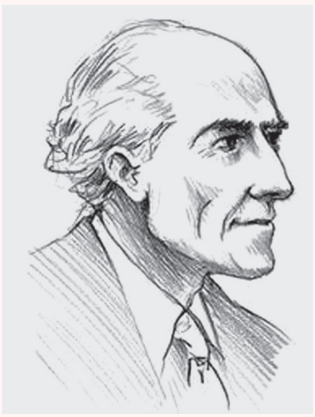
大数据热潮下关于抽样的冷思考

黄向阳

收集数据是统计工作的重点任务，而普查和随机抽样又是两种最基本的方法，这是目前的共识，也是统计入门课程的共同观点。这种格局的形成经历了一个复杂的演变过程。在 18 世纪和整个 19 世纪，从事政府统计和社会研究的学者都极少采用抽样方法，主流观点是数据越多越好，数据的种类也是多多益善，最好是天文地理人情世故无所不包。到 19 世纪末，数理统计学的主要奠基人高尔顿和皮尔逊重视的也是大样本 (large number) 而不是随机抽样方法。主要转折发生在 1900 年前后。在 1903 年的国际统计学会上，挪威统计学家凯尔倡导的随机抽样方法获得了多数人的赞同。数理统计学家奈曼则在 1930 年代从理论上证明了分层随机抽样方法的优越性，这样就从实践和理论两个角度确立了随机抽样方法的地位。在数理统计领域，费歇尔完善了基于随机抽样的小样本和实验设计，完成了统计学和概率论的结合。更有趣的一项进展发生在 1940 年代，物理学家乌拉姆 (Ulm) 和伟大的冯·诺依曼提出了蒙特卡洛方法，该方法的英文名 Monte Carlo 来自一家著名赌场，算是概率论对骰子和扑克的致敬吧。蒙特卡洛方法的基本思想是用算法或者计算机生成随机数表或者伪随机数表，随着计算机的发展，蒙特卡洛在很多应用场景和算法中变成了随机抽样的代名词。

进入 21 世纪的第二个十年，大数据 (big data) 突然爆红，在我国点燃大数据热潮的主要书籍则要首推 2012 年底翻译出版的《大数据时代：生活、工作与思维的大变革》一书，该书作者是维克托·迈尔－舍恩伯格（以下称呼他为舍恩伯格）。舍恩伯格在这本书的第一部分认为大数据时代已经来临，我们必须有一个思维变革，他列举了三个要点，第一点引用原文如下：要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本。这句



我也相信数据是多多益善，那叫大样本 (Large Number)。

——卡尔·皮尔逊

话被简化为要总体不要抽样，而随机抽样方法终将消失。那么，大数据时代会让随机抽样方法失去用武之地吗？也许，概率论和随机抽样的发展历程会给我们一些启示。

随机抽样是一种“合理猜测”的随机化策略

概率论起源于对机会游戏的数学思考，简单说来，就是数学家遇到赌徒的时候所发生的奇妙反应。最早研究此类互动的文献应该是卡丹诺在 1564 年发表的《机遇博弈》一书，不过这种模式即便在 500 年后的 20 世纪依然充满活力，战争、金融和赌博，为概率论的研究提供了源源不断的材料、灵感、金钱和博士学位。概率论提供了很多工具，随机抽样是应用面最广的工具之一，它让我们能够用随机化策略进行合理的猜测。

举个例子。在典型的球盒模型中，盒子是不透明的，里面放着 10 个球，分红白两色。游戏规则是：每次拿出一个球，然后放回去，多次重复这个有放回抽样。游戏任务是：根据看到的结果猜测白球的比例。这个简单的概率入门问题，包含了很多深刻的内容。

首先，白球比例有没有随机性？当然没有，盒子里的球都不会再有变化了，不是薛定谔的猫。就是说，我们要猜测的对象本身可以没有随机性，但是借助概率论，我们能够得到更好的猜测结果。这就是所谓随机化策略的具体应用。随机化策略的基本思想是：我们可以用概率论这种数学工具来处理本身不具备随机性的问题和对象。类似的场景，包括产品质量抽检和问卷调查，要调查的对象或者要估计的比例并不是随机变量。

