大数据与统计新思维*

李金昌

内容提要: 最近,《大数据时代》等几本书引起了广泛的关注,大数据正在改变着人们的行为与思维,那么以数据为研究对象的统计学该如何应对,本文基于对大数据的理解,认为统计思维需要发生三个方面的改变,即要改变认识数据的思维、收集数据的思维和分析数据的思维。其中,数据分析思维又要在统计分析过程、实证分析思路、推断分析逻辑等方面发生变化,同时统计分析评价的标准也要有所调整。围绕这些变化,本文提出需要从八个方面去积极应对大数据,以促使统计学科跟上时代的步伐。

关键词: 大数据; 统计思维; 统计学
中图分类号: C829.2　　文献标识码: A　　文章编号: 1002 - 4565(2014)01 - 0010 - 06

Big Data and New Mind on Statistics

Li Jinchang

Abstract: The latest publication of a book such as "*Big Data*: *A Revolution That Will Transform How We Live*, *Work*, *and Think*" has captured the public attention. With the big data changing the way people think and behave, how should the development of statistics, a discipline that aims at data, take its course? Based on its understanding of the big data, this paper puts forward three dimensions in which the mind on statistics need to be changed: the interpretation of data, the idea of data collection and the view of data analysis, where the process of statistical analysis, the mode of empirical analysis and the logic of inferential analysis ,and also the evaluation standards of statistical analysis should be adjusted. According to those changes, this paper suggests that the big data be actively dealt with from eight perspectives, in order to keep the science of statistics to abreast of the times.

Key words: Big Data; Mind on Statistics; Statistics

最近,译著《大数据时代》[1](英国 Viktor Mayer-Schǒnberger,Kenneth Cukier 著)和《驾驭大数据》[2](美国 Bill Franks 著),以及我国学者涂子沛[3]、郭晓科的《大数据》[4]等几本书引起了广泛的关注,其他各种媒体关于大数据的讨论也层出不穷,大数据已成为流行语。有人认为,大数据是一场新的革命,将横扫一切领域,重构世界。不少国家已将大数据作为国家发展战略,而商业领域更是将其视为下一个投资的宝库。毫无疑问,大数据时代已经来临,它正在悄悄地改变着人们的行为与思维,难以阻挡,无法抗拒。在计算机科学、电子商务等领域已率先在大数据技术开发与应用方面做出不俗成就的时候,以数据为研究对象的统计学该如何应对?无动于衷还是盲目追从?正确的态度应该是理性对待、积极跟进、改变思维、谋求发展。

一、对大数据的初步认识

到底什么是大数据,不同的学科领域、不同行业的从业人员肯定会有不同的理解。与传统意义上的数据相比,大数据的"大"与"数据"都有了新的含义,绝不仅仅是体量的问题,更重要的是数据的内涵问题。或许,"大"与"数据"根本就不能分开,只有把"大数据"当作一个整体概念来理解才有意义。那么从统计学的角度,我们该如何来理解大数据?笔者认为大数据不是基于人工设计、借助传统方法而获得的有限、固定、不连续、不可扩充的结构型数据,而是基于现代信息技术与工具可以自动记录、储存和连续扩充的、大大超出传统统计记录与储存能

* 本文为浙江省高校人文社科重点研究基地(统计学)成果之一。本文为第十七次全国统计科学讨论会特邀论文。

力的一切类型的数据。有人用 4V[4] (Volume, Variety、Velocity 和 Value) 来形容大数据的特征①, 最根本之处就是数字化基础上的数据化。通俗地说,大数据就是一切可记录信号的集合。

