



1 引言

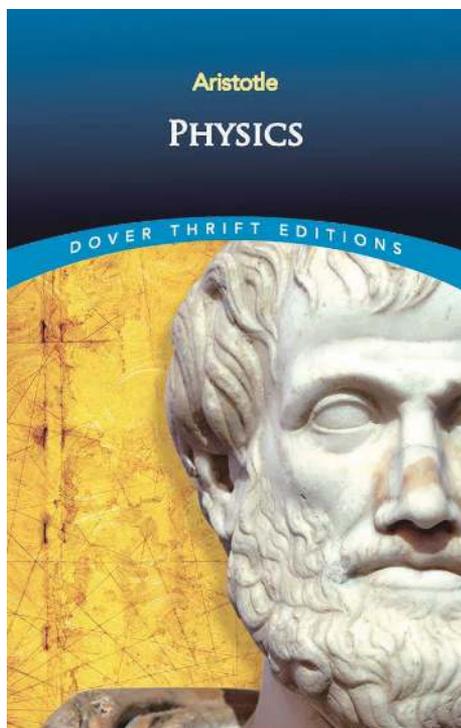
探求事物的原因，是人类永恒的精神活动之一。从古希腊的哲学到中国先秦的诗歌，都充满了对原因的追问和对因果关系的思考。比如，亚里士多德就在《物理学》(*Physics*)和《形而上学》(*Metaphysics*)两书中反复强调，我们只有知道了事物的原因，才能算真正理解这个事物。又如，屈原在《天问》开篇，就追问日月星辰运行的原因。

长期以来，人们一方面好奇地追问原因和结果的关系，一方面又苦于这些概念的模糊性。于是，这些话题在很长一段时间都仅仅局限在哲学和文学的范围内。精确地描述因果关系，尤其是用数学的语言来描述因果关系，则是非常近代的事情了。这一项思想飞跃，得益于现代统计学的发展。统计学家称之为“因果推断”(causal inference)。虽然因果推断在现代统计学的萌芽阶段就已经产生，但是它的发展并非一帆风顺：它长期被主流忽视、怀疑甚至攻击。直至最近四十年，尤其是最近十年，它才得到了广泛的认可和有力的研究，成为当今主流的研究方向之一。在最近的一篇文章中，Andrew Gelman 和 Aki Vehtari 评选了过去五十年中，统计学最重要的八个想法，排名第一的就是因果推断¹。当今世界，很多年轻的学者加入了因果推断的研究，他们来自统计学、经济学、社会学、政治科学、教育学、流行病学、计算机科学、哲学等等领域。毫不夸张地说，统计因果推断的研究迎来了它发展的黄金时代。

本文将回顾统计因果推断的历史背景，评述中国因果推断研究的现状，并且大胆推测它未来的发展前景。

¹A. Gelman and A. Vehtari, What are the most important statistical ideas of the past 50 years? 见 <https://arxiv.org/abs/2012.00174>。第一作者曾获得年轻统计学家的最高奖 COPSS 奖章。

2 哲学基础：因果推断何以成为可能？



亚里士多德《物理学》的一个英译本。这本书的 Book II 3 的开篇写道：“Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause)”，翻译成中文就是，我们探索的目标是知识，只有掌握了“为什么”，才算真正理解一个事物，即，掌握该事物的根本原因。

人们常常问关于原因和结果的问题。比如，某人死于肺癌，是不是因为他常常吸烟导致的？比如，我感冒症状减轻了，是不是因为服用了维生素 C 片导致的？比如，大学教育是否能够提高收入水平？类似的问题，充满了我们的日常生活。

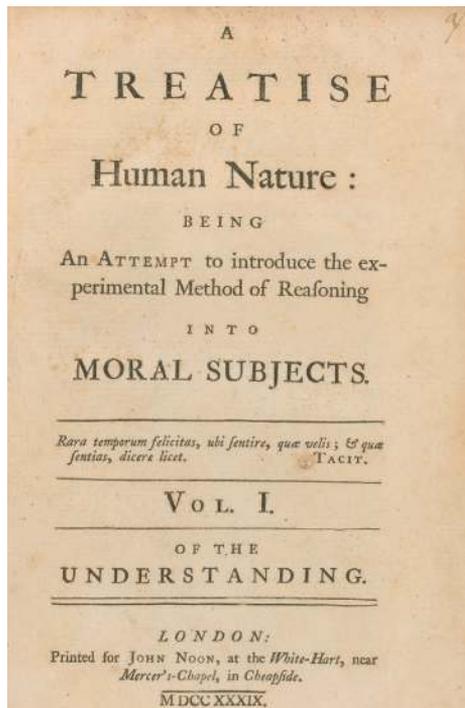
但是，这些看似直接了当的问题，却不容易回答。比如，有人吸烟，却没有得肺癌；有人不吸烟，却得了肺癌。比如，我可能仅仅喝白开水，感冒也会自己消失。比如，有人没有上大学，却做生意发了大财。当然，有点概率论常识的人很容易意识到，这些事件都带有随机性。从经验中，我们可能观察到吸烟的人更可能得肺癌；服用维生素 C 的人，平均来说，自我感觉感冒恢复得更快；上过大学的人平均收入更高。但是，这些统计上的“相关关系”是否就是“因果关系”呢？

