DOI:10.19343/j.cnki.11-1302/c.2016.02.001

大数据时代对传统统计学变革的思考*

朱建平 张悦涵

内容提要: 本文在大数据时代背景下,将统计学与大数据有机结合,剖析了大数据时代给统计学带来的变革,阐述了大数据为传统统计学带来的发展机遇。为了保持统计学旺盛的生命力,本文对统计学的发展提出了几点思考。

关键词: 大数据; 大数据时代; 统计学; 变革
中图分类号: C829.2 文献标识码: A 文章编号: 1002 - 4565(2016)02 - 0003 - 07

Reflections on Conventional Statistics in the Big Data Era

Zhu Jianping & Zhang Yuehan

Abstract: In this paper, statistics and big data are integrated based on the operation of conventional statistics. We analyze the changes and development opportunities brought by big data to statistics. Also, in order to keep active, we propose some suggestions on the development of statistics.

Key words: Big Data; Big Data Era; Statistics; Change

一、引言

美国百科全书把统计学的定义界定为“一门在不确定性方面为了做出正确的推断而进行搜集、分析定量数据的科学和艺术”,大英百科全书认为“统计学是一门搜集数据,分析数据,并根据数据进行推断的艺术和科学,最初与政府搜集数据有关,现在包括了范围广泛的方法和理论”,中国百科全书将统计学定义为“一门研究怎样有效地搜集、整理和分析带有随机性的数据,以对所考察的问题做出推断或预测,直至为采取一定的决策和行动提供依据和建议的学科”。由此可见,统计学与数据科学息息相关。

三年前,奥巴马的数据团队通过收集、存储和分析选民数据帮助其获得了总统连任; 马云领导的阿里巴巴早在 2008 年已把大数据作为一项公司基本战略。在不知不觉中,我们已经进入了大数据时代。大数据时代的建立,首先需要收集大量的广泛的数据,这些数据的来源渠道通常为现代网络渠道,如互联网、物联网等; 其次,需要有先进的存储设备,传统的存储设备已经不能容纳如此大量的数据; 再次是通过对大数据的分析,提取出有价值的信息; 最后需要将这些信息展示于公众。

通过对统计学发展过程的回顾,可以看出,统计学在应用的过程中生长和发展。它的生命力在于应用。在当今社会,统计起着“神经系统”的作用。统计是连接社会再生产各个环节、各个要素的中介,是商品生产和商品交换的先导,对经济活动起着灵敏有效的调节作用。例如,商品的生产和经营活动的依据是市场经济的统计信息。市场经济信息主要有两个方面: 一是客户对商品的需求,包括商品数量、品种、质量和规格的要求等; 二是各类商品生产、供应以及价格的变化。商品生产者生产商品的种类、数量、质量改进等问题都要根据市场需求信息和价格信息等研究决定; 商品经营者所决定的商品购买、库存、定价等问题也需要根据生产、需求信息及其变化趋势研究决定。在大数据时代,以上这些统计信息的获得不再局限于电话调查、问卷调查等高成本、低收益的方式,而是可以借助网络、移动通信等方式。同时,数据的质量也不再受到主观因素的限制。

* 本文获国家社会科学基金项目重大项目“大数据与统计学理论的发展研究”(13&ZD148)、国家统计局统计科学研究所研究基地项目“网络舆情分析的统计方法及其应用”(201407)资助。

