大数据还需要抽样吗

程开明　宋艺旋

大数据正多方位改变着人们的认识、思维和行为，维克托·迈尔·舍恩伯格提出大数据时代的三种思维变革：要全体不要抽样，要效率不要绝对精确，要相关不要因果。这一观点在社会上产生广泛影响，诸多学者和大众都表示赞同，认为以往因无法获取总体所以需要抽样，大数据时代“样本 = 总体”，要分析与某事物相关的所有数据，而不是依靠少量样本数据，毕竟抽样就意味着有抽样误差的存在。如此说来，大数据时代真的不再需要抽样了吗？通过对现有观点的梳理和理性思考发现，事实上大数据与抽样并非相互排斥，一定程度上存在着互补关系，大数据条件下合理利用抽样技术能够更有效地开展数据收集和分析，更好地释放大数据的能量，挖掘大数据的价值。

变革：从抽样到大数据

经典统计学观点认为数据分析是通过局部样本开展统计推断，以了解总体的规律性。在收集数据和分析数据能力受限的时代，为有效认识总体特征，抽样应运而生，产生了各式各样的抽样技术。抽样可分为随机抽样和非随机抽样，随机抽样又包括简单抽样、系统抽样、分层抽样、整群抽样和多阶段抽样等，非随机抽样包括拦截、指定、滚雪球等抽样形式。一般抽样调查是指随机抽样，即按照随机原则从总体中抽取部分单位进行观察，总体中各单位都有一个指定的概率被抽取，然后以样本统计量来推断总体参数，并可以计算和控制抽样误差的大小。

抽样调查具有经济性、时效性、准确性和灵活性等特征，能够解决全面调查无法或难以解决的问题，补充、订正全面调查的结果，对总体假设进行检验。虽然具有众多优势和作用，但抽样是在事先设定目的前提下展开工作，不管采用多完美的抽样技术，抽到的只是总体的一部分，难以完全准确地代表总体。大数据时代的抽样也面临一些实际困难（朱建平，2014）：（1）抽样框不稳定，随机取样困难。随着网络信息技术的迅速发展，人们获取信息的途径越来越便捷，更换工作、外出学习和旅游的机会和次数增多，人口流动性加快；另外企业经营状况变动较快，一些企业规模

日益壮大，有些企业破产倒闭，抽样框快速变动导致随机取样较为困难，抽样结果的精确性大打折扣；（2）事先设定调查目的，会限制调查的内容和范围。抽样调查往往是先确定调查目的，限于特定调查目的，调查范围往往受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，抽样结果不适合细分。抽样调查是在特定目的和一定经费控制下进行的，样本量限制使得对调查内容进行细分的结果通常因样本量太小而不具代表性。（4）纠偏成本高，可塑性弱。抽样过程中一旦抽样框出现偏误，想了解与事先调查目的不一致或目标总体的细分结果，往往需要重新设计调查方案，成本较高。

随着信息技术、网络技术取得巨大突破，体量大、类型多、结构复杂的海量数据扑面而来，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。随着人类社会行为记录的传感器化、传感器的网络化、网络的数据化，通过传感器、网络，汇集到存储设备，每时每刻都以某种形态记录着人类行为，存储于记录设备中，汇集成为大数据。大数据主要包括三大领域：一是社会网络数据，譬如微信、微博、Facebook 等形成的数据；二是人机交互数据，譬如网购、网游、网娱、工作、交通、医疗等人与机器交互形成的大量记录数据；三是狭义的传感器及机器数据，譬如 GPS 数据、智能电表、智能交通、计算与实验数据等。大数据的数据采集有别于传统的抽样调查，广泛使用的传感设备、信号识别技术，编译和可扩展的存储系统等，使数据收集的时效性增强，存储和积累更为方便快捷。