大部分西方哲学家都认为因果关系是一条本质的、似乎毋庸置疑的定律。但是，苏格兰哲学家大卫·休谟（David Hume, 1711-1776）曾经抛出了一条惊人的论点。简言之，他认为人类仅仅凭经验，只能认识事物之间恒定的前后相继关系（constant conjunction），并不能认识任何因果关系。很多哲学家都努力回应休谟的质疑，因为若是承认休谟是对的，那么知识何以成为可能？若人类的知识仅仅是经验性的前后相继关系，那么人类似乎没有拥有任何“心智的荣耀”²。

² “[T]he sole end of science is the honor of the human mind.” —— Carl Jacobi (卡尔·雅可比)

哲学家们对休谟的回应似乎都是徒劳的。我在学生时代曾经上过邓晓芒教授“康德哲学”的课，他就直言，休谟是驳不倒的。的确，休谟这样的彻底的怀疑论者，是无法驳倒的。我回顾休谟的高论，并非想卖弄哲学史，因为休谟是绕不开的：无论何时何地，只要谈及因果推断，就可能有人引用休谟的论点质疑你问题的合理性。也正是因为休谟这种近乎诅咒似的言论，使得因果推断的数学化步履维艰。

然而，上个世纪统计学的几项辉煌成果改写了思想史。如今人们已经不再羞于讨论因果关系，统计因果推断的语言，深入到了几乎所有的应用领域。这些成果也许并没有完全解决休谟的问题，但是它们给出了因果关系新的思考方式和推理框架。下面，我将分三部分回顾历史。



休谟的名著《人性论》对哲学史产生了深远的影响，他指出了归纳推理的缺陷，认为我们对因果关系的信念仅仅来自于习惯(habit)和传统(custom)。

3 统计学中“哥白尼式的革命”：内曼的“潜在结果”模型

1923年，耶日·内曼(Jerzy Neyman, 1894-1981)还是波兰华沙大学的博士生，他的毕业论文是“概率论在农业实验中的应用”³。在这篇论文中，他提出了用于因果推断的“潜在结果”(potential outcomes)的数学模型，并将它和统计推断结合起来。他的想法非常自然，数学结构也很简单。下面简单地回顾一下。

以农业实验为例，考虑 n 块田作为实验的对象，实验者想检测两种肥料对于产量的影响。用 i 表示第 i 块田， $Y_i(1)$ 和 $Y_i(0)$ 表示如果用肥料 1 和肥料 0 分别对应的第 i 块田的产量，那么 $\tau_i = Y_i(1) - Y_i(0)$ 就是肥料 1 相对于肥料 0 对第 i 块田产量的因果作用。实验者随机地分配肥料 1 或者肥料 0 到第 i 块田，

³ 内曼的论文是用波兰语写成的。1990年，D. M. Dabrowska 和 T. P. Speed 将论文翻译成英文，题目是 *On the Applications of the Theory of Probability to Agricultural Experiments*，发表于 *Statistical Science*。潜在结果的基本想法也许在历史中早就产生了，但是将它数学化、且正式地用于统计学，内曼的文章是首次。内曼是现代统计学的奠基人之一，他对假设检验、置信区间、抽样调查和实验设计等领域的研究，成为现代统计学的标准范式。我国概率论和数理统计学的先驱许宝騄教授是内曼在英国指导的学生之一。



年轻时的内曼。内曼是加州大学伯克利分校统计系的创始人(照片由该系提供)。

所以最终我们要么观测到 $Y_i(1)$, 要么观测到 $Y_i(0)$, 不可能同时观测两者。显而易见, 在这个模型下, 因果推断的本质困难就是无法同时观测 $Y_i(1)$ 和 $Y_i(0)$, 也就无法直接观测到 τ_i 。观测单个的 τ_i 太困难, 退而求其次, 我们可以考虑研究它的平均数:

$$\tau = \frac{1}{n} \sum_{i=1}^n \tau_i.$$

这个 τ 通常被称为平均因果作用 (average causal effect)。这可能是因果作用最简单的定义了。到此为止, 内曼引入了一些数学记号来定义“因果作用”。也许读者会觉得这平平无奇, 无非就是 $Y_i(1)$ 和 $Y_i(0)$ 。但是, 这些记号将开启一扇窗, 迎接新思想的曙光。

