社会科学研究方法

因果推断

元景阳

2025-08-04

# 前言

当亚里士多德在《物理学》中提出"四因说"时，他或许未曾预见两千三百年后社会科学面临的根本性挑战——如何在复杂的社会系统中构建兼具解释力与预测力的因果模型。从古典政治哲学对城邦政体的类型学归纳，到霍布斯《利维坦》中社会契约论的数学化演绎，再到当代计量经济学"可信性革命"的技术突破，人类对社会现象的理解始终在"理论建构"与"经验验证"的张力中螺旋上升。本书正是对这一认识论传统的当代回应，旨在构建连接政治哲学思辨传统与数据科学实证范式的桥梁。

本书的学术价值体现于对两种知识生产路径的有机整合：一方面继承古典政治学对制度演化的深层追问，另一方面融合现代计量经济学对因果识别的技术探索。这种双重性在章节设计中得到充分体现——首章追溯从霍布斯到马克思的认识论争辩，揭示社会科学方法论的哲学根基；次章系统梳理研究设计的规范框架，将休谟对因果关系的怀疑主义转化为可操作的变量控制策略；第三章则通过矩阵代数与概率论的严谨推导，为后续章节的因果推断奠定数学基础。最终章对"do算子"与潜在结果框架的深度解析，标志着传统政治哲学命题与现代数据科学的深度融合。

在数字时代的学术图景中，本书特别关注方法论革新与学科发展的共生关系。当结构主义学者强调"社会现象背后存在深层结构"时，我们通过因果图理论揭示这种结构的可识别条件；当后实证主义者批判"数据与理论的二元对立"时，我们以匹配法和双重差分技术展示如何在观察数据中逼近实验条件。这种将哲学反思转化为可执行的技术路径，正是本书区别于传统方法论著作的核心特征。

本书的写作过程受益于诸多学术前辈的智慧启迪。从陈强老师对计量经济学的系统阐释，到Angrist与Imbens在"可信性革命"中的方法论突破，从King、Keohane与Verba对定性研究的量化逻辑重构，到Pearl因果推断理论的革命性贡献，这些学术成果构成了本书的知识谱系。特别感谢杨立华老师《政治学与公共管理研究方法基础》为研究设计章节提供的方法论参照，以及马亮老师《学术祛魅》对概念化与操作化的精辟论述。

### 本书结构说明

全书遵循"哲学基础-方法论-技术工具-实践应用"的逻辑链条展开：

* 第一章追溯从方法论自然主义到后实证主义的哲学演进，揭示社会科学方法论的本体论基础
* 第二章构建研究设计的规范框架，系统解析选题策略与变量关系的逻辑建构
* 第三章夯实数理基础，通过矩阵代数与概率论推导计量模型的数学本质
* 第四章系统阐释因果推断的理论框架，比较潜在结果模型与结构因果模型的异同
* 第五章聚焦统计推断的技术实现，从OLS估计到前沿匹配算法进行实证解析

这种编排既体现知识生产的内在逻辑，也符合认知进阶的规律。每个章节均配备Stata操作示例，通过"理论阐释-数学推导-代码实现"的三维训练体系，帮助读者实现从概念理解到实践应用的跨越。

元景阳  
2025年8月

# 研究方法的哲学基础

## 存在论

世界如一枚硬币，既有可测量的金属材质，又有雕刻其上的图案花纹。想要真正理解世界，在掂量硬币的重量的同时，也要欣赏其花纹与光影。想象此刻的你站在桂林山水前。“美”是山水的客观属性，还是你内心的主观感受？这个简单的问题揭示了人类认识世界的两种基本方式——客观与主观。

### 客观—现实主义

客观—现实主义范式认为，社会现实是客观存在，独立于个人的意识和认知的。根据这一范式，社会现象有其独立的客观特征，可以通过科学方法进行观察、测量和分析。以此为基础或出发点，客观—现实主义在认识论上和方法论上，强调经验主义或实证主义，认为知识来源于对现实世界的经验观察和数据收集，并追求识别社会现象之间存在可识别的因果关系。这样的存在论可以称为客观主义、唯物主义、现实主义等，为了避免语境转换造成的词不达意，我们复合使用客观—现实主义。

### 主观—建构主义

主观—建构主义范式认为，社会现实是通过人类的主观认知和社会互动建构而成的。根据这一范式，社会现象的意义和现实是由个体和群体共同赋予和解释的，强调研究者理解人们如何通过互动和交流来建构人们所处的社会世界。就有了“社会世界是建构的而不是独立于人主观意识的客观存在”这一论断或认识。例如，身份是一种通过个体主观体验及社会互动形成的社会现象，涉及家庭、朋友等多元关系中的意义协商，并受文化差异（如性别、种族定义）、社会权力结构（如阶级地位）及规范标签（如刻板印象）的动态影响，具有流动性和历史情境性。

不同的存在论流派的争论由来已久，在科学革命后的当今世界，客观—现实主义无疑占据着中心与主导的地位，因此后文的主干部分介绍的也基本都是客观—现实主义一派。

## 科学革命的开端

### 方法论自然主义

肇始于文艺复兴时期的方法论自然主义是影响科学发展的最早的哲学流派，这一时期科学研究与经院哲学的诠释在方法论上分家，伽利略开启了以方法论自然主义弥合二者分歧的思路，即科学研究的方法论不涉及到宗教思考——超自然存在及终极真理仍然由教会掌握。人们认识世界的方式自此发生了重大转变，在渐进积累了一整个中世纪之后，那颗著名的苹果，砸落到一个聪明的脑瓜的那刻，其意义远大于物理课本上的万有引力定律，这一事件象征着这样一个时代来临：**从仰望神谕转向俯察人间，从依赖经典权威转向依靠观察、实验与理性推理**。

### 自然科学向社会科学的延申

亚里士多德在其哲学体系中提出了著名的“四因说”，系统阐述了自然界事物生成与变化的原因。他认为，真正理解一件事物，必须同时揭示其四种不同类型的因：其一为质料因（causa materialis），即构成事物的物质基础；其二为形式因（causa formalis），即事物所具有的结构、形状或本质特征；其三为动力因（causa efficiens），即促使事物生成、变化或运动的外在力量；其四为目的因（causa finalis），即事物存在或运动的目的与功能，是亚里士多德因果论中最具目的论色彩的一环。通过这一因果分类法，亚里士多德力图超越机械式的因果解释，建立一个涵盖本质、结构、动力与目的的多维解释体系，成为西方形上学和科学方法论的重要源流。

作为比较政治学研究先驱的亚里士多德开创性地对一百多个希腊历史上存在过的城邦政体进行了详细的分类，然而遗憾的是，亚氏首先构建了三个母本（君主、贵族与共和），而现实中存在的城邦政体只是母本的衍生。与其师柏拉图不同的是，亚里士多德并不试图以**纯粹的理念哲人王**统领人世，而是客观上承认了多元利益造就的多元共治和多元共赢。

中世纪的宗教为与王室和解（竞争）而逐步放弃了原始的蒙昧精神，科学的幼苗在柏拉图和亚里士多德两位的精神遗风的培育下逐步发芽，终于长成一棵果树，硕果挂满枝头。所以，当那颗苹果落下时，选择研究它为何落下、如何落下，开启了一场重塑人类理解自身与世界方式的伟大征程。

追随亚里士多德的脚步，中世纪的学者往往从目的论的角度出发，追问**苹果为何落下**——是否符合上帝的旨意或某种终极目的。而科学革命后的思想家们则更关注“**苹果是如何掉下来的**”，即试图通过观察和实验揭示支配自然现象的普遍规律。这种从“为什么”向“如何”的转变，标志着认识论的重大转向。

认清普遍、可量化的规律后的人们决定对自我生活的世界进行反思，从各个方面汇集而来的思考很快构建起经验主义的基本框架。有趣的是，尽管经验主义的思想渊源可以追溯到中世纪晚期的经院哲学家如**奥卡姆的威廉**，但作为一种系统的哲学传统，它在17至18世纪的英国得到了系统发展，并由此得名“**英国经验主义**”。与此同时，欧洲大陆上则兴起了以笛卡尔为代表的**理性主义**传统，强调通过先验推理获得知识。这两种认识论路径之间的张力贯穿了近代西方哲学的发展历程，也深刻影响了后来社会科学的方法论基础。

### 唯理主义与经验主义

随着自然科学的哲学转向，社会科学也开始寻求自身的认识论基础。其中，政治学作为一门独立学科的形成过程尤为典型。被誉为“君主宝鉴”的《君主论》始终是我们讨论政治学转型绕不开的书籍，在羞于谈论权谋却总以权谋私的政治环境中，一本较为直白的书最是难得。尽管如此，真正开启政治科学进程的更多还是霍布斯所著的《利维坦》一书。

此处转引包刚升老师对本书的评述：

霍布斯认为，人类因无限欲望陷入持续冲突。为摆脱混乱，需建立拥有绝对权力的国家。他强调人性中的自我保全本能和对权力的追求，认为**资源竞争、相互猜疑和荣誉感**导致自然状态下的生活**危险、贫困且短暂**。为此，霍布斯提出“社会契约”：人们为安全自愿转让权利，形成国家（即“利维坦”），以维持秩序、抵御外敌。国家可通过征服或契约建立。尽管国家可能存在问题，但其存在是避免无政府混乱的必要条件。

霍布斯在《利维坦》中结合个人主义与科学逻辑分析政治问题，为现代政治理论奠定基础。他的思想影响深远，即使后来如洛克等思想家也难以忽视其贡献。总体而言，霍布斯标志着西方近现代政治理论的重要转折，其关于**人性、权力与国家关系**的思考至今仍有影响力。

也就是从在这里开始，经验主义内部出现了分歧。不难看出，尽管霍布斯演绎出的**人性恶**的观点不无道理，但以**社会契约**来予以改善则不完全是一种现实的观察，更多是脑海中的推演。哈耶克在《自由秩序原理》（又译《自由宪章》）一书中也有过精彩的论述：

唯理主义（理性主义）传统假定，人生来就具有智识的和道德的秉赋，这使人能够根据审慎思考而形构文明；而进化论者（经验主义者）则明确指出，文明乃是经由不断试错、日益积累而艰难获致的结果，或者说它是经验的总和，其中的一部分为代代相传下来的明确知识，但更大的一部分则是体现在那些被证明为较优越的制度和工具中的经验……

洛克接过霍布斯的“社会契约”概念，却对其进行了合乎资本主义发展现实的革命性的改造。在洛克看来，自然状态并非全然“污秽、残酷与短暂”，而是存在着“自然法”的约束，人拥有**生命、自由、财产**等不可剥夺的“自然权利”。政府的建立，并非为了创造秩序本身，而是为了更好地**保护**这些先在的权利。当政府违背契约、侵犯权利时，人民有权**反抗**。洛克的学说，为自由主义奠定了基石：**个人权利优先、有限政府**和**基于同意的统治**。

在**认识论**层面，洛克在其著作《人类理解论》中提出了“白板说”，认为人的心灵如同一块空白的石板，所有知识皆来源于感官经验。这种观点体现了典型的英国经验主义传统。与之形成鲜明对比的是法国哲学家笛卡尔，他在《第一哲学沉思集》中提出“我思故我在”，强调理性思维先于经验，主张通过演绎推理获取知识。这两种路径不仅构成了近代哲学的基本张力，也深刻影响了社会科学中的研究范式之争，经验主义和唯理主义之间经久不息的较量正式开启。

## 社会思潮的产生与认识论的发展

社会思潮的形成是复杂而多维的过程，它不仅反映了特定历史时期的社会矛盾与人们的思想追求，也深刻影响了科学研究的方法论选择与发展路径。从文艺复兴到启蒙运动，从现代自然科学迈向现代社会科学，不同的哲学思想和社会理论相互交织、碰撞，共同塑造了我们今天所看到的知识版图。在主要的社会思潮之中，最具有影响力的分别是自由主义、保守主义和社会主义。

