“人工智能其实就是统计学”这个命题并不重要

朱建平

【摘　要】大数据时代，在AI迅猛发展的催促下，我们统计工作者应该清醒地认识到传统统计学的变革，以便更好地“武装”统计学，真正的起到长期推进AI发展的作用。本文抛砖引玉，从微观的角度剖析了传统统计学的变革，同时引发了思考。

【关键词】人工智能；大数据；统计学；资源共享

最近，我在网上收集并阅读了好多关于人工智能方面的文章，有一个热议的话题“人工智能其实就是统计学”。2011年诺贝尔经济学奖获得者ThomasJ.Sargent在由厚益控股和《财经》杂志联合主办主题为“共享全球智慧　引领未来科技”的世界科技创新论坛上表示：人工智能其实就是统计学，只不过用了一个很华丽的辞藻，其实就是统计学。好多的公式都非常老，但是所有的人工智能利用的都是统计学来解决问题。

人工智能（Artificial Intelligence），英文缩写为AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

目前，自然而然地“统计学与AI的关系”议题，就成为了讨论的焦点问题，一下子又把统计学推到了峰顶浪尖。实际上，AI是否属于统计学科，这个问题并不重要，我们也不要谈“人工智能其实就是统计学”。重要的是，在大数据时代下，在AI迅猛发展的催促下，我们统计工作者应该清醒地认识到传统统计学的变革，以便更好地“武装”统计学，真正的起到长期推进AI发展的作用。在此抛砖引玉，从微观的角度谈谈传统统计学的变革，给我们引起的思考。

怎样用新的数据烹饪出好的“食品”

经常和朋友聊起来，说时代发生了巨大的变化，火车站、机场，包括出差酒店的住宿，到处都在刷脸，现在这个技术已经很成熟了。最近到北京出差，在机场安检的过程中，新开辟了一条智能化安检线，它会将安检要求不满足的物品，自动地分离出来，提高了安检的速度，当时我很好奇，在那里看了好久。这些技术的实现及应用，所产生的数据，其类型发生了巨大的变化，扩展了传统统计的研究对象。

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。大数据及AI则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（包括图像和声音等）、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。现在的数据库很多都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。

在数据分析中，就好似食品烹饪过程的体现，数据就相当是食材，食材的类型或品种发生了变化，能否烹饪出好的食品，是对烹饪师一个巨大的挑战。对于统计工作者而言，摆在面前的压力可想而知。

| 怎样认识总体意义下的"样本"

有一次，朋友到我们单位访问，朋友问：你们单位的 wifi 和密码是什么？想用手机上网；我"开玩笑"地和他说："你如果用我们单位的 wifi 和密码上网，我们单位的网络平台可以把你手机的信息全都扒走"。这就是维克托·迈尔·舍恩伯格写的那本书《大数据时代——生活、工作与思维的大变革》中提到的一个重要观念：在一定的条件下，我们现在所获得的数据是总体，而不是样本。在八年或十年前，这个技术感觉很神秘。当今在 AI 促动下，这个技术是海量数据收集的重要手段之一，其打破了我们对样本概念的认知，使得样本概念更加深化，体现出了一个重要的理念，就是"明确平台，收集数据"，这充分体现了总体意义下的"样本"含义。

我们知道，统计学依赖于样本统计（普查除外），样本是按照一定的概率从总体中抽取并作为总体代表的集合体。大数据时代，特别是 AI 技术应用，使得样本的概念不再这么简单，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，呈现"总体即样本"的趋势，这一特点弥补了传统样本统计高成本、高误差的劣势；二是动态数据，比如在确定的网络平台下，数据是随着时间的推移而变化的，其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"与有别于传统的样本，因为这些数据并非局限于随机抽取的，更可以是选定的与分析目的相关的数据。对于统计工作者而言，如何分析此"样本"的代表性等问题，应该提到统计研究的议事日程中。

| 怎样从数据收集实现"资源共享"

从 2012 年到现在，国家统计局和 19 企业签署了框架性协议，在框架性协议的支撑下，国家统计局可以获得企业的数据，一为社会服务；二为企业服务。例如可以利用阿里巴巴的数据来修正和完善 CPI；可以利用百度的数据来预测二手房的房间问题。以此为案例，我们把国家统计局获取数据的外延剔除掉（和那些企业签署框架性协议，暂时不考虑），提取获取数据的内涵，可以提炼出一个获取数据的重要手段，就是"框架性协议支撑"。这是因为现在政府和政府之间、政府内部的职能部门之间、政府和企业之间的数据不能交易，那么构建智慧城市等，需要构建宏观和微观大数据平台，"框架性协议"就

