求数据处理速度快等特点,因果分析往往不可行。大数据时代统计研究分析的目的主要是研究现象间的相关关系以及建立在相关分析基础上的预测分析。

4. 转变统计研究工作思想。传统统计研究工作中,囿于计算技术的限制,总是希望用尽量少的数据和相对复杂的模型尽量获取有价值的信息。传统的统计抽样调查方法虽然在小数据时代有助于节省费用、了解总体信息,但可能存在抽样框不稳定、调查样本片面、调查结果经不起细分以及纠偏成本高昂的缺陷。在大数据时代,样本即总体,由于计算机超前的数据处理能力,可以通过分析处理大数据了解总体各方面的信息。另外,还需将传统统计质量管理控制中的事后检验转变为事先预测,以及转变尽量利用复杂模型的思想为巧用简单模型的思想。

参考文献

[1] http://zh.wikipedia.org/wiki/:大数据.维基百科,2012-10-5.

[2] Mckinsey Global Institute, Big Data: The next frontier for innovation, Competition and productivity,2011-5.

[3] http://www.networkworld.com/news/2012/051012-big-data-259147.html.

[4] http:// www.teradatamagazine.com/v11n01/Features/Big-Data/: Merv Adrian. Big Data [N/OL]. Teradata Magazine.

我国政府采购的价格监测*

王群勇　陈燕平

内容提要: 价格监管的缺失是导致我国政府采购价格虚高的重要原因,也是"政府采购价格应低于市场平均价格"这一法律条文变成一纸空文的根源。本文提出了我国政府采购的一个价格监测理论模型,依据采购拍卖理论测算采购预警价格。与单纯的将采购价格与市场平均价格相比,这种方法更充分地考虑了市场环境和投标环境的竞争因素。对天津市政府采购中心 2012 年 8 月至 2013 年 7 月份协议采购商品的实证分析表明,该模型具有良好的拟合能力和稳健性。蒙特卡洛模拟实验表明,该模型对于异常交易的发现率明显提高。

关键词: 政府采购; 价格监测模型; 采购预警价格

中图分类号: C829.2　　文献标识码: A　　文章编号: 1002 - 4565(2014)02 - 0018 - 06

Price Monitoring of Government Procurement in China

Wang Qunyong & Chen Yanping

Abstract: The lack of price supervision is a major reason for the artificially high procurement prices and it is also the resource of the impracticable legal provision that "the government procurement price should be lower than the average market price". The paper proposed a new price supervision model which computes the procurement warning price based on procurement auction theory. Compared with the method of monitoring the procurement price using average market price, our method considers the competing factors of market environment and bid environment. The positive analysis of the goods in Tianjin Government Procurement Center from August 2012 to July 2013 reveals that the model has high fitness and robustness. A Monte Carlo simulation verifies that the model improves the detection rate of abnormal procurements. The model provides an efficient and liable precautionary and warning mechanism for the government procurement in China.

Key words: Government Procurement; Price Monitoring Model; Procurement Warning Price

一、引言与文献

自 1998 年我国引入政府采购制度以来,政府采购迅速发展,2012 年我国政府采购金额突破了万亿元。中国政府采购担负着节约财政资金、引导产业调整、促进中小企业发展等多重使命,却屡屡因为价格虚高而饱受诟病。究其原因,既包括我国财政预算制度和政府采购法规的不完善,也包括各地采购机制设计的漏洞,归根结底是因为价格监管的缺失。价格监管不到位,围标、中标、串谋、腐败等问题才能

* 本文为中央高校基本科研业务费专项资金资助项目《政府采购的机制设计与绩效评估》(NKZXB1224)成果。

[5] http://www.emc.com/collateral/demos/microsites/emc-digital...2011/topic1.sw.

[6] Grobelink M. Big-data computing: Creating revolutionary breakthroughs in commerce, science and society [N/OL]. 2012 - 10 - 02.

[7] Brian Hopkins, Boris Evelson. Expand your digital horizon with big data [N/OL]. 2011 - 9 - 30.

[8] http://wenku.baidu.com/view/abfb3a1552d380eb62946d9d.html: 刘念真. 利用 Oracle 信息模型驾驭大数据.

[9] http://www.banyuetan.org: 大数据时代降临. 半月谈网. 2012 - 09 - 22.

[10] Viktor Mayer-Schönberger. 大数据时代 [M]. 杭州: 浙江人民出版社, 2012.

[11] 韦博成. 漫谈统计学的应用与发展(I) [J]. 数理统计与管理, 2011, 30(1): 85 - 97.

[12] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, Pardis C. Sabeti, Detectin Novel Associations in Large Data Sets [J], Science, 2011, 12: 1518 - 1524.