### 自由主义的双重面向

古典自由主义（如洛克、亚当·斯密）深受英国经验主义熏陶。它强调**个体的感官经验、自由探索和利益驱动**是理解社会的基础。自由市场被视为一个庞大的、自发的“发现程序”，通过无数个体的互动（试错），信息得以传播，资源得以有效配置。这种对社会秩序形成的理解，更倾向于进化理性，即秩序是无数个体行动无意中形成的复杂结果，而非某个单一头脑的设计。

然而，自由主义内部也存在强烈的**理性主义**倾向，尤其在涉及政治制度设计时（卢梭的社会契约论、孟德斯鸠的三权分立以及美国的联邦党人的政治实践），各种说辞无不使人相信人类理性能够设计出保障自由、限制权力的制度框架。边沁的功利主义更是将理性计算推向极致，试图用“最大幸福原则”作为衡量政策和社会制度的科学标尺。

约翰·密尔在《逻辑体系》一书中给出了归纳的因果五法，参考互联网哲学百科全书对此做出的论述如下：

密尔对归纳法的著名论述揭示了信念的经验基础。他提出了五种用于识别因果关系的实验探究方法，旨在从现象前后的情境中找出那些通过**不变法则**与现象相关联的因素。简言之，我们通过在多种情境下观察现象，来判断某种因果关系是否成立，以下五种方法为此提供了系统性依据：

* **求同法** ：若多个不同情境中某一现象都出现，且唯一共同因素存在，则该因素可能是原因。
* **求异法** ：若某现象在一个情境中出现，在另一个中不出现，而两者仅有一个因素不同，则该差异因素可能是原因。
* **求同求异联合法** ：结合前两种方法：一方面找现象出现时的共同因素，另一方面看该因素缺失时现象是否也消失，从而更可靠地确定因果关系。
* **共变法** ：若某因素变化时，现象以固定方式随之变化，则二者存在因果联系。
* **剩余法** ：从复杂现象中剔除已知原因引起的变化，剩余部分可归因于尚未解释的因素。

### 保守主义与怀疑主义

如果说自由主义（尤其是其古典形式）更多地根植于经验主义对个体行动和自发秩序的观察，那么保守主义的核心则是对**抽象理性设计社会蓝图**的深刻怀疑，其思想源泉同样与经验主义传统，特别是对**历史实践、传统智慧与制度演化**的重视紧密相连。埃德蒙·伯克在《法国革命论》中的论述，堪称保守主义对理性主义社会工程学的经典批判。

伯克并非反对变革本身，而是反对基于抽象理念（如“天赋人权”、“社会契约”）的激进革命。他认为社会是一个**复杂的有机体**，是历经漫长岁月、由无数代人的实践、妥协和智慧沉淀而成的。其**制度、习俗和传统**（即使存在不完美）包含了超越个体理性的“集体智慧”，是维系社会秩序和情感纽带的基石。盲目地用“理性”的手术刀切割这些传统，试图按照某种乌托邦蓝图重建社会，其结果必然是灾难性的——法国大革命后期的恐怖似乎印证了他的预言。

与此同时，在**认识论**层面，英国哲学家大卫·休谟对自由主义中的唯理主义一面的打击是毁灭的。休谟以其**彻底的怀疑主义**动摇了经验主义乃至所有声称具有必然性的知识的根基。他的核心洞见直指科学方法和理性社会设计的核心——何为因果，参考互联网哲学百科全书对此的论述如下：

根据休谟的观点，因果关系由三个基本观念构成：**时间先后、空间邻近和必然联系**。

* **时间先后**：原因必须先于结果。若B在A之前发生，说A导致B是不合逻辑的。
* **空间邻近**：因果事件需在空间上接近。例如，英国人扔石头的同时中国窗户破碎，并不意味着两者有因果关系。
* **必然联系**：仅有时序与邻近不足以构成因果。例如，喷嚏与灯灭同时发生并不等于因果关系。我们之所以认为有必然联系，是因为我们反复经验到类似A的事件总是跟随类似B的事件，从而形成心理习惯——即看到A出现便会预期B的出现。这种“期待感”构成了我们对必然联系的观念。

休谟指出，人们常犯的一个低级错误便是，由于既往的感知如此，便武断地认为因果关系当然存在，然而这种论证始终没有开端，即为**循环论证**。

值得注意的是，尽管休谟以其怀疑主义闻名，他对因果本质的质疑，反而促使后来的科学家和哲学家更加谨慎地对待变量之间的关联，从而推动了**实验设计与统计因果推断**的发展。

### 社会主义的实证精神

激进自由主义者的左转在社会思潮中形成了社会主义，而在认识论上则出现以孔德为首的**早期实证主义**和以马克思为首的**历史唯物主义**两大流派。

奥古斯特·孔德被视为社会学和科学哲学的创始人，他的代表作是《实证哲学教程》，参考中国大百科全书撰写简介如下：

孔德在书中提出**实证主义认识论**体系。他将人类理智发展分成三个阶段：

* **第一阶段**：神学的或虚构的，这一阶段总是临时性的；
* **第二阶段**：形而上学的或抽象的，纯粹是过渡性的；
* **第三阶段**：实证的或真实的，这是决定性的阶段，在这个阶段，相对性取代了绝对性，对规律的研究取代了对原因的研究。

三阶段规律同时支配着人类社会历史、科学史以及个人心智成长过程。与此相联系，孔德按照研究对象的普遍性递减和复杂性递增的顺序对科学进行分类：数学、天文学、物理学、生物学、社会学。

正是通过这样的顺序，休谟意图表明，科学已经进入到实证阶段，科学整体的基本精神就是实证精神，它是统一各门学科的基础。此外，孔德还提出“社会学”的概念，把社会学定义为一门以经验的实证方法研究人类社会现象的科学，并认为社会科学与自然科学在本质上不存在差别，因而可以用研究自然科学的方式研究社会科学。

**马克思**与孔德则正好相反，在《关于费尔巴哈的提纲》中，马克思宣言：“哲学家们只是用不同的方式**解释世界，而问题在于改变世界**。” 这标志着其与孔德式静态观察的根本决裂。社会规律并非等待发现的永恒法则，而是人类在物质生产实践中**历史性生成**的产物。

换句话说，作为社会权力的体现形式之一的“社会学”本身即是**统治阶级意识形态**的一部分，因而不自然地就将资本主义生产关系自然化，掩盖了剩余价值剥削的实质。

马克思在《<政治经济学批判>序言》中写道：社会的物质**生产力**发展到一定阶段，便同它们一直在其中运动的现存**生产关系**或财产关系（这只是生产关系的法律用语）发生矛盾。于是这些关系便由生产力的发展形式变成生产力的桎梏。那时**社会革命**的时代就到来了。随着**经济基础**的变更，全部庞大的**上层建筑**也或慢或快地发生变革。

无产阶级的历史使命，不在于维护既存的不公平的社会秩序，而要在革命中荡涤尘埃。

### 小结

科学革命标志着人类认识世界方式的根本转变。**方法论自然主义**的确立，使科学研究摆脱宗教束缚，转向观察、实验与理性分析。这一变革不仅推动自然科学的发展，也为社会科学奠定了方法基础。

随着研究对象从自然转向社会，经验主义与唯理主义的争论延伸至社会科学领域。霍布斯、洛克等人以不同方式探讨国家与人性的关系，休谟和密尔则系统化认识论哲学，使社会研究深化科学内涵。在思想层面，自由主义、保守主义与社会主义三大思潮相继兴起：

* **自由主义**强调个体权利与市场机制，融合经验主义与制度设计；
* **保守主义**重视传统与制度演化，反对理性主义的社会改造；
* **社会主义**则批判资本主义结构，主张通过历史实践推动社会变革。

这些思潮体现了不同的认识论立场，也塑造了现代社会的基本格局。在浩荡的历史长河流动中，人们重新思考社会的本质与未来的方向，终于为现代社会科学的发展提供了坚实的方法论基础。

## 面向当代的方法论哲学

参考《政治学与公共管理研究方法基础》一书，补充了如下内容：

### 实用主义

19世纪70年代，美国哲学家皮尔斯创了实用主义哲学，经由詹姆士、杜威等人的发展，成为20世纪影响美国科学研究及科学方法的重要哲学流派。

实用主义的英文“Pragmatism”一词由希腊语“πρᾶγμα”（行动）衍生而来，可见，实用主义哲学在产生之初就与行动密不可分。在方法论上，实用主义者将科学研究视为指导未来行动的实际工具，而非对问题的终极解答。实用主义哲学强调知识和观念的“有用性”。詹姆士在其著作《实用主义：某些旧思想方法的新名称》中指出：“一个观念，只要我们相信它对我们的生活是有益的，那么它就是‘真的’。”

同样地，在科学研究及科学方法上，实用主义并不认为存在唯一的适用方法或绝对标准。实用主义者愿意采纳任何东西，“既遵从逻辑，也遵从感觉，并且重视最卑微、最具个人性质的经验。要是神秘经验有实际的效果，它也愿意重视神秘经验。”接受或拒绝某一命题或真理的关键在于其所能带来的实际利益或实际价值。

### 逻辑实证主义

伴随着西方哲学与社会科学研究中的“语言学转向”，实证主义哲学也从早期实证主义逐渐过渡到逻辑实证主义的新形态。两种实证主义与在认识论上都坚持经验主义原则，即主张“知识来自对可感现象界的认识，任何知识的产生都应完全归于可证实的经验”。它们的主要区别在于，逻辑实证主义在经验主义认识论之外，同时强调了对“**概念意义的逻辑分析**”。逻辑实证主义者强调，科学概念和理论不仅要能观察验证，**在逻辑上也必须讲得通，能清晰地分析出意义**。有学者总结了方法论上逻辑实证主义的四个特点：

* 注重“科学语言的逻辑分析”；
* 坚持经验主义认识论；
* 注重归纳方法；
* 强调科学研究与科学方法在获取知识上的逻辑性与客观性的统一。

虽然逻辑实证主义在哲学上的影响力日渐式微，但**逻辑与实证**两种方法至今仍对科学研究与科学方法产生着深刻影响。

### 结构主义

严格意义上来说，结构主义并不是一个统一的哲学派别，而更像是一种研究方法。结构主义首先兴起于语言学领域。索绪尔的《普通语言学教程》被视为语言学的“哥白尼革命”，其中提出的共时性（synchronic）与历时性（diachronic）原则、语言（langue）与言语（parole）等观点，深刻影响了现代语言学的发展，并奠定了结构主义方法论的基础。此后，经过列维-斯特劳斯的发展，结构主义逐渐在人文学科中产生广泛影响。自20世纪50年代起，学者基于对方法论的共识形成了一股结构主义的哲学思潮，广泛影响着西方人文及社会科学，尤其是文化学、心理学、语言学及社会学研究的发展。

结构主义的认识论基础与实证主义存在明显差别。结构主义认为，经由感官所了解到的自然和社会现象只是表面的，而非真实的事实；隐藏在表面现象背后的深层次结构才是科学家应该了解的真实知识。语言学家瑞克森·吉布森将结构主义方法论的基本原则及主要特征归结为以下五个方面：

* **强调研究对象的整体性**  
  结构主义研究者认为自然和社会现象都是复杂的统一体，因此无法片面地对某一部分或某个方面进行解释，对于部分的理解需要放在整体中进行。
* **注重系统内部各要素之间的关系**  
  这种关系既包括研究中整体与部分的关系，也包括研究对象与外部世界的关系性。
* **对主体的消解**  
  对主体的消解基于结构主义对整体性及关系性的强调。结构主义认为，在社会科学研究中，研究者对“人”的本质的探索不是通过直接研究人类自身，而是研究“与人类有种种关系总和的整体”。正如马克思所说：“人的本质是一切社会关系的总和。”因此，结构主义研究更加关注基于人及其关系所组成的整体，整体一旦建立，人在研究中的主体性意义便让位于整体。
* **强调结构的自我调节与自适应**  
  自我调节是指系统内部各因子之间所进行的转换、调整与适应，以保证结构内部的连续与稳定。
* **注重结构的动态性与转换**  
  结构主义者并不认为结构是静态的、一成不变的，而是由不同的转换机制组合而成的动态机制。皮亚杰认为：“一切已知的结构，从最初级的数学群结构，到决定亲属关系的结构等，都是一些转换系统。”