如果说,传统统计研究的数据是有意收集的结构化的样本数据,那么现在我们面对的数据则是一切可以记录和存储、源源不断扩充、超大容量的各种类型的数据。样本数据与大数据的这种区别,具有什么样的统计学意义? 我们知道,样本数据是按照特定研究目的、依据抽样方案获得的格式化的数据,不仅数据量有限,而且如果过程偏离方案,数据就不能满足要求。基于样本数据所进行的分析,其空间十分有限——通常无法满足多层次、多角度的需要,若遇到抽样方案事先未曾考虑到的问题,数据的不可扩充性缺点就暴露无疑。而大数据是一切可以通过现代信息技术记录和量化的数据,不仅所蕴含的信息量巨大,而且不受各种框框的限制——任何种类的数据都来者不拒、也无法抵拒。不难发现,大数据相比于样本数据的最大优点是,具有巨大的数据选择空间,可以进行多维、多角度的数据分析。更为重要的是,由于大数据的大体量与多样性,样本不足以呈现的某些规律,大数据可以体现; 样本不足以捕捉的某些弱小信息,大数据可以覆盖; 样本中被认为异常的值,大数据得以认可。这将极大地提高我们认识现象的能力,避免丢失很多重要的信息,避免失去很多决策选择的机会。

这里,我们自然就想到了大量观察与大数据这两个概念中的"大"的区别。对于传统的统计研究方法而言,大量观察法是基础,是收集数据的基本理论依据,其主要思想是要对足够量的个体进行调查观察,以确保有足够的微观基础来消除或削弱个体差异对整体特征的影响,足以归纳出关于总体的数量规律。所以,这里的"大"是足够的意思。大量观察法的极端情况就是普查,但限于各种因素不能经常进行,所以一般情况下只能进行抽样调查,这就需要精确计算最小的样本量。基于大量观察法获得的样本数据才符合大数法则或大数定律,才能用以推断总体。而大数据则指不限量的数据,是基于现代信息技术的一切可以记录的全体数据,其特征之一就是尽量多地包含数据,它与样本容量无关,只与信息来源的数量与储存容量有关。因此,这里的"大"是全体的意思。

可见,统计学的研究对象没有变,变的是数据的来源、体量、类型、速度与量化的方式。这种变化对统计研究带来了什么样的挑战? 《大数据时代》提出了三个最显著的变化: 一是样本等于总体,二是不再追求精确性,三是相关分析比因果分析更重要[1]。这些观点具有很强的震撼力,迫使我们对现有的统计研究思维进行反思。尽管这些观点值得进一步商榷,但至少告诉我们这样一个道理: 统计研究对象的基础变了,统计思维也要跟着变化,否则统计研究的对象只是全部数据的 5%,而且越来越少,那又怎么能说统计学是一门关于数据的科学呢? 又怎么去完善和发展开展数据分析研究的统计方法论呢?

二、统计思维的变化

改变统计思维,是大数据时代的必然要求。否则,统计学科就有可能被大数据的潮流所吞没,至少会被边缘化,失去一次重要的参与推动历史变革的机遇。当然,统计思维的变化应该以一个永恒不变的主题为前提,那就是通过数据分析去揭示事物的真相,这个真相就是事物的生存规律、联系规律和发展规律。也就是说,数据分析要以数据背后的数据去还原事物的本来面目,以达到求真的目的。如果说,我们原来限于各种条件只能根据有限的样本数据去实现这个目的,那么现在我们则可以在很多方面借助大数据去实现这个目的,关键就看我们开展数据分析的能力有多大,或者说利用大数据、从一切数据中提取有价值信息的能力有多大——因为大数据无疑增加了统计分析的难度,而这又首先取决于我们统计思维能否适应大数据时代的变化。正如迈尔 - 舍恩伯格所说: 大数据发展的核心动力就是人类测量、记录和分析世界的渴望[1]。

那么,统计思维应该发生怎样的变化? 笔者认为主要要有如下三大变化:

(一) 认识数据的思维要变化

前面已经提到,与传统数据相比,大数据不仅体量大、变化快,而且其来源、类型和量化方式都发生了根本性的变化,使得数据杂乱、多样、不规整。

首先,从来源上看,传统的数据收集因为具有很

① 也有指 4V 是 Volume, Velocity, Variety 和 Veracity; 或者 Volume, Velocity, Variety 和 Vitality。

强的针对性，因此数据的提供者大多是确定的，身份特征是可识别的，有的还可以进行事后核对。但大数据通常来源于物联网，不是为了特定的数据收集目的而产生，而是人们一切可记录的信号（当然，任何信号的产生都有其目的，但它们是发散的），并且身份识别十分困难。从某种意义上讲，大数据来源的微观基础是很难追溯的。