大数据处理技术能够对超大规模的数据进行分析处理，对研究对象的特征既能做到总体把握，又能了解局部情况，使得大数据成为普受关注的数据采集方式。作为一种新的数据来源渠道，大数据的确会对一些传统抽样数据产生替代作用，譬如在居民收入支出大都是通过银行转账的地区，银行交易数据即可一定程度上替代居民家计调查数据。大数据条件下，数据收集及处理方式发生重大变革，数据收集将更多地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 (李金昌 , 2014)。

| 互补：大数据仍需抽样

进入大数据时代，能对全体数据进行分析，不一定要抽取样本，但这并不意味着抽样就要退出历史舞台。当所获取的“大数据”总体不能完全代表目标总体时，总有一部分个体被遗漏在外，此时大数据分析可能得到有偏的结果。首先，目前来看并非所有数据都可以通过网络信息系统获得，因为并不是所有产业都已实现智能化，还有很多数据只能通过传统的抽样方式获得；其次，即使是网络数据，某些情况下对总体进行分析也并非最优选择，例如当面临均匀度很大的总体时，随机抽取部分单位作为样本开展分析就可以得到理想结果，此时并不需要去费时费力分析总体。

大数据如浪潮般涌来，带来的既有信息也有噪声，通常还是噪声居多，使得数据分析很容易为假像所迷惑，造成规律的丧失和失真。大数据的生成与采集在人为设计的框架之下，也可能存在系统性偏差，譬如社交网络数据中人群的上网行为习惯、计算机知识、经济地位等都是左右数据生成的混杂因素（Crawford， 2013），导致大数据与真实总体之间可能存在明显偏差。所以，大数据条件下随着众多缺失、含有噪声甚至是错误的数据进入到数据库中，此时从中抽取部分样本数据更能有效地进行数据清洗，以挖掘出数据背后的真规律。耿直教授（2014）认为利用随机抽样数据可以矫正杂乱、非标准的数据源，将经过严格抽样设计获得的统计机构数据作为标准和框架能够对互联网数据进行矫正，而将互联网数据作为补充资源能够对统计机构数据进行实时更新。

总体来看，某些场合下大数据还不能完全代表总体，抽样仍然必要，但抽样环境已发生显著变化。大数据领域一般依靠高性能计算机采用分布式系统处理数据，面对大数据环境下高速网络中瞬息之间涌入的海量数据流，部分情况下计算机无法将信息完全存储下来并及时分析，此时一种合理的策略便是基于抽样建立起能够进行事后分析的汇总信息以保存核心数据。另外，从计算成本、便捷性角度考虑，抽样相比于全数据处理往往也是更优选择，因此即使在有能力处理全数据的计算环境下，对抽样依然存在着巨大需求。虽然抽样受条件、时间、资源等诸多因素限制，然而通过合理设计，它在大数据领域仍然能够发挥重要价值，甚至起到与大数据相互印证的作用。

由于大数据来源与种类的多样性，以及数据增加的快速性，在享受数据丰富性的同时也往往面临一些困境：存储能力够不够，分析能力强不强，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？解决这些困境通常需要对数据进行分类、筛选，有针对性地删除那些垃圾数据、不重要或次重要的数据。如果说以前有针对性地获取数据叫做收集，那么今后有选择性地删除数据就意味着收集。大数据时代的数据收集将更多的是从已有超大量数据中进行再过滤、再选择（李金昌，2014），这种再过滤、再选择其实是抽样的过程。

顺应大数据环境，抽样也要积极应对所面临的挑战：（1）大数据的大体量、非结构，且来源复杂，使得目标总体的抽样框难以构造，而没有抽样框也就难以计算样本单元的入样概率，抽出的样本多属非概率样本，难以应用传统的抽样推断理论，如何利用非概率样本对总体进行推断需要进一步探讨（金勇进，2016）。（2）针对大数据可能存在的系统性偏差，如何把获取的抽样数据作为大数据分析的对照基础与验证依据，以改善大数据的偏倚性，还需寻求有效途径。