由于大数据的产生，使得统计学的定义、思维方式、作用都不同于传统统计。毫无疑问，伴随着大数据时代的到来，统计学进入了一个发展的新阶段。

二、大数据时代下传统统计学的变革

大数据时代的到来，给统计学的发展带来了前所未有的机遇，但同时，也对统计学提出了更多的挑战。在此，本文将从以下 7 个方面阐述大数据时代下传统统计学的变革。

1. 样本概念的深化。

除普查以外，传统统计学离不开样本，样本是研究中实际观测或调查的一部分个体，一个可用的样本必须能够正确地反映总体情况。大数据时代，样本的概念不再这么简单，由于此时数据大部分为网络数据，因此可以将其分为两种类型：一是静态数据，即当客户在查看数据时已经被生成好了，没有和服务器数据库进行交互的数据，直接在客户端创建完毕，对于这种数据，样本等同于总体，这样无需去提取样本并检测样本的可用性，减少了成本，并且总体本身对总体的反映更为准确，减少了误差；二是动态数据，比如数据是随着时间的推移而变化的，此时，总体表现为历史长河中所有数据的总和，而我们分析的对象为“样本”，这里的“样本”与传统样本的概念不同，因其并非局限于随机抽取的数据，更可以是选定的与分析目的相关的数据。

2. 数据类型的扩大。

传统意义上的数据为结构化数据，即可以用常规统计指标或图表表现出来的定量数据或专门设计的定性数据，有固定的结构和标准。大数据是指不仅包括结构化数据，还包含非结构化数据、半结构化数据或异构数据，即一切可以记录和存储的信号，具有多样化的特点，并且传统的统计指标等不一定可以将其完整地表述出来；其次，大数据的存储不同于传统的数据存储方式，有固定的格式和结构，对于大数据的数据库来说，可以直接将所探测到的信号自动容纳到其中；最后，由于大数据大部分是指非结构化以及半结构化数据，因此对数据的识别和分类也是多样的，通常用网络信息系统作为识别工具。

3. 收集概念的扩展。

传统统计中，数据的收集需要根据统计分析的目的进行，过程包括设计调查方案、严格控制调查流程，因此具有低效率、高成本的缺点。在大数据时代，对数据的收集分为三步，首先是数据预处理，包括识别与整理；其次是数据分析，目的为提炼有价值的信息；最后为数据存储。我们拥有超大量可选择的数据，同时，在存储能力、分析能力、甄别数据的真伪、选择关联物、提炼和利用数据、确定分析节点等方面，都需要斟酌。然而，这并不代表大数据时代搜集的数据是万能的，我们仍然需要有针对性地搜集，不仅如此，还存在着安全性和成本的问题。因此，我们应该将传统方法中有针对性的收集数据的优点和现代方法中利用高效率的技术和广泛数据源的优点结合起来，收集一切相关数据。

4. 数据来源的不同。

传统统计中是根据研究目的去收集数据，数据来源通常是已知的，很容易对数据提供者的身份进行识别或进行事后核对。而大数据的来源则很难追溯，由于大数据的来源一般为信息网络系统，不具有很强的目的性，更是一切被人为记录的信号（尽管信号有其目的性，但多数为发散的），并且很难识别记录者的身份。在大数据时代，努力打造统计数据来源第二轨，就显得尤为重要。

5. 量化方式的变化。

传统数据为结构化数据，对数据的量化方式已经相当成熟，并且比较容易得到可以直接进行分析的数据结果。大数据时代主要面对的是非结构化数据，Franks 说过“几乎没有哪种分析过程能够直接对非结构化数据进行分析，也无法直接从非结构化的数据中得出结论”。目前，计算机学界已着手研发处理非结构化数据的技术，从统计角度直接处理非结构化数据，或将其量化成结构化数据，这是一个重要的研究领域。

6. 分析思维的改变。

我们从统计分析、实证分析、推断分析三个方面论述大数据时代传统统计学分析思维的改变。

第一，传统的统计分析过程分为三步，定性、定量、再定性。首先通过经验判断找到统计方向，即目的；其次对数据进行量化、分析、处理等；最后根据结果得出结论。大数据时代，统计分析过程为“定量—定性”，基础性的工作就是找到“定量的回应”，直接从各种“定量的回应”中找出有价值的、为我们所需要的数据，并通过分析找到数据的特征和数量关系，进而据此做出判断与决策。

