
处在十字路口的统计学： 人工智能时代的挑战与机遇

蔡天文

编者按：本文源于国际数理统计学会 (IMS) 主席蔡天文 (Tony Cai) 于 2025 年 8 月在美国田纳西州纳什维尔 (Nashville) 的 2025 年联合统计会议 (JSM 2025) 的主席致辞。翻译自 IMS Bulletin (Volume 54, Issue 8, 2025) 并略作修改。



引言

统计学领域历经逾一个世纪的奠基性成就,如今正处于关键的转折点。人工智能(AI)的迅速崛起、复杂与高维数据的爆炸式增长,以及来自科学界、产业界和社会各界日益增长的需求,不仅重塑了我们开展研究的方式,更重新定义了统计学家的角色。这些变革绝非单纯的技术层面调整——它们对我们的学科定位、研究方法,以及在飞速演进的科学生态体系中的立足之地,都提出了深刻的挑战。

当今数据呈现出前所未有的多样性,既包括结构化与非结构化数据,也涵盖数值、文本、图像与音频等多种形态。这一趋势反映了数据科学研究边界的持续扩展。尤为值得关注的是,人工智能本身正在成为数据的重要来源。以 ChatGPT、Claude、Gemini、DeepSeek、Sora 和 DALL·E 等生成式模型为代表的技术,正在大规模生产合成内容。最新预测显示,人工智能生成的数据最早或将于 2026 年超越人类生成的数据规模,到 2030 年更将占据全球数据总量的 80% 以上。

这一转变提出了一系列根本性问题:当数据由算法生成时,我们应如何分析、验证与解释?如何评估其可靠性与科学有效性?这些问题是统计推理的核心,而我们的领域必须在解答这些问题中发挥关键作用。

这是一个机遇与责任并存的时代。随着数据与算法以前所未有的速度演进,我们必须清醒地认识到:统计学是否能够在人工智能与数据科学的新时代中继续发挥引领作用,取决于我们是否主动调整、深化合作,并持续强调统计思维在科学决策中的基础性地位。

统计学在人工智能中的核心地位

尽管机器学习和人工智能近年来发展迅猛,其背后的许多核心算法本质上仍源自统计学。在广为应用的“机器学习算法”中,如线性回归、逻辑回归、决策树、朴素贝叶斯分类器以及支持向量机,均植根于经典统计理论。这些方法之所以经久不衰,正是因为它们建立在估计、不确定性量化与假设检验等基本统计思想之上。它们的持续成功清楚地表明:统计学并非人工智能的附属,而是其理论与方法的核心支柱。

统计学真正的独特之处,在于其对推断、可解释性、可重复性和不确定性的重视——随着人工智能系统的复杂性和应用范围不断扩大,这些原则变得愈发关键。现代人工智能往往倾向于深度学习和“黑箱”模型,而统计视角能让我们提出关键问题:模型是否具有泛化能力?其预测的可靠性如何?驱动模型决策的因素是什么?正则化、模型选择、偏差-方差权衡以及预测不确定性量化等概念,均源于统计学,它们至今仍是构建稳健、可解释人工智能的重要工具。

随着数据分析不断深入医疗、司法、金融、教育及科学发现等高风险领域,对隐私保护、责任追溯、公平性与透明度的要求愈发迫切。统计学家具备独特优势,能够在理论与方法层面确保这些系统在科学上有效、在伦理上可接受,并经得起严格检验。通过将统计严谨性系统性地引入人工智能的研发与应用,我们有助于推动数据驱动系统在效率之外,实现可信、公正与可持续发展。

人工智能对科研生态的重塑

人工智能正在深刻改变科学发现的组织方式与运行机制。一个具有标志性的例子是 2024 年诺贝尔化学奖部分授予计算机科学家、DeepMind 联合创始人兼首席执行官 Demis Hassabis,以表彰其在 AlphaFold 系统上的突破性贡献。AlphaFold 通过人工智能方法实现了复杂蛋白质结构的高精度预测,极大推动了分子生物学与生物医学研究的发展。同年,诺贝尔物理学奖授予被誉为“人工智能教父”的 Geoffrey Hinton,以肯定其在人工神经网络领域的奠基性工作。

这些进展标志着科学进入了一个新时代——在这个时代,人工智能不再仅仅是辅助工具,而是推动发现的核心力量。统计思维强调推断、模型验证、不确定性量化和可重复性,对于确保人工智能生成的见解可信、可解释且具有科学意义至关重要。随着人工智能系统融入研究流程,统计学家必须引领开发一套有原则的框架,对这些系统进行评估和指导。我们的工具不仅需要衡量性能,还需诊断模型失效的原因、纠正偏差,并确保结论在不同条件下依然成立。