够有隙可乘。建立采购价格监管是杜绝采购价格虚高的一道重要的防火墙。

对政府采购实施价格监管是国际上通行的做法。投标人在采购拍卖中可能会出现一些非理性的行为,相对于其成本和期望收益报出过高或过低的价格。采购价格过高或过低都是非正常的。McCaffer and Pettitt(1976)较早地分析了出现过高报价的原因。比如,投标人对项目并不是特别感兴趣,或者投标的目的只是保持对采购项目的熟识度,为以后投标做准备(Skitmore,2002)。这种投标被称作象征性投标(courtesy bid)。由于采购部门一般会采用最低价中标的规则,因此象征性投标一般不会影响最终采购结果,更多的讨论集中在采购价格过低的情形上。与报价过高不同,报价过低的企业面临着无法弥补成本而违约的风险,因此监测过低报价也是对采购人和投标者权益的保护。几种原因可能会引发这种过低的报价。比如,企业面临财务困境,亟需一份合约帮助其度过难关;或者是供应商缺乏投标经验,亦或严重低估了项目成本(Gunduz and Karacan,2009)。有时企业刻意以低价投标以保证其市场份额,或者进入一个新市场(Alexandersson and Hulten 2007),这种行为被称作掠夺性投标(predatory bid),是掠夺性定价在采购拍卖中的直接体现。Conti,Givanni 和 Naldi(2012)利用排序比较算法监测异常的低报价,模拟结果显示,准确报警率和错误报警率都随着投标人的增加而降低;报价分散程度越高,准确报警率越低,错误报警率越高。

统计学中将偏离大部分观测值的数值称为异常值(outlier),对异常报价的检测也即归于对异常值的统计检测。针对于异常报价,西班牙、意大利、德国、土耳其等很多国家制定了自己的执行标准。比如,西班牙和意大利按照最低报价与平均保价的偏离程度进行判断,但 Conti 和 Naldi(2008)的评估结果发现,这种方法严重依赖于投标人数和报价的分散程度。德国则是通过比较最低报价与次低报价来判断异常值,称作排序比较算法,不过德国联邦采购局并没有对这种方法发布过评估报告(Zanza,2004)。欧盟也已经建立了一种信号机制来监测过低的异常报价(EUWG,1999)。Conti、Giovanni 和 Naldi(2012)讨论了对检测出来的异常报价的处理方法,拒绝或者隔离。私人采购人可以任意选择其中一种,而公共采购人则受制于当地的法律法规。如果投标人数比较多,直接拒绝异常报价就成为经常的选择,这样不仅可以缩减采购时间,也免去了对异常交易进行详细调查所带来的负担。但直接拒绝不可避免地会冤枉一些貌似异常的正常报价,因此欧盟 2004 年规定,除非特殊情况,对于异常报价应该将项目暂时隔离,并进行仔细核查。

我国对政府采购价格的监测源于《政府采购法》,其第十七条明确规定,政府采购价格应低于市场平均价格,但没有具体的实施细则。比如,"市场平均价"的概念如何界定,市场价格由谁来进行市场调查,又由谁来进行统计和发布?这一系列问题目前都没有明确的答案。进一步来说,把市场平均价格作为政府采购的价格监测标准并不尽合理。比如,相同条件下 10 家供应商竞争形成的采购价格比 5 家供应商竞争形成的采购价格应该更低,如果两种情形下采用同一个监测标准,那么就存在供应商之间达成共谋的空间(当然,这并不意味着供应商数目越多越好)。类似地,相比较 10 台计算机的采购数量而言,100 台计算机的采购数量可能更容易产生一个更低的采购价格。因此,合理的价格监测应该充分考虑供应商数目、品牌数目、采购数量和投标规则等多种投标竞争因素。

[13] http://queue.acm.org/detail.cfm? id = 1988603. 2011 - 05 - 23: Pat Helland, If you have too much data, then good enough is not good enough [N/OL]. Acmqueue.

[14] Agrawal D, Bernstein P, Bertino E, et al. Challenges and Opportunities with Big Data-A community white paper developed by leading researchers across the United States [R/OL]. Computing Community Consortium 2012 - 10 - 02.

作者简介

朱建平,男,2003 年获南开大学理学博士学位,现为厦门大学经济学院教授、博士生导师、厦门大学数据挖掘研究中心主任。中国统计学会副会长、教育部高等学校统计学类专业教学指导委员会秘书长、中国统计教育学会常务理事。研究方向为数理统计、数据挖掘。

章贵军,男,现为厦门大学经济学院统计系博士研究生。研究方向为数据挖掘、计量经济模型。

刘晓葳,男,现为厦门大学经济学院统计系博士研究生。研究方向为数据挖掘、计量经济模型。

(责任编辑: 程　晞)