### 后实证主义

20世纪70年代中期，在社会科学领域掀起了定性与定量研究方法的争论，这挑战了当时以调查和统计分析方法为主的实证主义研究范式。后实证主义哲学思想正是基于对早期实证主义和逻辑实证主义的反思，以及“**人文主义**”研究立场对社会科学带来的挑战发展而来。

后实证主义避免了**对经验主义原则的过分强调**。正如杰弗里·C. 亚历山大指出的：“全部科学发展是一个双轮的过程，既为经验的论证亦为理论的论证所推动。”总体而言，在方法论意义上，后实证主义具有四个关键特征：

* 注重“经验环境与形而上学的双向互动”；
* 摆脱了片面注重某类研究方法的成见，**定性与定量相结合**的混合研究方法以及多元研究方法成为新的发展趋势；
* **价值中立**原则受到挑战；
* 不仅强调科学知识的“可证实性”，更强调其“**可证伪性**”（falsifiability）。

“可证伪性”由卡尔·波普尔提出，已经成为现代科学研究中检验理论可靠性的重要标准。后实证主义的方法论奠定了现代科学的基础。

## 科学研究的哲学发展路径

科学研究的哲学基础在存在论、认识论与方法论的互动中逐步深化，塑造了社会科学与自然科学的范式分野与融合。**存在论层面**，客观—现实主义主张社会现象具有独立于个体意识的客观规律，而主观—建构主义则强调社会现实由人类认知与互动动态建构。

这种本体论分歧直接映射至**认识论**领域：经验主义（洛克、密尔）坚持感官经验为知识源头，通过**归纳与实验**揭示因果关系；唯理主义（笛卡尔、卢梭）则推崇**先验理性与逻辑演绎**，主张通过抽象思维构建制度框架。两者的张力催生了休谟对因果必然性的怀疑，推动科学方法对变量关联的严谨验证。

**方法论层面**，从方法论自然主义到实证主义，科学范式逐渐形成经验验证与逻辑分析的统一标准；而结构主义与后实证主义则进一步突破主客观二元对立，强调理论可证伪性、混合研究方法及社会结构的动态性。

研究方法的哲学演进不仅反映了社会思潮的发展与分野，更揭示了现代社会科学在技术化与人文关怀间的平衡需求——既需借助定量工具捕捉规律性，亦需通过定性反思理解意义网络，最终在实用主义导向的多元方法论融合中回应复杂社会现象的复杂与多样。

99 王正绪，《厘清社会科学研究范式的基本差别》，[中国社会科学网](https://www.cssn.cn/skgz/bwyc/202409/t20240913_5777429.shtml)

包刚升，[哔哩哔哩：国家起源于人与人的战争状态？如何解读霍布斯的《利维坦》？](https://www.bilibili.com/video/BV1S2HsePERE)

哈耶克，《自由秩序原理》，三联书店

Mill, John Stuart条目，[Internet Encyclopedia of Philosophy](https://iep.utm.edu/milljs/)

Hume, David条目，[Internet Encyclopedia of Philosophy](https://iep.utm.edu/hume/)

《实证哲学教程》条目，[中国大百科全书网](https://www.zgbk.com/ecph/words?SiteID=1&ID=399321)

马克思，《关于费尔巴哈的提纲》，[中文马克思主义文库](https://www.marxists.org/chinese/marx/marxist.org-chinese-marx-1845.htm)

马克思，《<政治经济学批判>序言》，[中文马克思主义文库](https://www.marxists.org/chinese/marx/06.htm)

杨立华  等，《政治学与公共管理研究方法基础》，北京大学出版社

# 社会科学的研究设计

## 找准选题

### 导言

在网上冲浪时，我们常常见到如下的讨论：

张三：某大学发表高水平论文证明了低收入人群喜欢吃辣。

李四：\*\*\*不是常识吗？这些学者……（此处省略不文明用语）

作为社会科学的研习者，我们当然不会对研究意义产生虚无主义的印象，但一个不容忽视的问题的确摆在人们的面前，当前的社会科学界终于是陷入到一种为了“规范化”而“规范化”的境地，在学界“内卷”的体系之外，社会大众的评价似乎变得不再重要，研究的价值往往脱离现实。

这一点在量化研究之中的“显著性”和“稳健性”两大概念中体现得最为明显（简单理解即越显著越稳健越好），一些论文通篇粗制滥造，却执着于在量化部分雕花：你有一颗星显著（经济学认为此时犯错误的概率不超过百分之十），我就要有三颗星（不超过百分之一）；你用一个稳健性证明了你的结论可靠，我就拉来十个证明结论稳健性的方式，做完后发现只有三种方式做出来可以接受（也就是至少有“一颗星”），那就专门把这三个放在上面。

于是最终呈现给外行人的结果便是“学术狗屁”，即洋洋洒洒几十页，证明了一个凭借直觉或者有一点生活经验便可以得出的结论，且通常与迫切需要改善的社会现实并无干系，呈现出一种“garbage in, garbage out”的状态。

### 好问题的标准

好的研究问题应当符合五个关键标准，这些标准共同构成了一个层次分明的"金字塔模型"。

1. 问题必须从小切口的困境出发，聚焦具体现象而非宏大命题，你自己要能把握住而不是给别人画饼。
2. 问题需要与既有研究脉络形成对话，明确识别并填补学术空白，比如将普遍原理运用于中国实际。
3. 优秀的问题应当兼具理论野心与实践价值，能够通过小问题推动理论进步，同时回应现实关切。
4. 问题必须具有可执行性，与研究者的能力、时间和资源相匹配，避免好高骛远，若缺乏定量技能就应选择质性方法**（当然本文目的是为定量打基础，你不能放弃！！）**。
5. 真正的好问题应当能激发研究者持续的探索欲，形成自我解惑的内在动力，不能写出一个自己都不想阅读的作品。

按照三个层次区分并构成的完整的评价体系有：基础层要求具体可操作，中间层强调学术对话与创新，顶层注重理论与现实的平衡。需要牢记的是，提出好问题只是研究的起点，更重要的是通过扎实的方法论和持续努力将其转化为有价值的学术成果。

### 寻找好问题

本部分参考《政治学与公共管理研究方法基础》：

要找到真正有价值的研究问题，首先需要明确研究目的，即这项研究究竟要解决什么问题、达到什么目标。

研究目的通常包含研究背景和研究意义两个核心维度。研究背景可以从时代背景和理论背景两个层面进行阐述。时代背景关注的是现实社会中的重大议题，通常体现为重要政策文件、关键会议精神或领导讲话要点，这些因素往往决定了研究的现实紧迫性。理论背景则需要梳理学科发展的脉络，既要关注经典著作中的核心观点，也要把握近期学术热点和前沿动态，这样才能找准研究的理论定位。

研究意义则可分为学术意义和实践意义两个层面。学术意义主要体现在对学科理论体系的补充、完善或创新，能够推动学科知识边界的拓展。实践意义则强调研究成果对现实问题的解释力、预测力或解决能力，能够为政策制定、社会实践提供科学依据。

研究问题的形成是一个动态的思维过程，通常需要经历探索和评价两个关键阶段。在探索阶段，研究者需要通过多种途径激发问题意识：

* **系统地苦读文献**可以帮助研究者掌握学术脉络，发现已有研究的不足或空白。
* **深入体会现实**则要求研究者保持对实际问题的敏感度，从社会现象中提炼科学问题。
* **广泛与人交流**能够碰撞思想火花，不同学科背景的对话常常能催生创新视角。
* **学术顿悟**则是在长期积累基础上产生的灵感突破，往往能带来研究视角的重大转变。

在评价阶段，研究者需要建立严格的问题筛选标准，帮助我们逐步排除“灵光乍现”但“鸡肋”的问题。

* **真实性**问题要求研究问题必须来源于客观现实或理论矛盾，而非主观臆测。
* **重要性**标准衡量问题的学术价值和社会影响。
* **善意原则**强调研究应具有建设性，避免无意义的争议。
* **明确性**要求问题表述清晰准确，避免模糊不清。
* **聚焦性**意味着问题范围要适中，既不过于宽泛也不过于狭窄。
* **有趣性**标准关注问题的创新性和吸引力，能够激发研究热情。
* **可操作性**则评估研究问题在现有条件下的可行性。
* **简洁性**强调问题的表述要精炼，避免冗长复杂。

选择研究问题的策略主要有两种典型路径，研究者可根据自身条件和研究目标灵活选择：

* **扩散型问题**选择强调从核心问题向外拓展，适用于理论验证或应用推广类研究。这种方法需要严格遵循科学研究的规范流程，例如通过多案例比较来验证理论的普适性，或者采用不同方法进行稳健性检验以确保结论可靠。
* **漏斗型问题**选择则采取由面到点的思路，适用于复杂问题的逐步解析。这种方法的关键在于将宏大课题分解为若干可操作的子问题，通过层层递进的方式逐步深入。可以先界定核心概念，再分析影响因素，最后构建理论模型。这种化整为零的策略能够降低研究难度，提高可行性。两种方法各有优势，扩散型更注重结论的广泛适用性，漏斗型更强调研究的深度挖掘。

在实际研究中，两种思路往往需要交替使用，**初期可采用漏斗型方法聚焦核心问题，后期则可运用扩散型方法验证研究结论**。研究者应根据课题特点、自身专长和资源条件，选择最适合的问题切入路径。

## 理论构建

### 概念、构念和理论

本部分参考《政治学与公共管理研究方法基础》和《学术祛魅》

##### 概念

对客观现象的主观抽象，通过归纳同类事物的共同属性而形成，概念通常以字词或符号呈现，包含名词、内涵、指标和外延四要素，其中内涵的多维度需通过指标测量。清晰界定概念是理论建构的基础，其价值取决于三个因素：经验事实的支持、定义的精确性，以及与理论体系及其他概念的相关性——即该概念在理论解释中的必要性。例如《现代汉语词典》对"研究生"的定义即提炼了各类研究生的共性，如下：

大学本科毕业（或具有同等学力）后经过考试录取，在高等学校或科学研究机构学习、研究的学生。一般分为硕士研究生、博士研究生两级。有时特指硕士研究生。

##### 构念

在研究的概念化过程中，明确区分**概念（concept）**和**构念（construct）**至关重要。

**概念**是指我们试图理解的抽象对象或现象本身，通常是研究需要测量的核心目标。而**构念**则是为了研究和测量这个抽象概念，人为建立的结构化表达形式，可以理解为该概念的**多个维度或侧面**。尤其在社会科学的领域中，所研究的概念往往复杂且难以直接捉摸，远不如自然科学中许多可以直接观测的对象那么直观，因此更需要深入细致地剖析。

为了更精确地理解和使用构念，我们可以对其进行分类分析。一个关键的分野是区分**显在构念**和**潜在构念**。**显在构念**指那些相对外在、可直接观察或测量（如用具体仪器）的属性。而**潜在构念**则隐藏于内，无法直接观测，需要通过间接手段来衡量。有些潜在构念（如身高、体重）虽然存在于物理或生理层面，相对容易通过标准工具测量；然而，更多在社会科学中关心的潜在构念（如自尊、魅力、社会凝聚力等）则具有高度的主观性和抽象性，它们无法被直接“看到”或“摸到”，往往只能通过人们的**感知、陈述或行为表现**来推断。这使得这类构念带有一种“**只可意会，难以言传**”的特性。正是由于社会科学研究的核心对象多是这种复杂的潜在构念，研究者必须精心设计和运用各类问卷、量表、行为观察或其他策略化工具，才能有效地触及和测量我们真正关心的深层概念。