此外,随着科学研究日益数据密集化和算法驱动化,统计学的角色必须进一步拓展。我们必须倡导在人工智能辅助科学中采用严谨的标准、透明的方法和伦理保障措施。统计学家肩负着至关重要的责任——不仅要解读数据,还要在人工智能时代塑造知识的创造、评估和共享方式。

理论与实践的张力

理论严谨性与实际应用之间始终存在着一种持续且日益凸显的张力。长期以来,理论发展一直是统计学的标志性优势——它确保了推断的可靠性、不确定性的校准性和模型的透明度——但当前环境越来越重视实证性能、可扩展性和实际影响。这种转变在学术出版和模型实际部署中均有所体现,在这些场景中,成功与否往往仅通过基准测试和预测准确性来衡量。

然而,在这样的背景下,统计理论不应退居幕后。相反,它必须不断演进,并通过应对现代机器学习带来的紧迫而复杂的挑战,彰显自身的相关性。这些挑战包括:理解过参数化模型中的泛化能力、量化深度学习中的不确定性、设计兼具隐私保障的公平且稳健的算法,以及描述在数据和资源约束下的学习极限。

我们不应将理论仅仅视为学术练习,而应将其作为构建可靠、可解释且可重复模型的实践基础。完善的理论框架能帮助我们理解模型失效的时间和原因,并确保从数据中得出的见解具有意义、稳定性和可信度——这些特质在医疗保健、金融和公共政策等高风险领域至关重要。

归根结底,统计学的未来取决于一个充满活力且能灵活应变的理论基础——这个基础必须与应用紧密相连,并植根于现实世界的复杂性。我们需要与实践相结合的理论,也需要能为理论提供启发的实践。弥合这一鸿沟,对于我们领域的持续发展和保持相关性至关重要。

统计研究的新前沿

人工智能时代为统计学研究开辟了前所未有的空间。随着数据在规模、复杂性与多样性上的持续增长,统计学家在应对建模、推断和决策过程中出现的根本性挑战方面,处于独特的优势地位。这些新兴前沿并非经典主题的简单延伸——它们要求我们结合现代计算、社会和伦理背景,重新思考核心统计原则。此类前沿领域包括:

■ 人机混合数据

随着人工智能系统越来越多地生成内容并与人类互动,许多数据集如今都是人类和机器输出的混合体。例如,大型语言模型生成的回应会经过用户进一步整理,形成复杂的反馈循环。统计学家必须开发有原则的方法,对这类数据进行建模、分析和验证,同时考虑算法偏差和行为偏差。

■ 差分隐私与负责任的数据共享

医疗保健、金融、社会科学和政府数据平台等领域,对隐私保护型统计分析的需求日益增长。差分隐私提供了严谨的隐私保障,但它的实际部署——尤其是在联邦学习和流数据环境中——在最优噪声校准、隐私-效用权衡和不确定性量化等方面仍存在尚未解决的问题。

■ 联邦学习与分布式学习

数据越来越多地在分布式设备或机构网络中收集和存储。统计学家可以引领开发通信高效、隐私保护且能应对数据异质性的联邦学习算法。这些进展在健康研究等领域至关重要,因为隐私法规限制了数据的完全共享。

■ 迁移学习与多源整合

迁移学习和领域自适应旨在整合来自不同数据源的见解。这些工具在现实应用中至关重要——例如医学、生态学和经济学领域——在这些领域中,目标群体与训练数据往往存在差异。统计学家有能力为迁移学习的适用场景、实现方式以及异质性环境中的不确定性量化提供理论保障。

■ 深度学习中的不确定性量化

许多深度学习系统缺乏校准后的不确定性评估,这限制了它们在安全关键型领域的可信度。统计学家有越来越多的机会搭建统计建模与深度学习之间的桥梁,开发更可靠的不确定性量化方法、预测区间和诊断工具。

■ 稳健性、可解释性与公平性

人工智能的社会部署要求模型能够应对分布偏移、对抗性输入和数据缺陷。同时,用户、政策制定者和监管机构对模型的透明度和可解释性提出了越来越高的要求。统计方法——尤其是来自因果推断、稀疏建模和敏感性分析的方法——对于推进稳健且可解释的机器学习至关重要。

■ 自动化发现与科学推理

人工智能正不仅成为预测工具,更成为假设生成、实验设计和知识发现过程中的合作伙伴。随着这些系统在科学领域的部署,统计学家必须协助设计框架,确保科学严谨性、可重复性和推断有效性。

这些前沿领域及其他新兴主题共同重新定义了统计学在现代数据生态系统中的角色。如今,统计学家已成为整个数据生命周期中不可或缺的一部分——从设计和生成,到推断和产生影响。