其次，从类型上看，传统数据基本上是结构型数据，即定量数据加上少量专门设计的定性数据，格式化，有标准，可以用常规的统计指标或统计图表加以表现。但大数据更多的是非结构型数据、半结构型数据或异构数据，包括了一切可记录、可存储的信号，多样化、无标准、难以用传统的统计指标或统计图表加以表现。同时，不同的网络信息系统有不同的数据识别方式，相互之间也没用统一的数据分类标准。再者，现在有的数据库是非关系型的数据库，不需要预先设定记录结构即可自动包容大量各种各样的数据。

第三，从量化方式上看，传统数据的量化处理已经有一整套较为完整的方式与过程，量化的结果可直接用于各种运算与分析。但大数据中大量的非结构化数据如何量化（结构化）、如何从中提取信息、如何与结构化数据对接是一个崭新的问题。正如 Franks 所说“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。”[2] 更为重要的是，“量化”的含义恐怕也不一样了，即此“量化”不一定等同于彼“量化”，量化结果的表现形式自然也不相同。显然，我们不能套用已有的方式去量化非结构化数据。

可以说，大数据是杂乱、不规整、良莠不齐的，但我们不能因此而回避它、拒绝它，只能接纳它、包容它。我们需要将统计研究的对象范围从结构型数据扩展到一切数据，需要重新思考数据的定义和分类方法，并以此为基础发展和创新统计分析方法。从某种意义上讲，没有无用的数据，只有未被欣赏的数据，关键是我们从哪个角度看数据。

（二）收集数据的思维要变化

收集数据是开展统计分析的前提，“没有黏土，如何做砖？”以往，收集统计数据的思维是先确定统计分析研究的目的，然后需要什么数据就收集什么数据，所以要精心设计调查方案，严格执行每个流程，但往往是投入大而数据量有限。现在，我们拥有了大数据，就等于拥有了超大量可选择的数据——备选“黏土”的体量与种类都极大地增加了，所要做的最重要的工作就是比较与选择，因此我们的思维应该是如何充分利用大数据，凡是大数据源中能找到的数据就不再需要进行专门的调查。

但是，由于大数据来源与种类的多样性，以及数据增加的快速性，我们在享受数据的丰富性的同时也不得不面临这样一些困境：存储能力够不够，分析能力够不够（是否及时、充分），如何甄别数据的真伪，如何选择关联物，如何提炼和利用数据，如何确定分析节点？现在 TB 级的数据库已经很多，PB 级的数据库也不少见，以后还会出现 EB、甚至 ZB、YB 级的数据库。今天的大数据，明天就不再是大数据。这样一来，电子存储能力能否跟得上数据增加的速度就成为首要的问题。如果让数据库自动更新就有可能失去一些宝贵的数据信息，而到了一定级别以后扩充存储容量或对数据进行拷贝，其代价是十分巨大的，因此我们不得不对数据进行分类、筛选，有针对地删除那些垃圾数据、不重要或次重要的数据。如果说以前有针对地获得数据叫做收集，那么今后有选择地删除数据就意味着收集。也就是说，大数据时代的数据收集将更多的是从已有的超大量数据中进行再过滤、再选择。因此，我们要做好丢弃一部分数据的准备。

当然，并不是任何数据都可以从现成的大数据中获得，这里存在一个针对性、安全性和成本比较问题。因此，我们既要继续采用传统的方式方法去收集特定需要的数据，又要善于利用现代网络信息技术和各种数据源去收集一切相关的数据，并善于从大数据中进行再过滤、再选择。问题在于什么是无用的或不重要的数据？该如何过滤与选择数据？这就需要对已经存在的数据进行重要性分析、真伪识别和关联物定位。

此外，大的数据库可能需要将信息分散在不同的硬盘或电脑上，这样一来，在不能同步更新数据信息的情况下如何选择、调用和匹配数据又是一个问题。因此从某种意义上讲，从大数据中收集数据就是识别、整理、提炼、汲取（删除）、分配和存储元数据的过程。

（三）分析数据的思维要变化

基于上述两个变化，数据分析的思维必然要跟着变化，那就是要主动利用现代信息技术与各种软

件工具从大数据中挖掘出有价值的信息,并在这个过程中丰富和发展统计分析方法。

关于数据分析思维的变化,特别需要强调三点:

第一,传统的统计分析过程是“定性—定量—再定性”,第一个定性是为了找准定量分析的方向,主要靠经验判断,这在数据短缺、分析运算手段有限的情况下很重要。现在我们是在大数据中找矿,直接依赖数据分析做出判断,因此基础性的工作就是找到“定量的回应”,这在存储能力大为增强、分析技术与分析速度大为提高的今天,探测“定量的回应”变得越来越简单,所要做的就是直接从各种“定量的回应”中找出那些真正的、重要的数量特征和数量关系,得出可以作为判断或决策依据的结论,因此统计分析的过程可以简化为“定量—定性”,从而大大提高得到新的定性结论的可能性。

第二,传统的统计实证分析,一般都要先根据研究目的提出某种假设,然后通过数据的收集与分析去验证该假设是否成立,其分析思路是“假设—验证”,但这种验证往往由于受到假设的局限、指标选择的失当、所需数据的缺失而得不出真正的结论。特别是,一旦假设本身不科学、不符合实际,那么分析结论就毫无用处、甚至扭曲事实真相。事实证明,很多这样的实证分析纯粹是为了凑合假设。现在,我们有了大数据,可以不受任何假设的限制而从中去寻找关系、发现规律,然后再加以总结、形成结论。也就是说,分析的思路是“发现—总结”。这将极大地丰富统计分析的资源与空间,有助于发现更多意外的“发现”。

第三,传统的统计推断分析,通常是基于分布理论,以一定的概率为保证,根据样本特征去推断总体特征,其逻辑关系是“分布理论—概率保证—总体推断”,推断的评判标准与具体样本无关,但推断是否正确却取决于样本的好坏。现在,大数据强调的是全体数据,总体特征不再需要根据分布理论进行推断,只需进行计数或计量处理即可。不仅如此,还可以根据全面数据和实际分布来判断其中出现某类情况的可能性有多大,其逻辑关系变成了“实际分布—总体特征—概率判断”,也即概率不再是事先预设,而是基于实际分布得出的判断。按照迈尔 - 舍恩伯格的观点,这个概率判断就可用于预测了。

伴随着上述三大变化,统计分析评价的标准又该如何变化?传统统计分析的评价标准无非两个方面,一是可靠性评价,二是有效性评价,而这两种评价都因抽样而生。所谓可靠性评价是指用样本去推断总体有多大的把握程度,是以概率来度量的——有时表现为置信水平,有时表现为显著性水平。特别是在假设检验和模型拟合度评价中,显著性水平怎么定是一个难题,一直存在争议,因为所参照的分布类型不同其统计量就不同,显著性评价的临界值就不同,而临界值又与显著性水平的高低直接相关。然而在大数据的背景下,大数据在一定程度上就是全体数据,我们可以对全体数据进行计数或计量分析,这就不存在以样本推断总体的问题了,那么这时还有没有可靠性的问题?还要不要确定置信水平?怎么确定?依据是什么?如何比较来自不同容量数据库的分析结论的可靠性?

所谓有效性评价指的是真实性,即误差大小。这里又有两个相关的概念: 准确性与精确性。准确性一般是指一个观察值与真实值的吻合程度,通常情况下是无法做出测度的; 而精确性一般指样本统计量分布的离散程度,以抽样分布的标准差来衡量。很显然,精确性是针对样本数据而言的。也就是说样本数据既有精确性问题又有准确性问题,样本数据中的误差既包括抽样误差也可能包括非抽样误差。抽样误差可以基于抽样分布理论进行计算和控制,而非抽样误差只能通过各种方式加以识别或判断,但多数情况下由于样本量不是太大而可以得到较好的防范。但对于大数据,由于它是全体数据,因而不再有抽样误差问题,只有非抽样误差问题,也就是说大数据的真实性只表现为准确性而非精确性。然而由于大数据是超大量数据,再加上混杂性与多样性,因此其非抽样误差很难防范与控制,这就使得准确性评价问题变得更为困难——如何测度?标准怎样?