选择：大数据抑或抽样

大数据时代，争论是否需要抽样意义并不大，关键是要清楚“何时抽样、何时利用全体数据集”。分布式和实时处理技术的发展，让大规模数据分析成为可能，如果所获取的“大数据”是研究问题的合适总体，除非特殊情况下需要开展抽样，一般应该利用整个数据集开展分析，毕竟数据量越大反映的信息越多。但是，从效率和成本角度考虑，有时适当的抽样还是必要的。譬如一些商业领域的大数据处理耗费大量资源和时间，等得到结果后整个行业环境可能已发生翻天覆地的变化，这种情况下针对特定目的的抽样便有必要。当然，有些数理处理及分析问题无法通过抽样来降低处理的复杂程度，就必须利用一些专门为处理大数据而设计的存储、计算和分析技术来实现。

针对不同类别的问题，如何在大数据和抽样之间进行选择，可借助图1来加以说明。

图1中面临一个有确定目标函数的数据处理问题，A、B、C三条曲线分别代表三类问题（刘鹏和王超，2015）。

（1）A类问题。如果通过抽样能够显著降低数据处理的复杂程度，同时解决问题的效果（目标函数）没有太大的下降，那么应该采用抽样的方式来开展数据处理，此类问题可用图中的曲线A示意。由于通过较低的采样率就能解决问题，并不需要大规模分布式的计算架构，这类问题可归为传统数据处理问题，而非大数据问题。

（2）C类问题。一些数据问题基本上不可能通过处理一小部分数据来达到处理全体数据所能达到的效果，即随着采样率的降低，解决问题的效果会快速下降，此类问题是典型的大数据问题，用图中的曲线C示意。由于处理大规模的全体数据，传统存储和计算架构不再合适，需寻求新的大数据解决方案。

（3）B类问题。实践中的大数据和抽样之间可能并不泾渭分明，某些问题的处理效果随着数据量上升有一定的提升，但当数据大到一定规模后再增加数据量带来的效果提升并不明显，这类问题可用图中的曲线B来示意。解决此类问题，往往是选取一个有较大规模但并非全体的数据集来处理。

实际当中采取大数据还是抽样的方式，还取决于研究问题的性质。从概率论的角度可将事件分为大概率

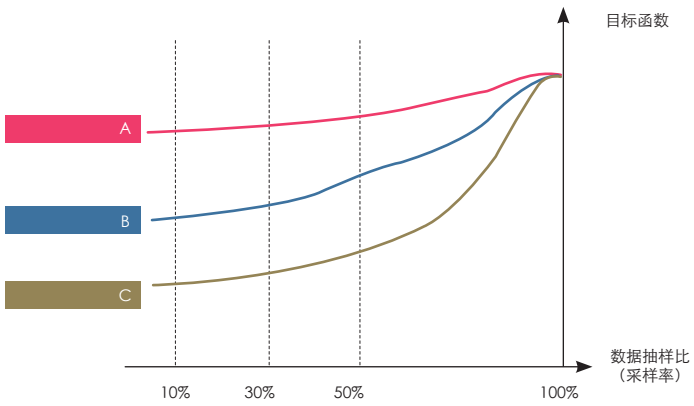

图　大数据与抽样的选择示意图

事件和小概率事件，前者通常选择抽样的方式进行处理，如新政策的支持率、售后服务的满意度等；后者应使用尽可能多的数据以发现这些“罕见”事件，如信用卡诈骗、非法操纵比赛结果问题，只有掌握了全体数据才能找出异常情况。如果所研究的问题以推断总体为目标，而检测所有个体实际上不可能或者会破坏样本，比如测试灯泡、汽车碰撞试验等，此时抽样是较好的选择。若所研究的问题以个体特征分析和应用为目的，则直接利用所获得的全体数据集进行分析更合适，如个性化推荐、精准营销等。还有一些问题本身不能采用抽样的方式来解决，必须采用全体数据，譬如排名或 top10 问题、异常值处理等。