其次，为什么我们要用样本而不是用总体，不就是 10 个球吗？答案是，在实际应用中，往往会遭遇各种约束，迫使我们根据不完整的信息进行推测。这些约束可以大致分为四类：规则（逻辑）、技术、成本和时间。四类约束可能并存，也可能有一个发挥主要作用。比如在球盒模型里的主要约束是规则，因为游戏规则不允许打开盒子来看，此类障碍在实际生活中往往表现为逻辑，即在逻辑上无法得到总体。我们将在第二部分对此做一点解释。

这里需要强调的是，不要用成本约束来替代其他类型的约束，很多时候，钱不是万能的。比如在没有电子计算机的时代，仅仅靠花钱或者人海战术是不可能在指定时间内完成计算任务的，也不可能在合理时间内完成人口普查数据的整理和分析。从抽样调查或抽样检验的实践来看，我们都很容易理解技术、成本和时间约束带来的压力，随机抽样方法无疑能够节省大量时间和金钱，下面我们将聚焦于大数据和机器学习，问一句：抽样老矣，尚能饭否？

逻辑上无法获得完整数据的应用场景

在有些应用场景中，逻辑上就不可能获得完整数据或者总体。这些场景分传统和现代两类。在传统应用场景中有所谓小 N 问题，即必须依赖少数样本的情形。典型的比如研制新药过程中的临床实验，需要找到合适的实验对象来测试新药的疗效，显然在这个阶段参与新药实验的人越少越好。当然，我们可以设想一个技术乌托邦：在遥远的未来，人类开发出强大的人体计算机模型，以至于能够在这个模型中测试新药的疗效，到那个时候，研制新药过程中的小 N 问题自然就消失了。不过，如果这种技术乌托邦是可能的，估计人类也几乎不可能患病了。总之，在最近的几十年里，一些传统领域的小 N 问题还是不可避免的，必

须依靠合适的实验设计和统计分析才可能得到合理的结果。

小 N 问题在最大胆的技术乌托邦中，在逻辑上是可以被消灭的，但是还有一大类问题在乌托邦里也无法获得全体数据。这就是所谓“流数据”。流数据在网络环境下指的是网站的用户数据，我们永远只能获得到当前为止的用户数据，如果要求预测在未来两个小时之内的用户行为，该怎么办？目前有效的方法和模型来自随机过程的统计分析，可以将历史数据视为抽样样本。这个方法也许存在诸多不足，但是无论如何我们都无法获得用户行为的“全体”数据。当然，这里也存在一个技术乌托邦，假设我们有个模型，可以前知五百年，后知五百年呢？不过，人类的好奇心会自然延伸到五百零一年，除非我们假设主脑计算机能够前知无限年，后知无限年，否则技术永远追不上人类的好奇心。

流数据还涉及一个重要的传统领域，统计质量控制。在工厂的生产现场，流水线上的产品真的像水流一样，必须按照时间先后经过质量控制岗位，负责控制质量的人要根据抽样结果来判断生产线的状态，决定是否需要调整甚至暂停生产线。在这类问题中，谈论“全体”数据也是毫无意义的空想。

隐藏在机器学习算法深处的随机抽样

在合理的技术发展假设之下，计算机的硬件和软件会不断提升能力，但是我们知道，技术体系越发达，体系自身产生的数据就越多。简单说来，在任何时刻，人类要处理的数据永远会压倒能够动用的数据处理能力。因此有效的算法，比如现在流行的各种机器学习算法，都必须想方设法地节省计算量。为了节省计算量，基于随机抽样的算法，或者说蒙特卡洛算法，就会成为不可或缺的手段。实际上，当前的大部分机器学习算法都有一颗蒙特卡洛的心。

以 2016 年战胜人类围棋高手的阿法狗 (AlphaGo) 为例，它的基本算法就源自蒙特卡洛。其大致原理是，围棋游戏中要评估的着数组合太多，因此用一种类似“胡乱挑选”的方式抽取一些组合来进行判断。当然，舍恩伯格会说，随着计算机算力的提升，围棋游戏终究能够用“死算”来解决，就像当年深蓝用死算方法在国际象棋中战胜卡斯帕罗夫一样。这个说法应该在逻辑上是合理的。不过这种合理是所谓归根结底的合理性，我们现在来看看比阿法狗更普遍的实际情形。