潜在结果 $Y_i(1)$ 和 $Y_i(0)$, 以及平均因果作用 τ , 在某种意义上, 都是假想的数字。仅有这些定义, 还不能说明这个模型的现实意义。问题的关键是: 我们能否根据观测到的数据推断 τ ? 内曼给出了肯定的回答。在随机化实验下, 第 i 块田接受肥料 1 或者肥料 0 是完全随机的。用 $Z_i = 1$ 表示第 i 块田接受肥料 1, 用 $Z_i = 0$ 表示第 i 块田接受肥料 0。随机化实验固定接受肥料 1 和肥料 0 的田的总数, 分别是 n_1 和 n_0 , 对应的 (Z_1, \dots, Z_n) 这个向量是 n_1 个 1 和 n_0 个 0 的随机置换 (random permutation)。如果第 i 块田接受了肥料 Z_i , 那么我们观测到的产量就是

$$Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

这个恒等式似乎显而易见: 从数学上讲, 它无非说明, 当 $Z_i = 1$ 时, $Y_i = Y_i(1)$; 当 $Z_i = 0$ 时, $Y_i = Y_i(0)$ 。但是, 我在和朱迪亚·珀尔 (Judea Pearl) 交流时, 他认为这是因果推断最重要的恒等式, 因为它联系了左边我们能够观测到的结果和右边的潜在结果。

最终能够被观测的数据是 $\{(Z_i, Y_i) : i = 1, \dots, n\}$ 。一个显而易见的估计量是

$$\hat{\tau} = \frac{1}{n_1} \sum_{i: Z_i=1} Y_i - \frac{1}{n_0} \sum_{i: Z_i=0} Y_i.$$

它是接受肥料 1 和肥料 0 下, 平均结果的差值。内曼证明了 $\hat{\tau}$ 是平均因果作用 τ 的无偏估计 (即 $\hat{\tau}$ 的期望是 τ), 计算了这个估计量的方差, 讨论了如何估计这个方差, 还提出了一个基于 $\hat{\tau}$ 的中心极限定理的置信区间 (即这个区间以指定的概率盖住真值 τ)。最后一步的中心极限定理在内曼的原文仅仅是一个直

觉的证明，一直到了 Paul Erdős, Alfréd Rényi 和 Jaroslav Hájek 工作的出现，这类中心极限定理的证明才被严格化⁴。

上面仅仅讨论了一个最简单的数学结构：两个组的随机化实验中的因果推断。现实中的随机化实验丰富多彩，如何在各种随机化实验中做因果推断取决于具体的实验设计方案。内曼本人于 1935 年在英国皇家统计学会宣读的论文，讨论了随机区组设计（randomized block design）和拉丁方设计（Latin squares design）的因果推断，引发了包括罗纳德·费希尔（Ronald Fisher）在内的统计学家的激烈争论。同时期，费希尔对随机化实验进行了深入的研究，虽然他并没有使用内曼潜在结果的记号，但是因果推断始终是他思考的对象。随后的几十年，随机对照实验（randomized controlled trial；RCT）成为美国食品药品监督管理局批准新药的黄金标准。最近二十年，大量的随机化实验出现在社会科学中，用来研究复杂社会问题中的因果关系。比如，麻省理工学院和哈佛大学的三位经济学家，Abhijit Banerjee, Esther Duflo 和 Michael Kremer，因为用实验的方法研究发展经济学，获得了 2019 年的诺贝尔经济学奖。

内曼生前对自己在统计假设检验方面的奠基性工作颇为自豪，认为那是统计学中“哥白尼式的革命”（Copernican Revolution）⁵。他并未预料他在因果推断的奠基性工作，也将产生深远的影响。这个影响则是由唐纳德·鲁宾（Donald Rubin）开启的。

4 统计学的拓荒者：鲁宾关于观察性研究中的因果推断的研究

从直觉上，也许大家不会对随机化实验中的因果推断感到惊奇。毕竟随机化实验保证了两个组在平均意义上是相似的，那么他们之间的区别就可以归因于不同肥料对产量的因果作用。但是，现实的统计问题，很多数据收集并非源自随机化实验——这类研究通常被称为观察性研究（observational study）。比如，如果要研究吸烟和肺癌的因果关系，基本的伦理不允许我们随机地让一部分人抽烟、让一部分人不抽烟。再如，研究大学教育对收入的影响，我们不能随机地让一部分人上大学、让一部分人不上大学。很多流行病学和社会科学的问题，本质上一定是观察性研究，人们也迫切地想从这些观察性研究中获得关于因果关系的知识。

虽然潜在结果模型成功地数学化了随机化实验中的因果推断，但是它长期并未用于观察性研究——内曼本人是持怀疑态度的，因为缺乏随机化，观察性研究有太多复杂性，比如抽烟的人和不抽烟的人，可能就是两群完全不同的人，不具有可比性。虽然他从未尝试用他的潜在结果模型分析观察性数据，但是他

⁴ 这方面的文献综述是：Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112, 1759-1769.

⁵ 见内曼的传记：C. Reid (1982), Neyman - From Life。注意，哥白尼和内曼都是波兰人。