| 策略：大数据与抽样的融合

获取数据是开展分析的前提，拥有大数据就拥有了超大量可选择的数据，接下来考虑的重点应该是如何充分利用大数据，凡是从大数据源中能够找到的数据就不再需要专门的调查。当然，在信息化、数字化、物联网等还不能全覆盖的情况下，仍然有很多数据需要通过抽样的方式去获取。与此同时，很多情况下尽管可对大数据进行全体分析，但考虑到成本与效率因素，抽样仍然是明智的选择。因此，大数据时代既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，又要采用传统的抽样方式去收集与处理特定需要的数据（李金昌， 2014）。

1. 充分利用大数据，打造政府统计“第二轨”。大数据为扩展统计资源获取空间、丰富统计产品功能、提升统计信息价值提供了共享平台，深度开发利用大数据，将其打造成为统计基础数据的“第二轨”势在必行。大数据促使统计思维的大转变，将无所不在的网络数据、物联网数据、行政管理记录、企业销售记录、搜索记录等海量数据逐步纳入统计范畴，融入到政府统计工作中来，有利于推动统计部门提升数据分析能力、提高统计服务水平，拓宽统计服务范围，更好地服务于经济社会发展。

2. 切实发挥抽样功能，与大数据相得益彰。在大数据时代，抽样仍然发挥着至关重要的作用，但体量更庞大、类型更繁多、结构更复杂的数据对抽样的要求也更高，抽样需要转变功能以顺应时代要求。一是统计机构通过抽样调查所获取的数据具有权威性，可作为对照基础和验证依据与大数据进行对比，开展抽样数据对互联网数据的校正与调整。二是把抽样数据看作从混杂的大数据中寻找规律或关系的线索，以便更好地进行大数据挖掘和快速探测分析。大数据应与抽样相结合，以实现高质量的数据收集、处理及分析。

3. 创新抽样方法，适应大数据环境的需要。针对大数据流环境，需要探索如何抽取足满足研究目的和精度要求的样本，探求新的适应性、序贯性和动态的抽样方法，探究大数据的案例抽样方法和基于事件的抽样方法、社会关系网络和图的抽样方法等（耿直， 2014）。大数据背景下的抽样很难构造清晰的抽样框，抽取的样本多为非概率样本，这就需要解决非概率抽样的统计推断问题（金勇进，2016）。另外，如何根据已获取的样本逐步调整感兴趣的调查项目和抽样对象，使得最近频繁出现的热门数据进入样本，也需要方法上的创新。

大数据时代，既要全体又要抽样，有时需分析与事物相关的所有数据，因为大数据更为全面；有时则应分析少量的样本数据，因为抽样更具效率。未来，大数据与抽样将相互补充，携手前行。

作者单位：浙江工商大学统计与统计学院

参考文献

[1]Crawford K. The hidden biases in big data[J]. HBR Blog Network, 2013(1):9–10.

[2]“大数据中的统计方法”课题组．大数据时代统计学发展的若干问题[J]. 统计研究， 2007(1):5–11.

[3]耿直．大数据时代统计学面临的机遇与挑战[J]. 统计研究，2014(1):6–9.

[4]金勇进，刘展．大数据背景下非概率抽样的统计推断问题[J]. 统计研究， 2016(3):12–17.

[5]李金昌． 大数据与统计新思维[J]. 统计研究， 2014(1):11–17.

[6]刘鹏， 王超． 计算广告：互联网商业变现的市场与技术[M]. 北京： 人民邮电出版社，2015.

[7]维克托·迈尔·舍恩伯格，肯尼思·库克耶． 大数据时代——生活、工作与思维的大变革[M]. 杭州：浙江人民出版社，2013.

[8]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):11–17.

[9]朱建平，张悦涵． 大数据时代对传统统计学变革的思考[J]. 统计研究， 2016(2):4–9.