一句话总结便是：概念要想落地，就需要找到概念指涉事物本身的共性加以把握，构念就是操作化的前瞻，也是找寻变量的开端。

##### 理论

描述的是概念间关系，其定义有如下几个表述：

《现代汉语词典》强调理论的实践性、知识性和系统性；《美国传统辞典》侧重解释功能、可检验性和预测性；默顿关注逻辑命题和经验一致性；艾尔·巴比突出对事实规律的解释；巴卡拉克强调概念变量系统；特纳注重观念性和解释功能；范埃弗拉提出理论包含因果规律等要素。

综合来看，理论是人类基于实践或研究形成的系统性、可检验的知识体系，用于描述、解释和预测自然或社会现象，包含概念、变量、机制、命题等要素。

近年来较为“出圈”的中层（中观）理论（默顿），以及一些广为人知的宏观微观理论也分别有一些定义和解释：

* **宏观理论**：高度抽象复杂，解释世界本质和基本规律，如相对论、马克思主义、结构功能主义等。
* **中观理论**：介于两者之间，研究特定现象或群体，如公共资源治理理论、国家理论等。
* **微观理论**：聚焦具体概念关系，由可验证的命题构成，如少婚导致少子之类的简单判断。

检验理论离不开操作化的环节，需要明确的是，操作化的目的不是为了证明“先验”理论的正确，即不能带着答案找问题。

### 操作化

##### 变量（variable）

又称为变项，是对概念内涵操作化和具体化的产物。变量，顾名思义，即变化的量，是具有一个以上不同取值（不同的子范畴、不同的属性，或不同的亚概念）的概念。至于只有一个固定值的概念，则称为常量。举例而言，“性别”这一变量就包括了“男性”与“女性”两个不同的子范畴；“受教育水平”这一变量就可以分为“小学及以下” “初中” “高中” “大学本科或专科” “硕士研究生” “博士研究生”六个不同子范畴；工资收入更是有诸多取值可以填写。

在科学研究中常常使用变量语言及变量思维来描述或解释研究现象之间的关系。变量的使用是为了观测概念，亦是为了通过探寻变量间的关系，更便利地研究不同概念间的关系。将概念具体化为可以测量的变量，并说明不同变量以及变量的不同属性（属性就是事物的本性或特征）之间的逻辑关系，是建构理论的关键环节。

##### 变量间关系

在科学研究中，能够产生新的理论贡献是学术界一直以来的期望，这就要求研究者对变量的选择及其关系做出理论说明。在建立理论的过程中，变量之间并不是孤立的、割裂的，而是存在着某种逻辑关系，一般情况下存在以下三种关系来解释变量间的机制或原理：

* **直接关系**，即那些用箭头把两个概念直接连在一起的图形所表示的关系，如X→Y，这表示变量X与变量Y之间存在着直接效应。
* **间接关系（中介关系）**，即包含中介链的关系。比如变量X对变量Y的效应被认为是通过一个中间变量Z来产生的，我们就把变量Z称为中介变量，它们之间的关系表示为X→Z→Y。
* **调节关系**，指的是自变量和因变量的因果关系机制中包含有调节变量的关系。需要注意的是，调节变量调节的是一个关系，而不是一个变量，如果用图形表示，调节变量调节的是X→Y中间的“→”。



变量间关系

需要明确的是，并非只有量化分析才需要用到变量思维，当代的政治科学名著中，如亨廷顿《变化社会中的政治秩序》，尽管没有用明确的假设H1一类指出，更没有用到统计分析，却仍然有一个完整的链条，总结一下大约是：

* **自变**量：政治权威。
* **因变量**：政治发展。
* **中介变量**：政治秩序。
* **调节变量**：政党、领导人、社会结构等等。
* **控制变量**：一国的历史与现在（国家异质性）、国际体系变化等等。

##### 命题（proposition）

是“对于一个概念的特征或多个概念间关系的陈述”。例如，“受教育水平更高的人相较于受教育水平更低的人文化消费更多”。

##### 假设（hypothesis）

是命题的一种特殊形式，是为了得到逻辑的或经验的结论并加以检验而做的尝试性假说。假设通常是试探性的（也可以理解为实验性的），是为了被检验的目的而提出的严格的假设或建议。从这个角度来说，研究者通过提出待检验的假设，在抽象理论与经验事实之间建立了联结。

在大多数的社会科学研究中，由于研究者的主要目的大都可以归结为解释或说明不同变量之间的因果关系，因而，提出和检验有关变量间关系的假设成为科学研究和理论建构中不可或缺的环节。一般而言，假设有以下三种陈述方式：

* **条件式陈述（conditional statement）**：其表达形式为“如果x，则y”，其中x代表前提或先决条件，y为结果。这种方式通常表示两个变量之间存在因果关系，但有时也指代相关关系。
* **差异式陈述（differential statement）**：其表达形式为“A组与B组在变量x上无（或有）差异”。在统计学中，无差异假设即“零假设”或“虚无假设”。
* **函数式陈述（functional statement）**：其数学表达式为y=f(x)，即y是x的某种函数。它说明，如果变量x发生变化，则y也相应发生变化，反之亦然。函数式陈述在自然科学研究中较为常见，在以因果推断为方向的社会科学中也有增加的趋势。

## 具体方法

### 研究对象

##### 研究对象的选择与界定

在开展研究时，明确研究对象是至关重要的第一步。研究对象的核心在于确定**分析单位**，即研究的基本观测点。常见的分析单位包括：

1. **个体层面**：以个人为研究对象，例如研究"个人收入差异的影响因素"，分析教育程度、工作经验等对个体收入的影响。
2. **组织层面**：以企业、学校等组织为对象，比如探讨"企业创新效率的决定因素"，考察研发投入、人才结构等组织特征的影响。
3. **国家层面**：以国家或地区为分析单位，如比较"不同产业政策对经济增长的影响"，需要收集各国政策和经济数据。

需要警惕的是**生态学谬误**，即错误地将高层次的分析结论直接推论到低层次单位。例如，用国家层面的收入不平等数据来推断个体幸福感，就可能犯这种错误——相对剥夺理论的最初研究就踏入其中。

##### 数据类型

研究设计时还需考虑**时空维度**的选择：

* **截面数据**：捕捉同一时间点不同地区的状况，适合进行横向比较。例如2023年各省教育投入比较。
* **时序数据**：追踪同一地区不同时间的变化，适合趋势分析。如某省2000-2023年教育投入增长。
* **面板数据**：兼具截面和时序维度，能同时分析地区差异和时间变化。如全国31个省区市2010-2023年的教育投入数据。

### 获取与分析数据

找准所需数据的类型后，进入到数据分析方法的环节，以教育为主题为案例举例分析：

##### 观察法

指研究者通过系统观察并记录教学过程中的现象来收集数据。例如，要研究教师提问方式对学生课堂参与度的影响，研究者可以在课堂上记录教师提问的类型（如开放式问题或封闭式问题）以及学生的回应情况（如举手次数、回答质量）。这种方法适用于自然教学环境下的行为研究，但存在两个主要挑战：一是观察者可能因主观判断产生偏差，二是若观察的课堂样本不足或仅限于特定班级，可能导致结论缺乏普遍性。

##### 实验法

通过主动干预来验证教学方法的因果效应。典型操作是将学生随机分组，对实验组采用新教学方法（如项目式学习），对照组维持传统教学（如讲授式），学期末比较两组学生的成绩差异。例如，研究翻转课堂的效果时，随机分配部分班级使用课前视频学习+课堂讨论，另一部分采用常规授课，最终通过标准化测试评估学习成效。实验法的核心要求包括随机分组（确保学生基础水平均衡）、控制无关变量（如教材、课时一致），以及实验过程可重复验证。这种方法能直接证明教学方法的优劣，但可能因实验环境与真实课堂的差异而影响外部效度；在某些情况下，干预违背研究伦理而无法开展。

##### 准自然实验法

当无法随机分组然后实验时，可利用教育政策或学校制度变化作为“自然实验”条件。例如，某地区推行“小班化教学改革”，研究者可比较改革前后同一学校学生的成绩变化，或对比改革校与非改革校的差异，使用双重差分法（DID）排除生源波动等因素干扰。另一种情况是，利用学校按分数线分班的规则（如“实验班”与普通班的分数线临界点），采用断点回归（RDD）分析教学差异对成绩的影响。这种方法依托现实教育场景，但需注意其他变量（如家庭支持、教师经验）可能混淆结论。

### 总体与样本

本部分参考《政治学与公共管理方法基础》，在讨论具体方法之前，我们需要明确几个核心概念：

##### 要素

研究的基本分析单位，也是信息收集的对象，可以是一个人、一个班级、一所学校等。

##### 总体

指研究对象的全部要素的集合，比如"某市所有高中生"就是一个总体。

##### 样本

从总体中抽取的部分要素的集合，比如从该市随机选取的500名高中生。

##### 样本规模

又称样本容量，指的是样本中包含的要素数量，比如上述的500人。

##### 样本代表性

指样本在多大程度上能够反映总体的特征。一般来说，当样本的关键特征（如性别比例、成绩分布等）与总体越接近，代表性就越好。

##### 抽样

无论选择哪种研究方法（质性研究还是量化研究），都要求我们从总体中抽取一定样本（否则真要累死），那具体怎么抽呢？抽样就是从研究总体中选择部分个体或要素作为样本的过程。要完成抽样，首先需要明确**抽样框**——也就是包含所有研究对象的清单或规则（比如全校学生名单），同时确定**抽样单位**，即实际抽取的基本单位（比如个人、班级或学校）。

抽样方法主要分为两大类：**非概率抽样**和**概率抽样**。

**非概率抽样**不依赖随机原则，而是根据研究需求、主观判断或现实条件来选取样本，适合探索性研究或特殊案例分析。常见方法包括：

* **方便抽样**：选最容易获取的样本（比如就近调查自己班的学生）；
* **目标抽样**：按特定标准选取（比如只研究成绩前10%的学生）；
* **滚雪球抽样**：通过已有样本推荐新样本（比如通过少数留学生找到更多留学生受访者）；
* **配额抽样**：按比例匹配总体特征（比如男女各抽50人）；
* **典型案例抽样**（选最具代表性的案例）、**极端案例抽样**（选表现异常好或差的个体）、**负面案例抽样**（故意选不符合理论的案例）等。

**概率抽样**则严格遵循随机原则，确保每个个体有已知的、非零的被抽中概率，适合需要统计推断的研究。常用方法有：

* **简单随机抽样**：像抽签一样完全随机选取；
* **系统抽样**：按固定间隔抽取（如每第10个学生）；
* **分层抽样**：先按特征分组（如年级），再在各层内随机抽；
* **整群抽样**：随机抽取完整群体（如抽几个班级而非个别学生）；
* **多阶抽样**：分多个阶段逐步缩小范围（如先抽城市，再抽学校，最后抽班级）。

##### 抽样误差

抽样过程不可避免地会产生**误差**，即样本统计量（如样本平均分）与总体参数（如全市高中生真实平均分）之间的差异。这种误差是随机产生的（**无法根除**），但可以通过科学的抽样方法和适当的样本规模来降低（**但不能躺平**）。

选择抽样方法时，非概率抽样灵活但可能偏差大，概率抽样科学但成本高。实际研究中常结合使用，比如先滚雪球找到目标群体，再用分层抽样确保代表性。

### 得出结论

本部分参考《政治学与公共管理方法基础》、《比较政治学：理论与方法》和《比较政治中的议题与方法》，为了确保我们得出的是一个流程科学结果合理的结论，我们还需要反复考量如下要求：

