DOI:10.12361/2661-3263-05-08-115562

大数据时代统计学发展分析

侯俊文
对外经济贸易大学统计学院，中国·北京　100105

【摘　要】本文研究工作的开展对大数据、统计学的来源及发展历程进行阐述，对大数据时代对统计学带来的影响进行分析。讨论大数据时代背景下统计学发展存在挑战以及策略。希望能够为相关部门及人员工作的开展提供参考和指导思路。

【关键词】大数据；统计学；挑战；发展策略

Statistical Development Analysis in the Era of Big Data

Hou Junwen
School of Statistics, University of International Business and Economics Beijing 100105

[Abstract] The development of this research work on the source and development of big data, statistics, and analysis of the impact of the era of big data on statistics. Discuss the challenges and strategies of statistical development in the era of big data. Hope to provide reference and guidance for relevant departments and personnel to carry out the work.

[Keywords] Big data; statistics; challenges; development strategy

1　大数据时代与发展历程

1.1 大数据时代内涵分析

近些年来，我国科学技术得到了飞速的发展。在此背景下，物联网、云计算等技术有了显著的进步。数据的增长呈惊人状态出现，出现海量的数据意味着大数据时代的到来。在当前背景下，对大数据时代统计学的发展方向以及发展趋势进行研究，有着非常重要的意义。实际上，大数据是高维便利及大样本的数据集合，针对样本问题，通过统计学原理，做分析抽样工作，满足其精度需求。针对部分维度较高的问题，应该通过统计学原理的方式进行压缩、降维、分解。站在另一个角度分析大数据，所涉及的内容具备多样化特点，为诸多领域的数据综合体，其中涵盖人文科学、自然科学。每个学科之间能够相互穿插、融会贯通。部分传统的统计学方法适用于分析单个计算数据。现阶段，大数据背景下使这一现象得以改变。大数据环境中涵盖了磁盘存储环境、数据流环境、多线条环境以及分布存储环境。现阶段，大数据背景下，首要目的是将数据转变成人们能够了解接受懂得的知识，从而对数据的源头及机制进行探索，并制定科学合理的对策，因想要将信息转变成知识，并非一朝一夕就可以完成的，需要漫长的实践过程。当前，信息量逐步增加，在此情况下，常规电脑内存量无法承载新处理数据，新型数据处理技术在此衍生出来。此类技术能够消除僵化层次结构，对数据的排列产生促进性，并使数据处理量得以提升。

1.2 统计学发展历程

人类的统计活动是在技术行为下衍生出来的。正因如此，统计发展历程可以追溯到原始时期。也就是说，直至今日，统计已经有上千年的发展历程。但人类统计时间直至理论层面的发展追溯到近代有300多年的历史。大体来说，统计学发展历程能够划分为古典记录统计学、近代描述统计学和现代推断统计学几个阶段。古典记录统计学的时间范围在17世纪50年代至19世纪50年代。在此阶段，统计学兴起，同时，形成了最初的研究规则及方法。近代描述统计学的时间范围自19世纪50年代直至上世纪20年代。在此阶段，描述特征指的是通过原本生物进化成研究领域。正因如此，历史上称其为生物统计学派。现代推断统计学时间范围在上世纪初至上世纪50年代。在此阶段，统计学有了显著的进步，在历史上称其为农业试验统计学派。在当前时代背景下，对于统计学的发展来说，存在着诸多的挑战和发展机遇。挑战是指当前传统统计学方法无法满足大数据时代发展需求。机遇指的是将统计学作为基础大数据开展数据分析、数据统计等相关工作，能够凸显大数据的可视化特点。