三、积极应对大数据

面对大数据,我们唯有积极应对,别无选择。如何应对,需要考虑以下几个方面:

(一) 需要改变总体、个体乃至样本的定义方式

传统的统计分析,是先有总体,再有数据,即必须先确定总体范围和个体单位,再收集个体数据,分析总体。但对大数据来说,情况完全不同了,是先有数据,再有总体。从某种意义上说,大数据的产生系统多数是非总体式的,即无事先定义的目标总体,只

有与各个时点相对应的事后总体，原因就在于个体是不确定的，是变化着的，是无法事先编制名录库的，这与传统的总体与个体有很大的不同。更为复杂的是，事后个体的识别也很困难，因为同一个个体可能有多个不同的网络符号或称谓，而不同网络系统的相同符号（称谓）也未必就是同一个个体，而且还经常存在个体异位的情况（即某一个体利用另一个体的符号完成某种行为），因此我们对于大数据往往是只见“数据”的外形而不见“个体”的真容。但对大数据的分析，仍然有一个总体口径问题，依然需要识别个体身份。这就需要我们改变总体与个体的定义方式——尽管它们的内涵没有变。与此对应，如果要从大数据库中提取样本数据，那么样本的定义方式也需要改变。当然，考虑到大数据的流动变化性，任何时点的总体都可以被理解为一个截面样本。

（二）需要改变对不确定性的认识

众所周知，统计学是为了认识和研究事物的不确定性而产生的，因为无论是自然现象还是社会经济现象，都时时处处充满着因个体的差异性而引起的不确定性，因为在大多数情况下我们缺乏足够的信息或缺乏足够的知识去利用有效信息[5]，而人们总是期望通过量化事物的不确定性去发现规律、揭示真相，认识不确定性背后的必然性。要研究不确定性就需要收集数据，在只能进行抽样观测的情况下，这种不确定性就表现为如何获得样本、如何推断总体（包括估计与检验）和如何构建模型等方面。对于大数据，仍然存在着个体的差异性，区别只在于它包括了一定条件下的所有个体，而不是随机获得的一个样本。这样，大数据的不确定性就不再是样本的获取与总体的推断，而是数据的来源、个体的识别、信息的量化、数据的分类、关联物的选择、节点的确定，以及结论的可能性判断等方面。可以说，大数据的不确定性只来自于其来源的多样性与混杂性，以及由于个体的可变性所引起的总体多变性，而不是同类个体之间的差异性——因为我们已经掌握了一定条件下的完全信息。

（三）需要建立新的数据梳理与分类方法

大数据的多样性与混杂性，以及先有数据、后有总体的特点，原有的数据梳理与分类方法将受到诸多的限制。传统的数据梳理与分类是按照预先设定的方案进行的，标志与指标的关系、分类标识与分组规则等都是结构化的，既是对有针对性地收集的数据的加工，也是统计分析的组成部分。但对于大数据，由于新的网络语言、新的信息内容、新的数据表现形式不断出现，使得会产生哪些种类的信息、有哪些可以利用的分类标识、不同标识之间是什么关系、类与类之间的识别度有多大、信息与个体之间的对应关系如何等，都无法事先加以严格设定或控制，往往需要事后进行补充或完善。面对超大量的数据，我们从何下手？只能从数据本身入手，从观察数据分布特征入手。这就需要采用不同的数据梳理与分类方法。否则，要想寻找到能有效开展数据分析的路径是不可能的。因此根据大数据的特点，创新与发展数据的梳理与分类方法，是有效开展大数据分析的重要前提。这里需要强调的是，能否建立起能自动进行初步的数据梳理与分类的简单模型？因为从技术上讲，我们已经具备了一定的对大数据进行多次迭代建模的算法。

（四）需要强化结构化数据与非结构化数据的对接研究

有效实现结构化数据与非结构化数据的对接，是数据概念拓展的必然结果。尽管大数据是超大量数据，但大数据不能涵盖所有的数据，因此传统意义上的结构化数据与大数据中的非结构化数据必将长期并存。大数据时代的来临，使得数据收集、存储与分析的能力大为增强，而且步伐越来越快，但出于针对性与安全性考虑，总有一些结构化数据要通过专门的方式去收集而不能依赖于公共网络系统（例如政府统计数据，专题研究数据）。这样，如何既能有针对性地收集所需的结构化数据，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，就成了一个新的课题，值得探讨的问题包括非结构化数据如何结构化或结构化数据能否采用非结构化的表现形式等。通过特定的方法，实现结构化数据与非结构化数据的转化与对接是完全可能的。但要实现这种对接，必须要增强对各种类型数据进行测度与描述的能力，否则大数据分析就没有全面牢固的基础。如果说传统的基于样本数据的统计分析侧重于推断，那么基于大数据的统计分析需要更加关注描述。