##### 控制变异

在社会科学研究中，控制变异是确保研究结论可靠性的关键环节。研究者需要通过系统的方法来最大化研究变量的差异，控制干扰因素，并减少测量误差。

1. **最大化系统差异**：指由于自变量变化而引起的因变量的变化幅度要达到最大（变化越大越能找到规律）。明确研究问题，确定最佳的自变量和因变量，同时应该考虑到自变量和因变量的所有情况，并选定合理的样本范围。
2. **控制外生变量** ：控制外生变量主要有三种方法：
   * **纳入法**：将可能影响结果的重要变量纳入分析模型。**核心变量**（如政策实施年份、受教育年限）必须包含，同时要控制**人口统计学变量**（如年龄、性别）和**环境变量**（如行业类型、宏观变量等）。
   * **非纳入法**：通过研究设计来控制变量。**同质法**即没有外生变量不一致导致的误差（一般这样只能找到有限个特征相同，不能保证完全相同，用得很少)；**随机分配法**确保**实验组**和**对照组**在各方面均衡（概率抽样）；**配对法**则根据相似特征（如经济发展水平、人口规模）来匹配比较对象。
3. **最小化误差差异**：测量误差的控制至关重要。采用标准化测量工具（如Likert五级量表）可以提高数据一致性。同时，通过多源数据交叉验证，也可以显著提升数据的准确性和可靠性。

##### 研究效度

* **构念效度**：指一个构念能够正确反映其所要表达、描述或测量对象内容和特征的程度，或者也可以定义为一个构念表达、描述和测量的精确性。
* **内部效度**：指基于特定研究对象而得出的结论本身符合特定研究对象实际情况的程度；而对探求精确因果性关系的定量研究而言，则是基于特定研究对象得出的变量(或构念)间因果关系的推论符合特定研究对象实际情况的程度，也即其因果关系推论对特定对象本身而言的可信度。
* **外部效度**：指将从特定研究对象得出的研究结论推广到具有不同分析单位、时间维度、研究区域、研究层次、研究尺度等的其他研究对象的可信度。
* **统计结论效度**：指通过统计检验对假设关系进行解释的可信度。

需要说明的是，这四种效度常互相制衡（如提高内部效度可能损害外部效度）；而追求统计结论效度实际上也可能犯了构念效度的损伤（试变量试到跑题）。

## 定性与定量的融合

上文虽然是定量分析的范式，但定性亦可参照一二，诸多时候定量分析的理论渊源实际上来自定性，更遑论KKV（加里·金/Gary King、罗伯特·基欧汉/Robert O. Keohane、悉尼·维巴/Sidney Verba）三人经典的《*Designing Social Inquiry*》中早已有详尽的解释（其繁体版译本名为《好研究如何設計？：用量化邏輯做質化研究》）。

对这一部分的理解有助于我们理解计量经济学的开端、发展与变革，从而更好理解结构计量的设计转向。值得注意的是，KKV倡导的"量化逻辑"并非要将质性研究简单数字化，而是强调研究设计中的因果推断逻辑必须严谨。这种思想与Angrist等人提出的"可信性革命"遥相呼应，二者都要求研究者超越方法论的二元对立，在保持方法适切性的前提下实现优势互补。当处理异质性处理效应时，定性访谈揭示的个体决策过程往往能修正定量模型的设定偏误；而大样本数据分析又能检验质性研究发现的外部效度。

99 董晨宇  等，《拆论文：从精读中领悟学术写作》，重庆大学出版社

杨立华  等，《政治学与公共管理研究方法基础》，北京大学出版社

马亮，《学术祛魅：实证研究十讲》，中国人民大学出版社

王正绪  等，《比较政治学》，复旦大学出版社

B.盖伊·彼得斯，《比较政治学：理论与方法》，上海人民出版社

托德·兰德曼  等，《比较政治中的议题与方法》，上海人民出版社

# 定量分析的数理基础

## 数学

定量研究对于不少文科专业的同学来说最关键的一个门槛便是理解数学公式，好在笔者本人也是文科专业出身，写这一部分足够契合（水平也菜）。

本文介绍了因果推断中常用的数学基础，包括微积分、线性代数、概率与分布、以及统计推断等内容。这些基础知识为理解和应用因果推断方法提供了必要的数学工具。

### 微积分

##### 函数

在数学中，函数是一种特殊的关系，它将一个集合中的每个元素（**自变量**）映射到另一个集合中的唯一元素（**因变量**）。函数可以用多种方式表示，包括解析式、图像、表格和算法等。函数的概念在微积分中至关重要，因为微积分的主要研究对象就是函数的变化率和累积量。

##### 导数

一元函数 的一阶导数定义为：

其几何意义是函数在某点的切线斜率。二阶导数定义为：

二阶导数反映了函数图像切线斜率的变化率，描述了曲线的弯曲程度，这一性质在数学中被称为**曲率（curvature）**。在经济学中，我们一般默认二阶导数恒为正或负值。



函数 与其一阶、二阶导数

##### 偏导数

偏导数是多元函数对其中一个变量的导数，而将其他变量视为常数。例如，对于二元函数 ，其对 的偏导数定义为：

类似地，对 的偏导数为：

偏导数在经济学中有着广泛的应用，例如在效用函数和生产函数中，偏导数分别表示商品和生产要素的边际效应。

##### 基本求导规则

* 线性法则：, 其中为常数；
* 乘法法则：；
* 除法法则：；
* 链式法则：。

常见函数的导数公式

| 函数 | 导数 |
| --- | --- |
| 常数 |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

##### 洛必达法则

洛必达法则的核心思想是通过分子分母分别求导，解决或型不定式极限问题：

适用条件包括：极限形式为或，在的去心邻域内可导且，求导后极限存在或为无穷。

证明

**证明思路**：通过取对数转换，构造型极限，应用洛必达法则求解。

**证明步骤：**

1. 设 ，取对数得
2. 令 ，转化为
3. 应用洛必达法则：
4. 还原得

##### 一元最优化

最小二乘法（OLS）和最大似然估计（MLE）都是最优化问题。无约束的一元最优化问题中，最大化问题的一阶条件为，二阶条件为；最小化问题的一阶条件同样为，二阶条件为。

##### 多元最优化

考虑无约束的多元最大化问题：

其一阶条件要求在最优值 处，所有偏导数均为0：

该条件称为**一阶必要条件**，在满足一阶条件的基础上，其充分条件为：

##### Hessian条件

记函数的Hessian矩阵为：

* 若 是**负定矩阵**，则 是局部极大值点。
* 若 是**正定矩阵**，则 是局部极小值点。
* 若 **半正定或半负定**，则可能为鞍点或边界情况。

其中，负定和正定矩阵的判定方法可通过特征值符号判断。



一个抽象的多元最优化示意图

### 线性代数

##### 矩阵

将 个实数排列成如下矩形阵列：

称 为 级矩阵（matrix），其中：

* ：行数（row dimension）
* ：列数（column dimension）

元素 表示第 行、第 列的元素。记法： 或者 。若所有元素均为 0，则称为零矩阵（zero matrix），记作 。

##### 矩阵的转置

将矩阵 的第 行变为第 列，得到其转置矩阵 ，维度为 ：

* 若 是对称矩阵，则
* 转置的转置还原：

##### 方阵

当 时，矩阵 称为 阶方阵：

主对角线元素为。

常见类型有：

* 对称矩阵：。
* 对角矩阵：非主对角线元素全为 0。
* 单位矩阵 ：主对角线为 1，其余为 0。

上述三种类型的矩阵限制条件逐项增加，可以看到单位矩阵一定是对角矩阵和对称矩阵；单位矩阵 在矩阵运算中作用类似于数字 1。

##### 行列式

**定义**：方阵的一个标量值函数

* 记法： 或
* 对于 矩阵 ：
* 其中：
  + 是 的所有排列集合
  + 是排列符号函数（偶排列为 ，奇排列为 ）

##### 行列式的求解

**定义：**

* 排列中“逆序对”的数量
* 举例：123（1<2<3，0），213（2>1<3，1），321（3>2>1，3）

**三阶行列式展开：**

* : 的排列
* : 排列的逆序数
* 共6项（例如：排列 对应 ）

##### n阶行列式

* : 到 的排列
* : 排列的逆序数
* 总共 项（如：3阶有6项，4阶有24项）

设矩阵：

行列式展开：

计算得：

所以：

##### 向量

**定义**：若 ，矩阵 称为 维行向量（row vector）；若 ，则称为 维列向量（column vector）。向量是矩阵的特例。

设列向量：

它们的内积（点乘）定义为：

**向量正交性**

**定义**：若 ，称 与 正交，即在 维空间中相互垂直。任何形如 的乘积求和，都可以很方便地写为向量内积 的形式。特别地，平方和 可写为 ：

##### 矩阵加法与数乘

**矩阵加法**：若 和 是同维矩阵，则：

**数乘**：实数 与矩阵 的数乘为：

矩阵的加法满足以下规则：

1. (加上零矩阵不改变矩阵)
2. (加法交换律)
3. (加法结合律)
4. (转置为线性运算)

##### 矩阵乘法

设 为 矩阵， 为 矩阵，则乘积 的元素定义为：

**注意**：矩阵乘法不满足交换律（）。

矩阵的乘法满足以下规则：

1. , (乘以单位矩阵不改变矩阵)
2. (乘法结合律)
3. (乘法分配律)
4. , (转置与乘积的混合运算)

##### 线性方程组

**定义**：考虑以下由 个方程、 个未知数构成的线性方程组

其中， 为未知数，可将上式写为：

直观上，如果可将此方程左边的方阵 “除” 到右边去，则可得到 的解。为此，引入**逆矩阵**的概念。

##### 逆矩阵

**定义**：设 为 阶方阵。若存在方阵 ，使得

则称 可逆（或非退化）， 是其逆矩阵，记作 。 可逆的充要条件是 ，且逆矩阵唯一。假设方程中的矩阵 可逆，则在该方程两边同时左乘其逆矩阵 可得：

矩阵求逆满足以下规则：

1. (求逆与转置可交换次序)
2. (求逆与乘积的混合运算)

##### 矩阵的秩

向量组的线性相关性

* 对于两个维列向量与：若是的固定倍数，则向量组中只有一个向量真正含有信息。
* 个维向量，若存在不全为零的标量使得：

则称向量组线性相关（linearly dependent）。若仅当时上式成立，则称线性无关（linearly independent）

向量组的秩：极大线性无关部分组所含向量的个数。

* 列秩（column rank）：列向量组的秩
* 行秩（row rank）：行向量组的秩
* 重要性质：行秩 = 列秩，统称为矩阵的秩（matrix rank）
* 满列秩：若矩阵的列秩等于