2　大数据时代对统计学的影响

在大数据分析、利用及研究过程中，统计知识具备诸多的应用形式。主要是依赖统计学搜索、分类、整合爆炸增长的信息数据。正因如此，在一定程度上，大数据研究工作的开展所运用统计学知识。但大数据并没有被统计学学科充分利用，导致这一现象的主要原因是大数据的使用模式、运用方式和统计学之间有着非常明显的差异性。统计学主要是对样本统计资源进行充分利用，样本将既定概率标准作为依据，在总体中进行抽样调查。随机抽样调查具备成本属性，例如资本投入及时间消耗等。在增加样本量的情况下，估计范围误差因总体样本量增加而出现变化，这是统计学所具备的明显不足之处。海量的数据信息及电子商务信息是大数据时代背景下具备代表性的内容。从总体角度出发，大数据呈现出总体样本数据化这一趋势，这一特征能够弥补样本统计存在的不足之处。在大数据背景下，整体样本统计包括所有样本容量，但因诸多情况下，数据具备半数据化及非结构性能特征，诸多数据资源表现为重视尾部分布状态、标准差、方差等标准化方法不具备重要意义。不稳定性及整体依赖性超过了时间内的整体假设。正因如此，应用概率论的范围逐渐变得狭窄。对此情况，统计学在对大数据技术进行充分利用，开展样本统计工作过程中，应该选择并融合整体数据资源。

2.1 大数据对样本及总体的影响

大数据具有其特有的多样性、庞大的体量、某些样本难以表达的规律，而大数据可以很好地展现；对于一些小的信息，有些样本很难捕获，而大数据可以很好地覆盖；而大数据，却可以通过一些样本，发现一些不正常的现象。在大数据时代，人类对客观事物和现象的认知能力将大大提高，各种重要信息也将不会被遗漏。也就是说，在大数据的时代，大数据不仅仅是一个范例，更是一个整体。

2.2 相关分析变化

在大数据时代，我们必须提高对相关分析应用的重视，这也是传统相关分析研究的一个难得机会。同时，从相关性分析的角度来看，如何保证相关的统计分析方法能够适应时代发展，在大数据时代发挥其应有的作用，而传统的相关分析也面临着严峻的挑战。

2.3 因果分析的变化

在大数据时代，相关分析的应用越来越受到人们的关注，这对于传统的相关分析来说也是一次难得的机遇。同时，从相关性分析的视角出发，探讨了在大数据时代，如何使相关统计分析方法更好地适应时代的发展，使其在大数据时代更好地发挥其作用。

3 大数据背景下统计学面临机遇与挑战

3.1 大数据背景对统计学挑战

（1）大样本标准变化。样本统计是统计学的重要组成部分，它是以抽样统计方法来研究事物的数量特征以及数量关系等。在大数据时代，样本即整体将向新的发展方向转化，大样本标准也会发生改变。通常，大的抽样是指抽样数量在30以上，而小的抽样数量则是小于30。传统的统计学以30为大样本，由于大量的数据资源噪声、多源、异质性等因素，这一标准明显不够完善，不能有效地过滤掉干扰信息和数据的影响，导致统计结果只能在一定范围内解释客观事物的规律。因此，以往的统计方法应该充分整合大数据时代的海量数据，提高数据源的多样性，提高样本数量；同时，要改变大的抽样标准，用较大的抽样数量取代以往的较少的抽样数量，以达到在大数据时代对数据准确性的需求。

（2）样本选取和形式的再确定。目前，85.0%以上的数据都是半结构化、非结构化的数据，很明显，以往的关系型数据库很难对这些数据进行有效的处理，而大数据可以利用非结构化的数据库，将非结构化的数据转化为结构化的数据，充分利用这些数据的潜力。如果传统的统计方法能够突破结构数据的限制，减少抽样标准，建立无结构的数据库，促进数据的多样性，则可以扩大统计的应用领域。

（3）统计软件需要升级开发。将统计模型、统计软件作为基础，传统的数据处理和分析是可以进行的，而统计软件是处理和分析数据的最好工具（SPSS、 SAS和 DPS），而统计模型是用来处理数据和分析数据的，而统计模型则则在不同的变量之间建立了数据的关系。如果现有的统计软件能够有效地借鉴大数据处理软件，在数据处理和存储功能的基础上，增加数据传输和存储功能，这些软件就会变成数据中心，实现数据的共享，从而有效地推动大数据在统计软件中的应用。对于一个庞大数据来说，原始资料应该有一个统一的标识，否则，以该标识为基础的数据中心将会出现数据无法辨识的问题。另外，数据标记也会在一定程度上增加使用者的操作难度，这也是大数据处理中的一个难题。