第二，传统的统计实证分析，思路是“假设—验

证”,即首先提出假设,接着按照统计方法进行数据的收集、分析、展示,最后通过所得到的结论对假设进行验证。事实证明,这种实证分析存在很大误差。大数据时代,实证分析的思路是“发现—总结”,为了更全面、深入地了解研究对象,需要对数据进行整合,从中去寻找关系、发现规律,然后再加以总结、形成结论,这将有助于发现更多意外的“发现”。

第三,传统的统计推断分析过程是以分布理论为基础,在概率保证的前提下,对总体进行推断,通常是根据样本特征去推断总体特征,推断是否正确取决于样本的好坏。现在,其过程变成了以实际分布为基础,根据总体的特征进行概率的判断,在静态或者动态的某个时点,大数据所需处理的对象为总体数据,不需要根据分布理论推断总体特征,而要根据计算方法进行推断。

7. 统计软件的增多。

传统统计学以统计模型和软件为基础进行数据分析处理,统计模型的作用在于对数据间的数量关系进行构建,统计软件是分析和处理数据的工具,需要研究者自主输入经过处理的数据,以及统计模型的公式等。常见的统计软件有 SAS、R、STATA、SPSS、MATLAB 等。大数据所依赖的数据分析技术为非关系型的,以数据中心为基础。若将统计软件与大数据结合起来,则统计分析的过程可以在很大程度上简化。

综上所述,大数据时代的来临,对传统统计学的变革从样本的定义方法一直到数据分析的思维与技术均有所体现。可以看出,大数据使我们对数据的利用取得了更大的主动权,将促使传统统计学的迅速发展。

三、大数据给统计学带来的发展

统计学的优势在于“以小见大”,但容易产生误差等问题,对于大数据来说,可以利用更多甚至是总体的数据,数据的限制因素已经成为历史。统计学可以与大数据进行合作,不仅可以做到以小见大,还可以做到由繁入简,在大数据的基础上大大提高统计效率、模型拟合度和推断准确性。本文将从以下 5 个方面阐述大数据给统计学带来的发展。

1. 统计质量得以提高。

针对统计质量而言,国际数据标准 SDDS 确定了两条规则作为评估统计数据质量的标准,我们可以据此归纳出四个原则,即: 适用性、准确性、时效性、平衡性,来把握统计质量的内涵。

适用性,是指收集的统计信息符合用户的需求。保证统计信息适用性的根本是使统计信息最大化地满足用户。大数据的广泛覆盖性能够在很大程度上满足适用性的原则。以 CPI 为例,传统的价格统计涉及的商品和销售点种类繁多,且随着社会的进步、经济的发展和人们消费观念的改变,对于动态的数据需要及时进行调整,这必定会产生很大的误差,使得统计工作者不能保证统计数据是否适用于用户的需求。而基于大数据的“在线价格指数”不再必须通过样本进行分析,统计数据可以包含所有的商品和线上销售网点,可以实现通过总体进行分析,使统计误差大幅度下降。

时效性,是从统计调查的各个方面缩短时间。另外,为了使用户及时掌握、使用统计信息,对于统计数据应预先公布发布日期,按时发布,并建立规范的发布制度。传统统计数据具有滞后性和低频率等缺点,而大数据由于其来源为信息网络,具有及时性和时效性的优点。仍然用 CPI 的统计数据举例,CPI 的发布频率为每月,如我国的 CPI 通常在每个月 9 日发布上个月的 CPI,由此可见,CPI 的发布存在滞后; 而“在线价格指数”能够根据市场的变化对价格进行即时的更新与汇总,提高了统计信息的时效性,并且“在线价格指数”的频率可以从每月提高到每天甚至更短时间,据此分析出来的通货膨胀规律相比传统统计的准确率大大提高。