（五）需要转变抽样调查的功能

对于传统的数据收集而言，抽样调查是最重要的方式。尽管样本只是总体中的很小一部分，但由

于依据科学的抽样理论,科学设计的抽样调查能够确保数据的精确度和可靠性。但抽样调查毕竟存在着信息量有限、不可连续扩充、前期准备工作要求高等缺陷,很难满足日益增长的数据需求。现在有了大数据,我们应该利用一切可以利用的、尽量多的数据来进行分析而不是仅局限于样本数据。但这是否意味着抽样调查可以退出历史舞台呢? 笔者认为还为时过早,在信息化、数字化、物联网还不能全覆盖的情况下,仍然还有很多数据信息需要通过抽样调查的方式去获取。与此同时,尽管我们可以对大数据进行全体分析,但考虑到成本与效率因素,在很多情况下抽样分析仍然是不错的或明智的选择。当然,抽样调查也要适当转变其功能以便进一步拓展其应用空间: 一是可以把抽样调查获得的数据作为大数据分析的对照基础与验证依据; 二是可以把抽样调查作为数据挖掘、快速进行探测性分析的工具——从混杂的数据中寻找规律或关系的线索。

(六) 需要归纳推断法与演绎推理法并用

哲人培根说过"知识就是力量"[6]。统计研究的任务就是为了发现新的知识,归纳法则是发现新知识的基本方法。因此,归纳推断法成为最主要的统计研究方法,使得我们能够从足够多的个体信息中归纳出关于总体的特征。当然,归纳推断的依据通常是样本数据,即在归纳出样本特征的基础上再推断总体。对于大数据,我们依然要从中去发现新的知识,依然要通过具体的个体信息去归纳出一般的总体特征,因此归纳法依然是大数据分析的主要方法。正如 C. R. 劳指出"'从数据中提取一切信息'或者'归纳和揭示'作为统计分析的目的一直没有改变。"[5]但是,大数据是一个信息宝库,光重视一般特征的归纳与概括是不够的,还需要分析研究子类信息乃至个体信息,以及某些特殊的、异常的信息——或许它(们) 代表着一种新生事物或未来的发展方向,还需要通过已掌握的分布特征和相关知识与经验去推理分析其他更多、更具体的规律,去发现更深层次的关联关系,去对某些结论做出判断,这就需要运用演绎推理法。演绎法可以帮助我们充分利用已有的知识去认识更具体、细小的特征,形成更多有用的结论。只要归纳法与演绎法结合得好,我们就既可以从大数据的偶然性中发现必然性,又可以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[7]。

(七) 需要相关分析与因果分析并重

《大数据时代》认为,我们只须从大数据中知道"是什么"就够了,没必要知道"为什么",并且指出"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"[1]。毫无疑问,从超大量数据中发现各种真实存在的相关关系,是人们认识和掌控事物、继而做出预测判断的重要途径,而大数据时代新的分析工具和思路可以让我们发现很多以前难以发现或不曾注意的事物之间的联系,因此大力开展相关分析是大数据时代的重要任务。但是,我们仅仅停留于知道"是什么"是不够的,还必须知道"为什么",正所谓"既要知其然,更要知其所以然",只有这样才能更好地理解"是什么"——为什么需要把手电筒与蛋挞放在一起。只有知道原因、背景的数据才是真正的数据。因此探求"是什么"背后的原因始终是人类探索世界的动力,因果分析是人类永恒的使命。哲学家德谟克利特早就指出"与其做波斯国王,还不如找到一种因果关系。"[8]如果我们只知道相关关系而不知道因果关系,那么数据分析的深度只有一半,一旦出现问题或疑问就无从下手。而如果我们知道了因果关系,就可以更好地利用相关关系,就可以更好地掌握预测未来的主动权,就可以帮助我们更科学地进行决策。当然,因果分析是困难的,正因为困难,所以要以相关分析为基础,要更进一步利用好大数据。相关分析与因果分析不是互相对立的,而是互补的,两者必须并重。