##### 行阶梯形法

： 通过初等行变换将矩阵化为行阶梯形，非零行数即为秩。

计算矩阵 的秩

**步骤 1：消除第一列下方元素**

初始矩阵：

执行初等行变换：

结果为：

**步骤 2：消除第二列下方元素**

继续对当前矩阵进行变换：

执行变换：

得到行阶梯形矩阵：

非零行有 2 行，因此 。

##### 多重共线性与研究设计（KKV）

**定义**：一个解释变量可被其他变量预测，影响对各自独立效应的识别（不依赖线性假设）。

**问题实例**（军备合作）：

* 假设1：大小不同的国家更易合作。
* 假设2：不相邻国家更易合作。
* 若观察值仅限于“大小不同且相邻”或“大小相同但不相邻”，则无法区分两个变量的影响。

**解决方案**：

* 收集更多理论相关案例（如大小相同又邻近的国家）。
* 或从其他分析层次寻找可观测意涵。

##### 二次型与正定性

对于对称矩阵 ，二次型定义为：

**分类**：

**应用**：计量经济学中常用 度量标准化距离。

## 统计学

### 概率与分布

##### 无条件概率

**定义**：无条件概率是概率论中最基础的概念之一，表示事件发生的概率，记作。例如，我们可以研究股市崩盘的概率。无条件概率的特点是独立于其他事件，即不考虑任何其他事件的影响。

##### 条件概率

条件概率的数学定义为：

其中表示事件和同时发生的情况（例如“太阳雨”事件）。一个典型的应用场景是：已知经济衰退时，计算股市崩盘的概率。

##### 独立事件

对于两个事件和，以下两个条件是等价的：

当这些条件满足时，我们称事件和是相互独立的。

##### 全概率公式

对于一组互斥且完备的事件（其中），全概率公式可以表示为：

这个公式在概率计算中非常有用，特别是在事件可以分解为多个互斥情况时。

##### 离散型概率分布

离散型随机变量的取值与概率可以用如下表格表示：

其中且。常见的离散型分布包括：

* Bernoulli分布（两点分布）
* Binomial分布（二项分布）
* Poisson分布（泊松分布）

##### 概率密度函数

连续型随机变量的概率密度函数(PDF)满足：

##### 区间概率

随机变量落在区间内的概率为：

##### 累积分布函数

累积分布函数(CDF)定义为：

##### 期望

离散型随机变量的期望：

连续型随机变量的期望：

##### 方差

方差的定义：

##### 协方差

协方差的定义：

##### 相关系数

相关系数的定义：

相关系数具有以下性质：。

##### 条件密度函数

条件密度函数的定义：

##### 条件期望

条件期望的定义：

##### 正态分布

若随机变量的概率密度函数为：

则称服从正态分布，记作，其中为期望，为方差。

性质

* 对称性：关于对称。
* 峰度：超额峰度为0。
* 线性变换：。

应用：正态分布是中心极限定理的基础，广泛用于误差建模和假设检验。

##### 卡方分布

若，则：

其中为自由度。

性质

* 期望：。
* 方差：。
* 可加性：独立卡方变量之和仍为卡方分布。

应用：卡方分布在方差分析和卡方检验中有重要应用。

##### t分布

若与独立，则：

性质

* 对称性：关于0对称。
* 厚尾：时尾部显著厚于正态。
* 收敛性：时。

##### F分布

若与独立，则：

性质

* 非对称：仅取非负值。
* 倒数关系：。
* 与t分布关联：。

应用：F分布在方差齐性检验和ANOVA中有重要应用。

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

常见概率分布的比较

### 统计推断的核心概念

##### 总体与样本

* 总体(Population)：研究对象的全体。
* 样本(Sample)：从总体抽取的部分个体（容量）。
* 理想样本：独立同分布(iid)随机样本。

##### 参数估计

对于参数，其估计量：

* 样本统计量（随机变量）。
* 估计值：给定样本时的具体取值。

常见估计量包括：

##### 估计量的评价标准

偏差的定义：

当时，称为无偏估计(unbias estimator)，反之则为有偏估计(bias estimator)。

##### 均方误差(MSE)分解

均方误差可以分解为方差与偏差平方之和：

证明过程：

其中交叉项为0，因为：

##### 偏差-方差权衡

统计学的核心命题是均方误差可以分解为：

极端案例对比：

* 无偏高方差估计量（如高阶多项式回归，易过拟合）
* 有偏低方差估计量（如一次回归，可能欠拟合）



拟合示例图

## 社会科学中的统计推断

### 描述变量

如何描述变量。这看似是一个简单的目标，但其内涵远比表面深刻。第二章讨论了如何建立研究问题，以及实证研究如何帮助我们理解世界，本章的前半部分回顾（预习）了基本的数学知识，此刻我们讨论变量描述。

实证研究问题本质上可归结为描述统计变量的密度分布——定量实证研究的核心就在于此。所有有趣的实证研究发现——无论是物理学、社会学、生物学、医学、经济学、政治学等领域——都有一条共同的主线。这条主线精巧地建立在概率基础之上，其形态表现为统计变量的密度分布，将整个宇宙中所有实证知识永恒地联系在一起。

为了理解数据，我们必须掌握如何观察并描述它们。这需要描述变量的类型及其分布形态。本节专注于单变量描述（后一节将讨论变量间的交互关系）。虽然单变量描述看似不如多变量分析有趣，但其重要性不容忽视。

在实证研究中，**变量**指对同一测量指标的一系列观测结果：如433名上海人的月收入、1999-2025年中国每年的企业并购数量、744名儿童的神经质心理评分、532朵花的颜色、《人民日报》上世纪某段时间内的头条新闻标题等。成功描述一个变量意味着能够清晰解释观测结果，而无需让他人查看所有原始数据。

描述变量的第一步是确定其类型。最常见的**变量类型**包括：

##### 连续变量（Continuous Variables）

连续变量可在一定范围内取任意值。例如上海人的月收入可以是12,000元，也可以是20,000元，或介于两者之间的任何值。这类变量没有“下一个最高”的概念，因为其变化是连续的。

##### 计数变量（Count Variables）

计数变量记录事件发生的次数或物体的数量。例如中国每年的企业并购数量。计数变量不能为负值，也不能取分数值。当计数变量取值较多时，其性质接近连续变量，常被当作连续变量处理。

##### 有序变量（Ordinal Variables）

有序变量的取值有“多”“少”之分，但缺乏精确的量化标准。例如神经质评分的“低”“中”“高”三个等级。虽然“高”高于“低”，但高多少并不明确。教育程度的“小学”“中学”“高中”“大学”也是有序变量——完成高中教育意味着比初中教育程度更高，但“高多少”无法精确量化。

##### 分类变量（Categorical Variables）

分类变量记录观测对象所属的类别。例如花的颜色（白、橙、红）。这些类别没有高低之分，仅是不同类别。在社会科学研究中极为常见（如宗教信仰、种族、地理位置等）。**二元变量**是分类变量的特例，仅有两个取值（如“是/否”）。其优势在于便于处理实验效应（是否接受治疗），且多分类变量可转化为多个二元变量（如将政治面貌转化为“是否团员”“是否党员”等）。

##### 定性变量（Qualitative Variables）

定性变量是非数值型、非分类的变量集合。如《人民日报》标题文本。这类变量通常包含大量难以概括的细节，常需转化为其他变量类型进行分析（如统计标题提及女性政治人物的次数，或使用AI模型进行特征评分）。

确定变量类型后，下一步是考察其**分布**——描述变量各取值出现概率的特征。

##### 分布描述

对于分类或有序变量，可通过给出各类别/取值的百分比来描述分布。完整分布可通过频率表或条形图展示。

美国高校学位类型分布

| 授予的主要学位类型 | 数量（N） | 百分比（%） |
| --- | --- | --- |
| 两年制以下学位 | 3495 | 47 |
| 两年制学位 | 1647 | 22 |
| 四年制及以上学位 | 2282 | 31 |
| 总计 | 7424 | 100 |



美国高校毕业生平均收入分布图

连续变量的分布描述更为复杂。由于观测值重复概率低，需采用**直方图**（将数据范围划分为区间并显示各区间的观测比例）或**密度图**（显示当直方图区间无限变窄时的极限形态）。密度曲线下面积代表概率，如图[3.2](#fig:earnings)所示，16%的观测值位于40,000至50,000美元区间。

当完整分布信息量过大时（尤其对连续变量），需用少量数字概括分布特征。最经典的例子是**均值**——将所有观测值相加后除以观测数。均值代表数据的中心趋势，试图生成一个代表性数值。

##### 百分位数（Percentiles）

**X百分位数**指X%的观测值小于该值的点。例如身高数据的第5百分位数为5英尺4英寸。中位数（第50百分位数）是另一种中心趋势度量，代表典型观测值（对偏斜数据更稳健）。最小值和最大值（第0和100百分位数）显示变量的取值范围，其差值称为**极差**。

##### 变异度量

变量间变异程度不同（如子女人数变异较大，而眼睛数量变异较小）。**方差**是基于均值的变异度量：

1. 计算均值
2. 各观测值减均值（得离均差）
3. 离均差平方
4. 求平方和
5. 除以观测数减1（得样本方差）

为消除平方单位影响，常取方差的平方根得**标准差**（恢复原始单位）。标准差有助于评估观测值的相对位置（如收入38,000美元者比均值高37.6%个标准差）。

##### 四分位距（IQR）

是第75与25百分位数之差，覆盖样本中50%最接近中位数的观测值，对极端值不敏感。

##### 偏度（Skew）

偏度描述分布向一侧倾斜的程度。例如年收入分布有长右尾（少数极高收入者），称为**右偏**。左偏分布则有长左尾。对称分布则两侧尾部相近。

处理偏斜数据的常用方法是对数转换（使数据更接近正态分布）。自然对数转换的优势在于系数解释直观（对数增加0.01约对应原始变量1%的增长）。

统计学严格区分**真相**（理论分布）与**数据**（观测分布）。我们真正关心的是理论分布（生成数据的分布），而非样本统计量。随着样本量增加，观测分布会趋近理论分布。

##### 常见理论分布

* **正态分布**：对称分布（如身高、智商）。即使严格意义上不适用（如身高不能为负），但当近似良好时仍被广泛采用。
* **对数正态分布**：取对数后呈正态分布的右偏分布（如收入、财富、公司规模等）。其优势在于可通过对数转换消除偏度。

##### 从数据推断理论分布

我们可通过假设检验评估理论分布的可能性：

1. 选择理论分布的某个特征（如均值）
2. 基于理论分布性质和样本量，确定该特征在随机样本中的抽样分布
3. 计算观测数据的对应特征
4. 评估观测结果在理论分布中的出现概率
5. 若概率极低，则拒绝原理论分布假设

篮球运动员得分数据样本中有100个数据，得分均值为102，标准差为30（，，），假设理论均值成立吗？

1. 抽样分布的推导

根据中心极限定理（Central Limit Theorem, CLT），当样本大小 足够大时（一般 ），样本均值的抽样分布近似正态分布：

* 分布的均值等于总体均值（在 下，）。
* 分布的标准差（称为标准误差，) 由公式 计算。

给定数据：n = 100, s = 30

计算标准误差：

因此，在 下，样本均值的抽样分布是正态分布 （即均值为 90，标准差为 3）。这可以写为：

（注： 表示正态分布，其中第二个参数是方差；标准差为 ，方差为 。在推断中，通常直接使用标准差。）

2. 检验统计量（z-分数）的计算

观测样本均值 。我们计算其与假设均值 的差异，并以标准误差为单位标准化：

z-分数为 4，表示观测均值 与假设均值 相差 4 个标准误差（即 4 个标准差单位）。这反映了在 下，观测值的相对位置。

3. p-值的解释

p-值是在 成立时，观测到与样本均值一样极端或更极端结果的概率。这是一个双尾检验（因为 ），因此：

在标准正态分布下：

* （右尾概率）。
* 双尾 p-值 。

因此，p-值远小于 0.0001（即 ）。这表明，如果 为真（），观测到样本均值 （或更极端）的概率极低（小于万分之一），属于统计学上的极端事件。

4. 结论

给定显著性水平 （常用阈值）：

因此，我们有非常强的证据拒绝原假设 。结论是：篮球运动员的平均得分不太可能是 90，更可能高于此值。

### 描述关系

对于大多数研究问题，我们不仅对单个变量的分布感兴趣，更关注数据中两个或多个变量之间的关系。两个变量之间存在关系意味着什么？这种关系显示了了解一个变量能为我们提供关于另一个变量的信息。

例如儿童的身高与年龄。通常，孩子年龄越大，身高越高。因此知道一个孩子13岁而另一个6岁，就能较好地猜测谁更高。我们可以称身高与年龄的关系为“正相关”——一个变量的较高值往往对应另一个变量的较高值（年龄越大身高越高）。也存在“负相关”关系（年龄越大哭闹越少），以及“零相关”关系（儿童年龄与居住可能性无关）。其他关系可能时正时负，或初期强正相关后期弱正相关。若变量是分类变量，则不存在“高低”之分，只有“差异”（大孩子更可能骑自行车上学）。