准确性,主要是估算值与“真值”之间的差异度。实际上所谓“真值”是不可知的,一般目标为保证统计误差在可接受的范围内变动,据此保证统计的准确性,通过分析抽样误差、人为误差、计数误差、模型设计误差等多个对准确性产生影响的因素,测算统计估值的变动系数、标准差、协方差等。由于大数据的全面性,因此可以通过减小统计过程中的人为误差保证统计结果的准确性。例如,传统样本搜集方法中,受调查者意识到自己在接受调查会有意对真实情况进行掩饰,这会导致调查所得数据无法真实反映现实。大数据可以在受调查者无意识的情况下收集他们的信息、获得数据,如手机现在已经成为居民必不可少的工具之一,当移动通讯用户带着手机进行出行、吃饭等一系列日常活动时,移动通信商就已经在用户无意识的情况下通过跟踪定位手机

采集到了用户的位置信息。这种方法获得的数据显然比传统调查方法所获得的数据更为真实准确，从而在此基础上的统计分析结果更为可信。

平衡性，即协调性，在统计学中指数据的协调能力，造成数据平衡性缺失的原因有很多，比如数据使用者对数据的理解与数据发布者有差异。大数据时代通过网络数据资源，有助于数据平衡性的提高。根据 SDDS 的第二条规则，在公布统计数据的同时，在统计框架内公布有关总量数据的分项，并公布有关数据的比较和核对方法与结果，有利于支持和鼓励使用者对数据进行核对和检验，借此提高数据平衡性。

2. 统计成本得以降低。

统计成本是进行一项统计调查或开展统计工作所实际付出的代价，是统计工作过程中耗费的人力、财力和物力的总和。下面从调查方法与数据利用率两个角度来阐述大数据时代统计成本的降低。

首先，从收集数据的方法来看，传统的统计数据收集方法主要依靠调查，如调查问卷、电话采访，或者通过查询统计报表。开展普查，可能就要动用全国的力量。这些方法都存在缺点，准确性得不到保证，并且统计成本相当可观。在大数据时代，数据的获得途径为信息网络、移动通信等，因此从统计成本的各个要素来看，大数据时代的统计成本会大幅下降，而且可以得到更大规模、更高准确性的数据。

其次，从所得数据的利用率来看，传统统计中，统计资料的失效过期是一个长期无法得到改善的事实，即使是依靠巨大的财政以及社会投入取得的普查资料，由于其开发方式单一、传递被动以及向公众发布的手段方式的局限，也得不到及时广泛的利用。而在大数据时代，对数据的初始收集没有很强的目标性，首先，数据可以服务于多个研究目的，无需再根据目的来重新收集数据；其次，数据被多次利用意味着数据价值的增加；再次，相比于传统统计，每个统计目的收集数据的成本会大幅下降。

最后，统计成本还体现在公众获取方面。对此，SDDS 制定了两项规划：一是成员国要预先公布日历表，据此进行统计数据的公布。预先公布统计数据发布日程表不仅可以使使用者根据自己的实际情况合理安排利用数据，还表明统计工作管理制度的完善和数据编制的透明。二是必须保证有关各方同时收到所发布的统计数据。统计数据作为一项公共产品的基本特征之一就是官方统计数据的公布，公众的基本要求就是及时和机会均等地获得统计数据。因此 SDDS 通过此项规定体现公平的原则并满足公众需求。数据发布时可依次提供概括性数据、详细的数据，当局应至少提供一个地方使得公众可以进入并有权使用数据，保证公众可以在第一时间获得发布的数据。SDDS 的目的是指导成员国并对其提供一套在数据收集和公布方面的标准，使各国在向公众提供具有全面性、时效性、可行性和准确性的数据时遵守共同的依据。在大数据时代，无论是数据的获取、分析还是发布，皆通过网络进行，SDDS 的规划变得更为可行。