成了获取数据的重要方法之一，其核心有四个字"资源共享"。这是在共享经济环境下，对传统数据收集概念的巨大扩展。

传统统计中，收集统计数据的思维是先确定统计分析研究的目的，然后根据需要收集数据，所以要精心设计调查方案，严格执行每个流程，往往投入大，而得到的数据量有限。在大数据时代，AI 在推进社会前进过程中，给数据的收集提出了新的挑战，使得收集数据的概念得到扩展，即收集数据就是识别、整理、提炼、汲取、分配和存储元数据的过程，其某个环境的实现，都是对传统统计数据收集研究带来了机遇。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，并不是任何数据都可以从现有的数据中获得，还存在安全性、成本性、针对性的问题。对于统计工作者而言，在采用传统的方式方法去收集特定需要的数据基础上，又如何扩展思路，利用现代观念、现代技术收集、获取一切相关的数据，同时怎样实现资源共享，也是统计工作者予以实现的目标之一。

| 怎样打造和利用数据来源的“第二轨”

最近在指导MBA学员毕业论文，有一个学员从事游戏软件开发，并负责公司的游戏产品营销，他毕业论文的主要研究内容是游戏产业的全球市场营销策略，其中游戏产品的定价研究就显得尤为重要，为了科学地制定游戏产品价格，针对开发的游戏产品，收集了大量的不同竞争产品的同等道具价格，例如：木材、粮食、铁矿、石矿、金币等等。在讨论文章的过程中，我问：“这些收集是通过调查得到的吗？”，他说：“不是，是通过互联网得到的数据，是开发商和使用者在游戏研发和玩游戏过程中记录下来的数据”。这些数据完全打破了传统统计数据的来源渠道，对数据的考察和验证带来的极大的挑战。

传统的数据是带着问题来收集，因为具有很强的针对性，因此数据的提供者大多是确定的，其身份特征是可识别的，有的还可以进行事后核对和验证。而随着AI技术的深入发展，海量数据的来源则很难追溯，由于这些数据通常来源于互联网、物联网，或者在云架构的支撑下，已经形成了极大的数据库，这些数据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。对于统计工作者而言，在充分利用好传统统计数据来源的基础上，针对社会发展的需求和时代发展的特点，要努力打造并利用好数据来源的“第二轨”——云架构、互联网和物联网等。

| 怎样提升和加快统计量化方式的转变

最近几年，接触了好多公司老板，他们的管理理念各有千秋。有一个老板管理很苛刻，要求职工上下班按指纹，他说：“发现有个别职工很淘气，上班按了指纹就出去，到下班的时候回来再按一个指纹，结果搞一个全勤”；

我说：“那你怎么办啊！”；

他说：“我在公司门口安装了一个监视器，可以时时记录职工的进出情况”；

他接着又说：“慢慢的发现，监视器获得的数据是连续数据，很难发现异常现象”；

我又问他：“那你采取什么方案呀！”；

他说：“按了一个人脸识别器，这样可以获得时点数据，很容易得到异常现象，以便及时处理”。

这里我想说的是，从指纹识别、监视器到人脸识别器，管理中所用的设备逐渐提升，就人脸识别而言，识别就是要找出每个人的差异性，从统计角度看，分析差异性的有一个重要指标，就是方差。那么识别结果，每一个人的“脸”产生的数据集，其方差如何计算呢？这打破了传统的统计计算规范。

传统数据为结构化数据，其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。在AI研发、延伸和扩展的同时，面临着大量的非结构化数据，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，计算机学界着手研发处理非结构化数据的技术逐步推进，也取得了不少阶段性成果。对于统计工作者而言，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，也势必促进统计学的发展。

AI的发展离不开统计学，统计理论和方法的深入研究也离不开AI的促进。统计工作者应该知道，当今我们对数据的利用，取得了更大的主动权，我们要把这个主动权真正的利用好，用在促使统计学迅速发展方面，让统计学产生的新理论、新方法在历史发展中形成更深刻的烙印。

作者单位：厦门大学管理学院MBA中心
厦门大学数据挖掘研究中心