（4）实质性统计方法大数据化

在大数据时代，统计所面临的问题不仅在于样本、软件，还有新的统计方法，这一方面在统计方法上的突出体现，如经济统计、卫生统计、生物统计等。在大数据时代，数据是一种资产，它的发展程度还很低，绝大多数都被互联网、搜索引擎、统计机构相关 IT 公司、所垄断[1]。

3.2 大数据时代对统计学机遇

（1）加强统计质量

科学、合理地运用大数据，可以提高统计质量。传统的统计资料存在着频率低、滞后等问题，而大数据的适时性可以弥补这一缺陷，增强统计的时效性。就拿CPI来说，CPI的公布频率是按月计算的，但往往都是延迟的，比如，我们国家是在每月的9号公布一个月的 CPI，然后动态收集和汇总市场价格，提供第一手的数据，同时也能提高发布的频率，从而有效地分析市场的通胀规律。

（2）统计成本下降

随着统计成本的降低，大数据可以被频繁地重复使用，它所收集到的数据不再是单一的功能，而是可以满足不同的需要[2]。对采集数据应用的次数逐步增多，数据所具备的潜在价值被更全面的挖掘，而采集数据所产生的成本并不会受数据应用的次数所影响，故各式各样用途的平均统计成本将得到显著地降低。以谷歌为例，它可以通过分析用户获取的信息来评价流感的蔓延，但是这只是大量的数据中的一部分。从这一点可以看出，随着数据的使用越来越频繁，统计费用也在不断降低[3]。

（3）统计学可发挥作用范围扩大

在大数据时代，很多数据都是从一些以前认为不能被数据化的产业中提取出来的，就像是从网络上的用户的搜索记录中提取出了相关的健康数据，比如消费者的消费偏好数据。可以从社会网络使用者的社交网络中获取相关的信用数据、财产状况等信息。若是能够收集到足够的数据，那么他就可以利用统计数据进行分析。从这一点可以看出，在大数据时代，统计学所能利用的领域有了很大的拓展[4]。

（4）延伸统计学科体系

统计在将大数据引入统计后，可以分为总体统计和样本统计两种类型，其中样本统计主要是指大数据中的全部样本特征，也就是总体统计是以大数据为基础的统计；而后者则是以随机现象和大量现象数据为研究对象，这就意味着，抽样统计是以数理统计和概率论为基础的传统数据统计。将统计的总体统计和样本统计结合起来，可以有效地解决统计中的统计问题，解决样本统计中的数据收集问题，从而有效地扩展统计学科的体系[5]。

4 结束语

总之，大数据对传统的统计工作提出了挑战，同时也为传统统计的高效发展提供了一个很好的机会。在大数据时代的发展趋势下，我们应该充分意识到，作为传统统计学的一种补充，而非更替。基于样本统计和预测分析内容的传统统计学，仍然在社会统计和经济分析中占有重要地位。

参考文献：

[1]李智明.浅谈大数据时代统计学的挑战与机遇[J].教育教学论坛,2020(13):95-96.

[2]袁明.统计学在大数据时代的应用[J].财富时代,2019(12):242.

[3]杨宗霖.大数据时代对统计学的挑战[J].今日财富,2019(17):219-220.

[4]唐勇.浅谈大数据时代对传统统计学变革的思考[J].大众投资指南,2019(17):262.

[5]潘玥.大数据时代下统计学的挑战与发展[J].科技风,2018(03):26.

作者简介：

侯俊文，（1995-），男，汉，山西晋城人，大学本科，研究方向：统计学。对外经济贸易大学统计学院在职人员高级课程研修班学员。