##### 条件分布

上一章讨论的是变量的无条件分布（也称边缘分布）。条件分布则是指给定另一个变量特定值时，某个变量的分布。

以条件概率为例：随机一个人是女性的概率约50%，但名叫**翠花**的人是女性的概率远高于50%。我们称之为“在名为翠花的条件下是女性的概率”。条件分布同理，不过是从单一概率扩展到整个分布的变化。

图[3.3](#fig:vitaminE)展示了服用维生素E剂量的分布，按是否进行剧烈运动分组。可见运动者与非运动者的分布存在差异——运动者服用剂量更大。这种分布差异表明维生素E摄入与运动存在关联。



上月运动情况对维生素E摄入量的影响

该概念同样适用于分类变量。图[3.4](#fig:vitaminEbarchart)显示：不吸烟人群服用维生素E的比例明显高于吸烟人群，这与Oster关于健康意识驱动行为的假设一致。



是否吸烟与是否服用维生素E的关系

##### 条件均值

掌握条件分布概念后，我们可计算该分布的任何特征。本章重点讨论条件均值——给定变量的特定值，变量的期望均值。

对于离散变量，计算简单：只需取该值所有观测的均值。图[3.5](#fig:VitaminE_by_BMI)显示维生素E服用比例随时间阶段（推荐前/中/后）的变化：推荐期间呈现正相关，推荐取消后转为负相关。



按体重指数值范围分列的服用维生素E的比例

连续变量则更复杂。解决方法有两种：

1. 将连续变量划分为若干区间（分箱），计算各区间内均值
2. 使用局部均值或LOESS曲线（局部加权散点平滑）进行平滑处理

图[3.6](#fig:LOESS_Curve)的LOESS曲线清晰显示：BMI越高，维生素E服用率越高，初期关系强烈后期趋于平缓。这种方法通过移动窗口计算局部均值，避免分箱法的任意性缺陷。



按体重指数(BMI)划分的服用维生素 E 的比例(LOESS)曲线

##### 回归分析

另一种方法是假设变量间存在某种数学关系（通常是直线），即回归分析。图[3.7](#fig:VitaminE_BMI_linear)用直线拟合维生素E与BMI的关系，斜率0.002表示BMI每增加1单位，服用概率上升0.2个百分点。



用拟合直线表示按体重指数计算的服用维生素E的比例

普通最小二乘法（OLS）是最常用的直线拟合方法，通过最小化残差平方和确定最佳直线。其斜率计算公式为协方差除以的方差，直观反映了“的变化中有多少与协同变化”。

当直线拟合不足时，可采用：

* 多项式回归（如二次曲线）
* 非线性转换（如对数、概率单位转换）

##### 条件均值（“控制变量”）

通过回归残差，我们可以剥离已解释部分，专注于未被解释的变异。将这个方法应用于双变量，就能实现“控制变量”——消除对和的影响后，分析二者的纯净关系。

多元回归是最高效的实现方式，在仅考虑BMI时，我们估计：

现在加入BMI、性别和年龄后，得到：

在控制年龄和性别后，BMI的效应从0.002降至0.001，说明原关系部分由这些变量解释。同时，年龄每增加1岁服用概率提高0.2个百分点，女性比男性高1.6个百分点。

这种方法在数学上对应Frisch-Waugh-Lovell定理：通过回归残差间的回归，等效于直接进行多元回归。这让我们能清晰分离不同变量的独立影响。

99 陈强，《计量经济学及Stata应用(第2版)》，高等教育出版社

Nick Huntington-Klein，《The Effect》，CRC Press

# 因果推断的源流原理

## 计量与因果

### 计量经济学的历史与发展

1930年代，以计量经济学会成立为标志，经济学中的实证方法成为一个独立学科出现。

在计量经济学内部又可以分为两大部门，1950-60年代，宏观计量大发展，Klein、Stone等学者推动了以结构计量或模型为基础的计量经济学(Model-based Econometrics)。研究者首先先验地定好各变量之间的函数关系，再代入当下的宏观变量进行验证和改进，由于石油危机的出现，路径依赖下的结构计量遭遇滑铁卢，断点的产生使得精巧的模型与复杂的现实间的宏观难以逾越，更多人选择抛弃结构而转向设计。

潜在结果框架（Potential Outcome Framework）的发展可以分为以下几个阶段：

1. 生物学实验设计的萌芽：早期应用可追溯至18世纪林德医生（James Lind）的坏血病对照实验。他通过将船员分组并给予不同治疗方案（如柑橘类水果），首次以实验设计验证了因果效应。这一案例虽未形式化潜在结果概念，但体现了“反事实比较”的核心思想。
2. 统计学的理论奠基：20世纪20-30年代，Neyman在随机化试验中首次提出潜在结果的数学表述，但直到1990年其论文英译后才广为人知。Rubin(1974)将潜在结果扩展至观察性数据，正式建立Rubin因果模型，解决了非随机实验中的因果识别问题。

即使在结构计量之中，潜在结果的概念也已经有所体现。挪威经济学家 Trygve Haavelmo 在1943年研究联立方程模型（SEMs）时，探讨了供求模型中的识别问题（identification）。他区分了供求函数中的“任何想象的价格 π”和“实际观察到的均衡价格 p”，并将实际价格下的供求量视为特定条件下的实现结果。这一区分隐含了潜在结果的思想。因此，Haavelmo 被视为计量经济学概率论基础的奠基人之一（Imbens & Wooldridge, 2009；Rubin, 2008）。

当然，计量经济学后续的发展并没有沿着这一方向继续深入。到了20世纪中叶，主流方法转向直接对观测结果建模，将潜在结果与分配机制混杂在一起，导致以下问题：

1. 因果效应识别变得难以理解；
2. 需要引入越来越多的强假设；
3. 模型对设定误设（misspecification）非常敏感。

Leamer (1983) 指出："计量经济学的艺术就是，研究者在计算机终端中拟合许多（甚至上千个）统计模型，从中选择一个或几个符合作者预期的估计结果在论文中进行报告。"他发现"我们正处于一种令人沮丧和不科学的境地。没有人将数据分析看作严肃的事情，或者更准确地，没有人把别人的数据分析当回事。"他提议将计量经济学中的"谎言和欺骗"剔除出来，并提出了"敏感性分析"(sensitivity analysis)作为解决方案。

LaLonde (1986) 利用美国1970年代进行的一项就业培训的随机化实验数据，考察传统计量经济学方法是否能够模拟随机化实验的结果。他以随机化实验作为基准(benchmark)，利用观测数据作为控制组，运用回归、固定效应、Heckman选择模型等计量经济学常用方法估计了培训对收入的影响，发现这些方法都无法复制随机化实验的结果。这一研究得出了在观察研究中计量经济学方法无法可信地估计因果效应的悲观结论。

Dehejia和Wahba (1999) 利用倾向指数匹配方法重新考察了LaLonde (1986)探讨的问题。他们发现，尽管LaLonde尽量通过手工方式使两组个体相似，但构造的控制组个体与实验中的干预组个体仍然具有较大的特征差异。通过倾向指数匹配方法，他们获得了与干预组更为相似的控制组，发现估计的因果效应与随机化实验的结果非常相似。

1990年代左右，Card (1990)、Angrist (1990)、Card和Krueger (1994)、Angrist、Imbens和Rubin (鲁宾，1996)、Angrist和Lavy (1999)、Abadie等 (2003, 2010)的一系列研究，将经济学实证研究重新回到科学发展的开端，使内部有效性更好，实证结果更可信，从而引发了一场经济学经验研究的"可信性革命"(Angrist和Pischke，2010)或"自然实验革命"(Imbens，2010)。

2019年，Abhijit Banerjee、Esther Duflo和Michael Kremer因利用随机化实验研究全球减贫问题而获得诺贝尔经济学奖；2021年，David Card因利用自然实验研究劳动经济学问题、Joshua Angrist和Guido Imbens因因果推断方法论贡献而获得诺贝尔经济学奖。这些发展标志着基于设计的计量范式(Design-based Econometrics)的兴起。

当下的经济学经验研究正在经历了一场深刻的“可信性革命”，这场革命的核心在于从根本上改变了研究者进行实证分析的方法论取向。这场革命主要包含两个关键内容：首先是更加注重因果推断而非简单的相关性分析，其次是特别强调研究设计在实证分析中的核心地位。

具体而言，可信性革命引入了潜在结果框架（Potential Outcomes Framework），为因果关系的定义和识别提供了严谨的理论基础。在这一框架下，研究者致力于通过巧妙的研究设计，使观测性研究能够尽可能模拟随机化实验的理想条件。这种转变使得经济学实证研究能够获得更接近真实因果效应的估计结果。

“设计”是这场革命的核心所在。传统的估计方法如普通最小二乘法（OLS）、最大似然估计（MLE）和广义矩估计（GMM）本质上都只是统计工具，它们本身并不能保证因果关系的识别。相比之下，匹配方法（Matching）、工具变量法（IV）、双重差分法（DID）和合成控制法（Synthetic Control）等新兴方法则更加注重研究设计环节。这些方法通过精心设计的研究方案，模拟随机化实验的条件，在完成设计后，即使是简单的OLS回归也能得到可信的因果效应估计。

这种以设计为导向的研究范式主要致力于解决实证分析中的两个关键问题：

1. 消除混杂因素导致的偏差（Confounding Bias）
2. 样本选择偏差（Selection Bias）。

通过这种转变，经济学实证研究的内部效度得到了显著提升，研究结果的政策指导意义也大大增强。正如Angrist和Pischke（2010）所强调的，这场革命使得经济学实证研究重新回归科学方法的本源，为学科发展注入了新的活力。

至此，计量经济学家重新认识到潜在结果框架的重要性，并推动了一场被称为“可信性革命”的学术运动。这场革命强调因果推断的严谨性和可验证性，代表人物包括：Guido Imbens（斯坦福大学）；Joshua Angrist（麻省理工学院）；Alberto Abadie（MIT）；David Card（加州大学伯克利分校）。他们倡导以实验设计的思维方式来处理因果推断问题，被广泛称为计量经济学的“实验学派”。

### 因果关系与求异法

相关不等于因果，这成了大众普遍的认知，然而在实验设计为导向的现代计量当中，由于事先的分析，相关终于是为因果开辟了道路，为了更好理解实验设计，我们有必要首先习得因果关系如何识别。

二战期间，美国空军希望研究应优先在轰炸机的哪些部位加装防护装甲，以提高飞机的返航率。为此，他们收集了所有成功返航的飞机，并详细记录了这些飞机上遭受弹孔的位置。数据显示，机翼、机尾和机身等部位的弹孔较多，而发动机部位则几乎没有弹孔。根据这一观察，不少人认为应当在弹孔密集的区域加装装甲。

然而，统计学家亚伯拉罕·瓦尔德（Abraham Wald）提出了完全相反的建议。他指出，数据仅来自成功返航的飞机，而那些被击中发动机却未能返航的飞机并未被纳入观察范围。因此，发动机弹孔少恰恰意味着这些部位一旦被击中，飞机便难以生还。因此，应当将防护重点放在看似“未中弹”的发动机部位。

这个案例清晰地揭示了一个重要问题：我们所观察到的数据并不总是完整或无偏的。在社会科学研究中，这种“样本选择偏差”（sample selection bias）极为常见。我们往往只能接触到某种“幸存者样本”，而无法观察到那些由于某种机制而未能进入样本的个体。因而，仅凭已有数据得出的统计关联，未必反映真实的因果结构。

在社会科学研究中，类似瓦尔德的“反向思考”极为重要。许多情况下，我们所拥有的数据并不足以直接支持一个可靠的因果推断。我们需要意识到，哪些变量能够进入我们的分析视野，本身就可能受到制度、选择机制或测量限制的影响。