统计学界与国际数理统计学会(IMS)的使命

国际数理统计学会(IMS)在引领全球统计学界应对这一深刻变革方面具有独特地位。IMS 约 45% 的成员位于北美,25% 位于亚洲,23% 位于欧洲。IMS 既有广泛的影响力,也有责任在促进科学卓越的同时,搭建跨越地理、学科和代际界限的桥梁。

鉴于这一全球使命,IMS 刚刚启动了一项新的国际新研究者会议(INRC),该会议将在欧洲和亚洲交替举行。这一举措是在北美新研究者会议(NRC)长期成功的基础上推出的——北美新研究者会议(NRC)在支持早期职业统计学家、提高研究知名度和促进专业网络建设方面发挥了关键作用。将会议拓展至欧洲和亚洲,既认可了国际统计界日益增长的实力,也体现了IMS致力于在全球范围内提供公平获取资源、指导和合作机会的承诺。

为补充这一举措,IMS 还推出了一项新的年度会议系列——统计机器学习前沿会议(FSML)。该会议的时间安排具有战略意义,将在新研究者会议(NRC)之后、联合统计会议(JSM)之前,于当地一所大学举行,成为连接早期职业发展与更广泛统计和数据科学界的重要纽带。其目标是促进统计学与机器学习之间的有意义互动,为研究人员提供一个充满活力的论坛,探索理论、方法与实际应用交叉领域的最新进展。

这些举措共同凸显了 IMS 对培养下一代统计领袖和深化全球参与的决心。通过为早期职业研究人员创造专属空间, 并鼓励跨学科合作, IMS 加强了其国际影响力, 并将各种不同的声音纳入专业社区的核心。凭借坚实的基础和全球范围内的热情支持, 这些新举措有望复制——甚至超越——其北美同类会议的成功。

除上述举措外, IMS 在塑造我们学科的未来方面还发挥着至关重要的作用。该机构可以:

- 促进有意义的跨学科合作——尤其是与计算机科学、工程学、生命科学和社会科学的合作——确保统计思维为人工智能、数据科学和新兴研究领域提供指导。
- 通过指导项目、培训研讨会、差旅资助和高可见度的出版机会, 加强对早期职业统计学家的支持。
- 引领教育改革, 鼓励课程设置中纳入算法思维、计算工具、有效沟通和应用问题解决能力, 同时兼顾数学和统计理论。
- 通过在 IMS 期刊、会议和推广活动中突出有影响力的研究成果, 彰显统计学在人工智能和数据科学中的作用。

这些努力不仅是可取的——更是必不可少的。随着数据继续塑造科学、政策和产业的未来, IMS 必须既充当卓越标杆, 又成为包容、创新和产生影响的催化剂。通过深思熟虑的领导和大胆的行动, IMS 能够确保统计学始终处于科学事业的核心, 并继续在应对我们这个时代最重要的挑战中发挥关键作用。

结语

统计学并未在人工智能浪潮中被取代, 而是被赋予了强化与规范人工智能发展的关键使命。数据科学和人工智能领域的加速变革不仅带来了挑战, 也带来了机遇——统计学家有机会引领、塑造并守护科学、技术和社会中基于证据的推理的未来。

我们正站在学科发展的十字路口——这不仅关乎我们的学科, 更关乎在这个日益数据驱动的世界中, 我们作为知识的获取、传播和应用守护者的角色。我们今天做出的选择, 将深刻影响统计学的发展方向、科学研究的诚信基础, 以及循证推理在未来社会中的地位。

作者简介: 蔡天文 (Tony Cai) 现任美国宾夕法尼亚大学沃顿商学院 Daniel H. Silberberg 讲席教授及统计与数据科学教授; 宾夕法尼亚大学应用数学及计算科学教授; 宾夕法尼亚大学医学院生物统计, 流行病学, 及信息学系资深学者。2017-2020 年任沃顿商学院副院长。2006 年当选国际数理统计学会会士 (IMS Fellow)。2008 年获得世界统计学考普斯奖 (COPSS Presidents' Award), 2017 年任泛华统计学会 (ICSA) 主席, 2019 年获泛华统计学会杰出成就奖, 2023 年获美国统计学会 Noether 杰出学者奖, 2024 年当选美国科学促进会会士 (AAAS Fellow) 并担任国际数理统计学会 (IMS) 主席。曾任国际统计学顶尖刊物统计年刊 (Annals of Statistics) 主编及 JRSSB, JASA, AoS 等多个权威学术期刊的编委会成员。蔡教授的主要研究方向是大数据分析, 包括机器学习、高维统计、大规模统计推断、统计决策论、函数数据分析、非参数函数估计、以及在基因组与金融工程的应用。他的研究得到美国国家科学基金会、美国国家卫生研究院、及沃顿商学院全球倡议基金的资助。