3. 统计学科体系得以延伸。

大数据时代，对于统计学的发展应该用发展、辩证的眼光去看待，统计学应当在大数据的思想框架下构建新的学科体系。将大数据总体统计的思想和方法纳入统计学学科体系是非常必要的，例如，在统计学的教学内容中，将样本统计和总体统计相结合。样本统计对样本的要求是能够正确地代表总体，这就要求总体的观察单位必须是同质的，在现实生活中这种理想情况不容易达到，而基于大数据的总体统计恰好能够弥补样本统计的这一不足之处。

数据挖掘又称数据采矿，是数据库知识发现中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中的信息的过程。涉及机器学习、人工智能、模式识别、数据可视化等模块，也属于数据处理的范畴。因此，统计学应该对其进行充分的利用，将统计学原理应用到数据挖掘的技术中。当今大数据时代，统计学也应与计算机紧密结合，以数据挖掘为契机，进一步延伸和完善统计学科体系，培养具有现代统计技术、数据挖掘技术与计算机技术的复合人才。同时，统计学不仅要注重与其他学科的结合，更需要注重自身学科的提高，在原理、技术、方法等方面认真钻研，与时俱进，谋求创新与突破。

4. 统计学作用得以扩大。

传统统计由于成本、观念等的影响，主要用于行业和部门统计，为行业和部门制定与完善政策服务。在大数据时代，统计学不仅可以在统计领域得到更为快速的发展，更可以将统计原理与方法应用到其他学科，如金融、医学、计算机等，使统计学发挥更大的价值。

以数据分析为例，我们应该看到，计算机同数学

一样,都可以作为统计分析的工具。数学可以为统计学提供更坚固的理论基础,计算机则可以使统计分析更加方便快捷,并能够解决复杂的数据处理问题。当今社会是一个信息社会,由于计算机和网络的普及,使信息传递的质量发生了根本性变革。离开了计算机,统计学的发展也会停滞不前,为了使统计学不至于落在时代后面,我们应该为统计专业的学生开设必要的计算机课程,如数据库、数据结构、统计软件、算法设计、程序编码等等。我们也应将课程计划不再局限于传统统计学,也应设计原本是统计学科外发展起来的如计算机定向数据分析方法等。这无疑会大大丰富统计学发展的内涵,更大地发挥统计学的作用。

5. 统计学专业就业需求得以提升。

大数据对统计专业学生的就业起到了相当大的改善作用。当今社会,无数的行业,包括政府、企业、个人都希望能从大数据这座金矿中挖掘出对自己有价值的金子,但只懂得行业知识对于数据挖掘来说是远远不够的,还需要与专业的数据分析技能相结合,就是统计工作者和数据分析师。在大数据时代,他们利用自己的专业优势,将各种不同类型的数据转换为有价值的信息,对各行各业的发展都起到了促进作用,并可以提高各行业专家的思维水平及其工作效率。在未来,统计工作者和数据分析师的作用不容小觑,他们的地位也必定会得到大幅度提高。

众所周知,我国统计工作领域的三大巨头是政府统计、部门统计、民间统计。传统意义上,政府及各个部门是统计学学生就业的首选。然而,随着大数据时代的来临,越来越多的毕业生选择发展空间更为广阔的民间统计。民间统计相对于政府统计来说,涉及范围十分广泛,包括各类统计咨询公司、统计调查公司、统计研究院等,介于市场和企业、行业之间。通过民间统计,企业可以做出更为理智的决策,市场可以得到更好的发展。民间统计的发展前景十分广阔,不难想象,随着大数据时代的来临,统计学作用的提高,民间统计必会成为统计专业毕业生选择就业的主要渠道之一。

四、大数据时代下对统计学的几点思考

大数据时代的来临使得大量的数据呈现在人们眼前,我们突然发现一切社会现象从本质上来说就是一个统计的规律,到最后都是一个统计的规律。因此,大数据时代给统计学带来新的生命力,同时也引发了对统计学的再思考。