(八) 需要统计技术与云计算技术融合

尽管用于收集和分析数据的统计技术已相对成熟、自成体系,但其所能处理的数据量是有限的,面对不可同日而语的大数据、特别是其中大量的非结构化数据,恐怕单凭一己之力是难以胜任的,只能望"数"兴叹。首先遇到的问题就是计算能力问题,这就要求我们在不断创新与发展统计技术的同时,还要紧紧依靠现代信息技术、特别是云计算技术。云计算技术主要包括虚拟化、分布式处理、云终端、云管理、云安全等技术[9],或者说以编程模型、数据存储、数据管理、虚拟化、云计算平台管理等技术最为关键。借助云计算技术可以将网格计算、分布式计算、并行计算、效用计算、网络存储、虚拟化、负载均

大数据时代对统计学的挑战*

邱 东

内容提要:本文首先探讨了面对大数据潮流应持有的科学态度,然后从大数据能否淹没整个世界、信息与噪声能够泾渭分明吗、统计学与数据科学究竟是什么关系、大数据潮流对统计学究竟产生了什么样的影响等四个方面论述了大数据对统计学的挑战。

关键词:大数据;信息;噪声;数据科学;统计学
中图分类号:C829.2　　文献标识码:A　　文章编号:1002-4565(2014)01-0016-07

The Challenge of Statistics in the Age of Big Data

Qiu Dong

Abstract: This paper discusses the trend to big data which is due from scholars to scientific attitude, and then discusses the challenges of big data from four aspects as following: Can big data cover the whole world? Can Information and noise be quite distinct from each other? What's relationship between statistics and data sciences? What kind of impact generated on the trend of big data?

Key words: Big Data; Information; Noise; Statistics; Data Sciences

一、除了机遇还有挑战

世界潮流,浩浩荡荡,不可阻挡,国人讲究识时务者为俊杰,信息时代,数据爆炸。大数据大势当前,究竟采取什么样的态度才是真正的"识时务"?

大数据时代并不会自动生成,总是需要不断地提出和解决大数据发展所遇到的问题和矛盾,才会有切实的进步。事物发展的不同阶段有不同的"时务",需要不同的应对。

2009 年,大数据成为互联网信息技术行业的流行词汇。而早在 1980 年,著名未来学家 A. 托夫勒出版《第三次浪潮》,其中已将大数据赞颂为"第三次浪潮的华彩乐章"。此间 30 余年,能不能看作大数据发展的萌芽期?多数人对数据爆炸还懵懵懂懂,世界需要赛博世界(Cyber world)的开拓者,需要大数据潮流的预示者,需要导师,需要先声夺人。

一旦人们接受大数据汹涌而来的现实,就需要既讲机遇,也讲挑战。我们固然仍需要启蒙,需要科普,需要科学理论和方法论的"二传手",但不需要跟风,不需要屏蔽了部分信息的"偏息图",不需要抓住一点不及其余的"唯数据论",不需要"应运而生"的投机者。我们更需要切实有学术增加值的数

* 本文为第十七次全国统计科学讨论会特邀论文。

衡等传统计算机技术与现代网络技术融合起来,把多个计算实体整合成一个具有强大计算能力的系统,并借助 SaaS、PaaS、IaaS、MSP 等商业模式把它分布到终端用户手中。云计算的核心理念就是不断提高"云"处理能力来减少用户终端的处理负担,使用户终端简化成一个单纯的输入输出设备,并能按需享受强大的"云"计算处理能力。可见,统计技术与云计算技术的融合是一种优势互补,只有这样统计技术才能在大数据时代一展身手、有所作为,才能真正把统计思想在数据分析中得到体现,实现统计分析研究的目的。

数据创造统计,流量创新分析。由于各个应用领域的不断变化,特别是数据来源与类型的不断变化,使得统计学还难以成为一门真正成熟的科学。因此,在数据分析的世界里,不断提高驾驭数据的能力是统计学发展的终身动力。