各种机器学习、深度学习或者人工智能的算法，本质都是求优化问题的最优解。目前的深度学习模型最多可能需要求解包含上亿个变量的最优化问题，现有算力很难有效解决此类问题。从长远观点来看，如果计算机发展到能够轻松求解包含一亿个变量的最优化问题，恐怕到那个时候，数据已经堆积到需要求解 100 亿个变量的最优化问题了吧。总之，算力应该永远落后于需求。在这些场合下，普遍流行的算法是随机梯度下降，其基本原理是随机抽取若干方向来试探优化的方向。这就是说大多数机器学习算法都有一颗蒙特卡洛的心，或者是芯？

全体数据与过拟合问题

刚刚接触数据处理的人，往往会有总体数据拜物教。前面说的是总体数据可望不可及，这里要强调的是，即使有了总体数据，可能还是离不开随机抽样。原因在于，如果直接用全体数据建模，就会出现数据科学家的噩梦：过拟合。过拟合是数据建模中的主要陷阱之一，其基本形式是，如果用手头全部数据建立

模型，那么这个模型在现有数据上可能表现很好，但用模型来预测新数据的效果可能很糟。最早发现并处理过拟合现象的是统计学家，计算机背景的数据科学家后来又独立发现了这种可恶的现象，可见，学科之间的隔阂会造成多少人力、脑力和电力的浪费。

处理过拟合现象的基本方法是把手头的全部数据分成三个部分：训练集、测试集和验证集。划入训练集的数据才会被用来建模。那么，让哪些数据进入训练集呢？当然要通过随机抽样。因此，随机抽样方法构成了数据科学工作流程的必然组成部分。它意味着，即使我们可以获得全体数据，建模的时候仍然需要使用随机抽样得到的样本做训练集。当然了，如果坚持全知全能的乌托邦假设，模型也是多余的。不过，如果是全知全能，我们还要大数据干什么？

根源在于低估了概率革命的地位

为什么在 21 世纪的数据狂潮之下，会重新出现总体和样本之争（上一次发生在一个世纪以前的 19 世纪末）。我觉得有两个可能的原因。

只有用随机化策略，才能破解大自然的奥秘。

——费歇尔

其一是“二选一”的心理陷阱，好像不是用总体，就是用样本，其实建模活动中最重要的原则是权衡而不是站队。统计学家乔治·博克斯有一句名言：“所有模型都是错误的，但其中有些是有用的。”既然如此，则非要在总体和样本、因果模型和相关关系、确定型模型和概率论模型之间做个了断，难免庸人自扰之嫌。

其二是低估了概率革命的地位，即低估概率论改造科学方法的力量。研究科学史的学者们在 1980 年代提出了概率革命 (probabilistic revolution) 的概念，即概率论的深入研究和普遍应用已经改造了科学研究和工程应用，甚至可以认为概率建模已经和传统的确定论模型并驾齐驱了。两种方法谁也吃不掉谁。当然，概率革命的说法还存在一定争议。但是，概率论已经成为数据科学的基础学科，这一点是可以确认的。用统计学家费歇尔的话来说，人类在和大自然的博弈中，只有通过“随机化策略”才能破解大自然的隐瞒手段。

由此看来，要让概率论失去实用价值，恐怕只有全知全能的上帝才能做到。而要用概率论模型解决实际问题，随机抽样又几乎是不可避免的。从实用角度来看，至少在未来几代人的时间里，随机抽样算法仍将是数据科学的核心算法之一，退一步说，即使得到了“全体”数据，我们还是离不开随机抽样。

在数据科学家的世界里，算法一定是非常多样化的，混合各种算法才是大数据时代的王道。哲学家兼数学家罗素曾经说过：参差多态，乃幸福之本源。为什么偏偏是罗素来告诉我们这个道理呢？可能是因为他还获得过 1950 年的诺贝尔文学奖吧，自然对人性和生活有更深刻的洞察。

作者单位：中国人民大学统计学院