1. 完善总体、个体及样本的定义方式。

传统的统计分析,首先要从总体中抽样,然后通过研究样本的性质等确定总体的特征。因此是从总体中获得数据,即必须先确定总体范围作为研究目标,然后通过抽样获得样本数据,进而分析总体。大数据则与之相反,是先有数据,后有总体,没有事先定义的目标,更无需根据目标确定总体,只有在某一时点的全部数据所对应的总体概念。因为个体的不确定性,数据是动态的,无法事先依靠数据库的单位对其进行编制,更为复杂的是,由于数据随时间是流动的,这个时点与下一个时点的数据是不同的,因此也很难在事后识别个体。在网络系统中,同一个个体可以有多个称谓和符号,同一个称谓或符号也可以代表多个个体,而且个体异位的情况也不少见(即某一个体以另一个个体的名义完成某种行为),因此我们对于大数据往往看得到总体数据的外形,但对于个体很难考究。但是对于大数据分析来说,对个体进行身份识别仍然是必要的,这就需要一个总体口径。需要我们改变传统意义上总体与个体的定义方式,进而,传统意义上样本的定义方式已经不能够让我们从大数据库中提取样本数据了。当然,由于大数据的动态性,在任意时点的总体,都可以将其作为一个截面样本。

2. 转变抽样调查的功能以拓展其应用空间。

对于传统统计学来说,收集数据的重要方法之一是抽样调查。但抽样调查毕竟不是全面调查,需要有足够好的样本才可以正确反映总体的特征,具有不稳定性、误差大等缺点。进入了大数据时代,如前文所提到的,可以对总体进行分析,不一定要抽取样本,但这并不意味着抽样调查可以退出历史舞台了。首先,目前来看,并不是所有数据都可以通过网络信息系统获得,因为并非所有的产业都已经实现智能化,还有很多数据只能通过传统方法获得,即抽样调查; 其次,即使对于网络数据,在某些情况下,对总体进行分析也并非最优选择,例如,当遇到均匀度很大的总体时,随机抽取部分作为样本进行分析就可以得到我们想要的结果,而且比分析总体更为省时省力。当然,此时抽样调查的有些功能需要转变以适应大数据时代: 一是统计机构抽样调查所获得

的数据具有权威性，可以作为对照基础和验证依据与大数据分析进行对比。在大数据时代，尽管网络数据具有很多优点，但也并非完美，数据的偏倚性是不可避免的，在进行统计分析时，会有很多不可用的数据。此时可以将二者进行互补，抽样数据对互联网数据进行校正，互联网数据作为抽样数据的补充。二是可以把抽样调查看作从混杂的数据中寻找规律或关系的线索，以便更好地进行数据挖掘、快速探测分析。这需要从源源不断的数据流中抽取足以满足统计目的和精度的样本，及时调整已经获得的样本，使得热门数据与感兴趣的数据进入样本。

3. 如何使结构化数据与非结构化数据对接。

在大数据时代，数据的概念从结构化数据扩展为结构化数据和非结构化数据。数据概念的拓展必然会导致如何有效实现结构化数据与非结构化数据的对接问题。通过特定的方法，这是完全可能的。但在实际工作中，各种类型的数据是大数据分析的基础，因此我们必须要提高对各种类型数据的描述和测度的能力。传统的统计分析侧重推断，而基于大数据的统计分析则更加侧重描述，这是为了更为准确地进行推断。如何既能针对统计目的和需要收集结构化数据，又能从大量非结构化数据中挖掘出有价值的信息，使两者相辅相成、有机结合，是一个有待探讨的新课题，如何实现非结构化数据结构化使其更好地进行分析，以及结构化数据是否能用非结构化表示使其更加通俗易懂等，都是这个课题中需要解决的问题。

4. 采用新的梳理与分类方法处理大数据。

传统统计学按照预先设定的方案进行数据梳理与分类，所根据的指标以及所得到的分类都是结构化的，对数据进行梳理与分类是数据预处理的必要步骤，是统计分析的组成部分。但对于大数据，由于数据的来源、数据的形式与数据的表现方式等都是多样化的，若想如传统统计学一样在研究之前对信息的种类、分类的依据标识、标识之间的关系、类与类之间的区别度等进行严格的设定，是不可能实现的，只能在对数据进行预处理后，根据数据本身的特征进行补充与完善。此时传统的数据梳理与分类方法已经不再实用，必须开创与发展适应大数据时代的数据梳理与分类方法，并据此开辟新的数据分析的路径。这是对大数据开展分析的重要前提。

5. 不确定性的来源和表现产生差异。

在经济学中，不确定性指经济行为者实现不能准确地知道自己的某种决策的结果。或者说，只要经济行为者的一种决策可能产生的结果大于一种，就会产生不确定性。对于企业来说，不确定性对其产生的影响可能仅影响一次营销活动，也可能对企业的生存构成威胁。而统计学学科的根源就是不确定性。传统统计学若要研究不确定性，首先是进行数据收集，然后进行抽样。其不确定性表现在多个方面：样本的获得、模型的选取、对于总体代表性的推断等。在大数据的研究中，可以针对总体，不一定要通过随机获得的样本进行分析，但是个体之间仍然存在着差异性，并且数据的来源、对个体的识别、模型的选取、数据梳理与分类方法、结论的判断等都存在着多样化的趋势，这些都构成大数据研究中的不确定性。总而言之，由于在大数据时代我们已经掌握了一定条件下的完全信息，此时的不确定性只来自于数据来源的复杂多样、总体的动态化。

6. 相关关系分析与因果关系分析并重。

维克多(Vikor Mayer-Schǒnberger)在其《大数据时代》一书中认为“通过给我们找到一个现象的良好的关联物，相关关系可以帮助我们捕捉现在和预测未来”以及“建立在相关关系分析法基础上的预测是大数据的核心”。毫无疑问，人们认识和掌控事物、继而做出预测判断的重要途径，是从超大量数据中发现各种真实存在的相关关系，而大数据时代衍生出的对统计分析的思路与技术的创新开阔了我们的眼界，使我们发现了很多之前没有发现的事物之间的联系。因此大数据时代的重要任务之一是大力开展相关分析。然而，大数据时代的要求不仅限于“是什么”，还有“为什么”，只有这样，才能更好地理解“是什么”。真正的数据需要知道其原因和背景。因果关系是重要的，它决定了数据分析的深度。只知相关性而不知因果，那么数据分析的深度只有一半，一旦分析的方向偏离我们的预想，就会不知所措。而通过因果关系，可以帮助我们更好地利用相关关系，如有些事物有共同的原因，或可以导致共同的结果，据此可以推断这些事物之间的相关性。这可以帮助我们做出更为理智的决策，甚至更好地预测未来。深入挖掘事物的因果关系，以便更进一步利用好大数据。同时，因果关系的基础是相关关系。我们不能将相关分析与因果分析对立，而需将其并重，使其互相补充。

7. 结合多种统计方法全面驾驭大数据。

在传统统计中,最主要的研究方法就是归纳推断法,分析样本数据的特征,在此基础上推断出总体的特征。对于大数据,归纳法依然是大数据分析的主要方法,我们依然要通过具体个体的特征归纳出总体的特征,依然要从个体的信息中发现新知识。但是对于大数据,若只是重视一般或者总体特征的归纳,那就太浪费了。有些类别甚至是个体,或者某些异常值,都可能推断出全新的结论或做出全新的预测。因此我们还需要对个体的信息进行深入挖掘,还需要根据已有的分布特征和相关知识经验去对其他更具体的规律进行推理分析,更深层次地去挖掘事物之间的关联关系,据此对新事物做出判断,即演绎推理法。演绎法可以帮助我们在已有知识的基础上,对其进行更深入的挖掘,防止研究中忽略某些重要、细小的特征。归纳法与演绎法的有机结合,可以从大数据的偶然性中发现必然性,并利用全面数据的必然性去观察偶然性、认识偶然性、甚至利用偶然性,从而提高驾驭偶然性的能力。

8. 统计思维与现代信息技术相结合。

统计技术对于传统意义上的收集和分析数据已经相当成熟、自成体系,但当数据量很大,尤其是包含很多非结构化数据时,统计技术是远远不够的。计算问题是首要问题,由于数据量大,导致计算量大,这就需要我们将传统的统计技术与现代信息技术紧密结合,以便将统计学与大数据更好地结合。

五、结束语

大数据的产生对统计学具有划时代的意义,大数据以其价值性、多样性、大量性、高速性的特征弥补了统计学高成本、高误差的劣势,但这并不意味着统计学的时代结束了,我们对大数据的搜索、聚类、分类等还需要依赖统计学的方法,因此大数据离不开统计学。大数据时代的到来,提高了统计质量,降低了统计成本,使得统计学发挥作用的领域增大,并且使统计学科得以延伸,提高了统计学科在自然科学和社会科学中的地位,这是大数据给传统统计带来的机遇。在大数据时代,传统统计学也面临着挑战,要求其改变对样本的认识,改变对不确定性的认识,建立新的数据梳理与分类的方法,强化结构化数据与非结构化数据的对接,转变抽样调查的功能,结合归纳演绎法与推断演绎法,并重相关分析与因果分析以及结合统计思想与云计算技术。我们应该牢牢抓住大数据带来的机遇,积极应对挑战,将大数据与统计学有机地结合,在未来的科学发展过程中,保持统计学旺盛的生命力。

参考文献

[1] Lynch C. Big data: How do your data grow? [J]. Nature, 2008, 455(7209).

[2] Rifkin J. The third industrial revolution: How lateral power is transforming energy, the Economy, and the World [M]. New York: Palgrave Macmillan, 2012.

[3] Bughin J. Chui M, Manyika J. Clouds, big data and smart assets: Ten tech-enabled business trends to watch [J]. McKinsey Quarterly, 2010(8).

[4] Lavalle S, Lesser E, Shockley R, et al. Big data, analytics and the path from insights to value [J]. MIT Sloan Management Review, 2011, 52(2).

[5] MacKinsey Global Institute. 2011. Big data: The next frontier for innovation, competition and productivity. June 2011. Lexington, KY: McKinsey & Company.

[6] Dabenport T H, Barth P, Bean R. How big data is different [J]. MIT Sloan Management Review, 2012, 53(5).

[7] 朱建平,章贵军,刘晓葳. 大数据时代下数据分析理念的辨析[J]. 统计研究,2014(2).

[8] 袁卫. 机遇与挑战——写在统计学科成为一级学科之际[J]. 统计研究,2011(11).

[9] 李金昌,大数据与统计新思维[J]. 统计研究,2014(1).

[10] 朱建平著. 世纪之交中国统计学科的回顾与思考[M]. 中国经济出版社,1999, 12.

[11] 曾鸿,丰敏轩. 大数据与统计变革[J]. 中国统计,2013(9).

[12] 肖红叶. 中国经济统计学科建设 30 年回顾与评论——基于三大框架事件的研究[J]. 统计研究. 2010(2).

[13] 朱怀庆,大数据时代对本科经管类统计学教学的影响及策略[J]. 高等教育研究,2014(3).

[14] 程开明,庄燕杰. 大数据背景下的统计[J]. 统计研究,2014(1).

作者简介

朱建平,男,河南浚县人,2003 年获南开大学理学博士学位,现任厦门大学管理学院 MBA 中心教授、博士生导师、厦门大学数据挖掘研究中心主任,浙江工商大学现代商贸流通体系建设协同创新中心首席专家,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。研究方向为数理统计、数据挖掘。

张悦涵,女,河北保定人,现为厦门大学数据挖掘研究中心和经济学院统计系硕士研究生。研究方向为数据挖掘、计量经济模型。

(责任编辑: 方　原)