另一个典型案例是家长择校问题。许多家长简单地将学校升学率等同于教育质量。例如，国际学校学生普遍表现优异，但这可能源于其学生家庭背景的优势。这里，家庭背景作为第三变量，同时影响了学校选择和学业表现。这一现象称为“虚假相关”（spurious correlation），其根源在于存在一个共同影响因变量与自变量的“第三变量”（confounder）。若未能识别并控制这些混淆变量，我们所进行的因果推断将缺乏有效性，甚至可能产生误导。

为应对这一挑战，科学研究提出了“控制实验”的基本逻辑。其核心思想是：通过人为操纵某一变量（自变量），并在其他条件保持一致的情况下，观察结果变量（因变量）的变化，从而识别是否存在因果效应。

大航海时代是人类历史上地理大发现的开端，也是全球化进程的一个关键阶段。在这一英雄辈出的时代背景下，坏血病成为横亘于航海探索之路上的巨大障碍。据历史统计，在1500年至1800年期间，约有超过200万航海人员死于坏血病，平均每年约6700人。因此，坏血病可被视为大航海时代的一种“时代性疾病”。

一个典型的例证出现在1740年英国海军的一次环球航行中。在此次航行中，舰队起航后的十个月内，1900名船员中有多达1400人因坏血病死亡。更具启发性的是，比较同期不同国家海员的病亡数据发现，英国海员的坏血病死亡率明显高于法国与西班牙海员，这引发了关于疾病分布与饮食文化之间关系的深层次讨论。

值得注意的是，与西方国家的情况相比，中国明代郑和七下西洋的史实中，并无坏血病流行的明确记载。此一现象或许与中国古代航海路径沿岸补给频繁、船队后勤保障完善以及传统中餐在营养结构上的均衡特征有关。中国传统饮食重视蔬菜摄入与膳食搭配，这或在无意中避免了坏血病的发生，形成一项文化上的“防病机制”。

对坏血病病因的科学认知得益于苏格兰医生詹姆斯·林德（James Lind）的先驱性探索。林德出身平凡，仅受过中等教育，后以军医助理身份加入英国皇家海军。在一次于比斯开湾的巡航任务结束后，他在萨利斯堡号军舰上首次实施了系统性的对照试验，堪称人类历史上最早的随机对照实验之一。

林德将坏血病患者随机分为六组，分别给予不同饮食干预措施，包括苹果酒、硫酸药水、醋、海水、柑橘类水果以及混合饮品等。实验持续仅五天，便观察到第五组（摄入柑橘与柠檬者）病情显著好转，部分患者甚至基本康复。该实验最终确立了柑橘类水果（富含维生素C）对治疗坏血病的疗效，为现代营养学奠定了基础。

基于上述实验结果，可以初步推断维生素C缺乏是坏血病的根本原因。由此产生了一个引人思考的问题：为何英国海员比法国与西班牙海员更易罹患该病？一个可能的解释在于饮食结构的差异。英国传统饮食长期以来缺乏对新鲜蔬果的重视，而法国、西班牙、意大利等国更注重饮食品质与营养均衡。相对而言，英国海员更难通过日常饮食摄取维生素C，这可能是其坏血病患病率偏高的主要原因。

这一差异最终促使英国海军强制要求远洋船队配备柠檬汁，以作预防之用，进而使英国海员在民间获得了“Lemmy”的外号。

将视野回归中国历史，可以发现郑和船队鲜有坏血病记录，这与中国古代饮食文化注重蔬菜与谷物搭配密切相关。此外，其航线多沿亚热带与热带海岸展开，停靠频繁，可实现充足补给，从而避免因长时间缺乏维生素C而诱发疾病。这种频繁靠岸补给机制，本质上形成了一种“天然干预”，进一步解释了坏血病在明代航海实践中未曾成为重大威胁的原因。

林德实验的设计过程展示了现代随机对照试验的基本逻辑：样本随机分组、前测基线评估、干预变量施加与后测评估。具体而言，实验组前测与后测分别记作 M1 与 M2，控制组为 M3 与 M4。要评估干预效应，可使用倍差法（Difference-in-Differences, DID），即 (M2 - M1) - (M4 - M3)，从而排除时间趋势与样本特质对结果的干扰。

在实验方法论上，这种设计解决了两个关键问题：选择性偏误与虚假相关。随机分组保障了实验组与对照组在均值意义上的可比性，从而有效隔离了第三变量的影响。若仅分析实验组的前后变化，无法排除由于样本内在特质所引起的偏差。因此，控制组的设置是因果推断成立的必要条件。

在社会科学领域，尽管受限于实验操作的可行性，随机对照实验依旧被视为推断因果关系的“黄金标准”。其核心在于通过随机分组机制解决潜在的选择偏差问题，通过对照组消解第三变量的干扰，从而提高内部效度。此外，随机抽样机制虽然主要服务于外部效度，即实验结果的可推广性，但其本身并不解决因果识别问题。因此，内部效度的确立仍需依赖控制性实验本身的设计严谨性。

在前述自然科学研究的基础上，我们进一步探讨“求异法”（method of difference）这一因果推断的经典方法。求异法强调通过比较不同情境下变量的异同，识别导致结果差异的关键原因。具体而言，当两个案例在所有因素上相似，唯独某一关键因素不同，而该差异恰好对应了结果的不同，便可推断该因素可能是因果变量。

然而，求异法的有效运用前提，依赖于清晰且严格的概念界定。若研究者未能精准划分变量及其属性，或将性质截然不同的政治事件混为一谈，那么所谓“唯一不同的因素”便难以成立，因果推断也会因此失去说服力。

在展开概念性分析之前，我们不妨通过一个具体案例来引出讨论。2020年，《World Politics》杂志刊发了一篇关于革命政体（威权主义）的研究(Lachapelle et., 2020)，旨在解释为何某些革命政体具有更强的“韧性”。所谓“韧性”，是指一个政体具有更长的存续时间以及较低的崩溃概率。作者认为，革命政体之所以具有显著韧性，根源在于它们通常继承了四项关键的制度性遗产。

第一，这些政体往往在革命过程中塑造出一个高度凝聚、彼此信任的统治精英集团。第二，它们建立或重组了一支对统治阶层高度忠诚的武装力量。第三，革命政体拥有一套强有力的压制机制，能够有效控制社会和镇压反对力量。第四，革命在其爆发与胜利过程中通常彻底消灭了主要的敌对政治力量。因此，一旦建立，革命政体便具备了相对稳定和长期运作的制度基础。

由此可见，作者通过制度遗产的分析逻辑，论证了革命政体之所以较为持久，乃是因为其在初创阶段即获得了一整套支撑其运行的制度资源。该解释路径在逻辑上具有较强的说服力。

进一步追问，我们会发现这项研究背后的核心概念是“革命政体”本身。然而，在现实政治中，社会抗争、社会运动乃至暴力革命的形式多种多样，并非所有形式的变革都可被归类为“革命政体”。为此，作者对“革命政体”提出了严格的定义，并设定了四个必要条件：

1. 革命必须是自下而上的社会运动，而非由现有精英集团发动或主导的体制内变革。因此，若革命源自统治集团内部的政治博弈，则不符合其定义。
2. 必须通过实质性暴力推翻旧有政权。这里的暴力并非指象征性的抗议或街头示威，而是指对国家机器造成实际破坏并最终颠覆政权的物理性暴力。
3. 必须建立新的国家机构，尤其是新的军事力量。若革命未能重组国家军队，或者继续沿用原有官僚结构，则不足以构成制度性断裂。
4. 必须伴随激进的社会变迁，尤其是财富再分配与阶级结构的重构。以中国革命为例，其彻底的土地改革和新阶级体系的建构即构成激进社会变迁的重要体现。

唯有同时满足上述四项标准，某一政治事件才可被归入“革命政体”的范畴。基于这一界定，作者排除了大量表面上看似“革命”的政治事件，包括但不限于以下几种情形：第一，虽有社会动荡，但未造成政治结构的根本改变；第二，虽有暴力冲突，但国家权力体系未发生实质性转型；第三，政权更替虽然发生，但是在国家支持下进行的制度性接班；第四，那些由上而下推动的“颜色革命”或“精英政变”；第五，法西斯政体的建立也不满足其四项标准。

据此，作者排除了诸如欧洲的颜色革命、阿拉伯之春、埃塞俄比亚的政变事件，以及纳粹德国和墨索里尼意大利等案例。最终，在1900年至今的全球范围内，仅有18个事件满足其对“革命政体”的严格定义。这意味着，在其研究框架中，研究总体仅包含18个单位。

在此基础上，若研究者挑选4个案例进行深入比较分析，则这4个案例大致代表了全部案例的四分之一，具有较高的代表性。这种设计思路相较于泛化数百个事件的做法，更加严谨，亦更能经得起学术检验。

因此，这一研究不仅展示了定义工作的重要性，更提醒我们：关键概念的界定并非随意为之，而是必须嵌入在理论构建和研究设计的整体结构之中。尤其在因果分析中，若想运用“求异法”（method of difference）来识别因果机制，就必须以严密的概念操作为前提。

该案例充分说明：理论概念的建构过程，实际上与研究方法是交织在一起的，定义的过程本身即是一个理论化与方法化的过程。因此，对于初学者而言，理解和掌握概念界定的逻辑，不仅有助于明确研究对象，也为整体研究设计提供了坚实基础。相信经过本例的分析，大家对于理论与方法之间的内在关联已有更加深刻的理解。

在质化研究之外，量化研究通常也要求我们掌握类似的逻辑，图[4.1](#fig:lind)中展示了最为基础的带有混淆因素的因果图。



林德医生实验的因果图

因果图（Causal Diagram），也称为有向无环图（Directed Acyclic Graph, DAG），其理论发展主要源于计算机科学家朱迪·珀尔（Judea Pearl）在20世纪90年代的工作。珀尔致力于将因果关系形式化，使其能够在统计学和人工智能领域进行严谨的推理。

在1990年代之前，统计学主要依赖相关性分析，而因果关系的讨论往往缺乏严格的数学框架。珀尔提出的结构因果模型（Structural Causal Model, SCM）和因果图，首次将因果关系表示为变量之间的有向无环图，其中节点代表变量，箭头表示因果影响。

1995年，珀尔进一步引入do-算子（do-calculus），使研究者能够区分观察到的关联（）和干预后的因果效应（）。这一突破使得因果推断能够在观测数据（而非实验数据）中进行。2000年后，因果图在流行病学、经济学、政治学等社会科学等领域广泛应用，成为因果推断的核心工具之一。同时，珀尔的理论与鲁宾的潜在结果框架相互补充，共同推动了现代因果科学的发展。

## 结构式到简约式：因果图与潜在结果模型

### 路径

在珀尔看来，因果关系事实上分为三种层次，首先是关联（relation），其次是干预（treatment），最后是反事实（counterfactual）。珀尔用了上节提到过的因果图说明这三种层次，为了理论的贯通，我们采用珀尔书中的叙述结构，并暂时脱离社会科学：

在《为什么》一书中，珀尔意图找出一条能让人工智能摆脱概率而走向必然乃至创造必然的数学方法，他预言这一天的到来同样也是人工智能类似人类走出伊甸的开端——即人工智能亦具有三种能力——观察、行动以及想象。

目前，依赖直觉进行观察的统计学仍不时落入各样悖论之中，解决悖论，获得一个准确的估计量乃是我们踏入干预乃至反事实的前提。

##### 蒙提·霍尔悖论（Monty Hall Problem）

作为经典的概率悖论，源自电视节目《Let’s Make a Deal》中的游戏环节。该问题设定为三扇门，背后分别藏有一辆汽车和两头山羊。参赛者初始选择一扇门后，主持人会打开一扇未被选择且藏有山羊的门，随后询问参赛者是否更换选择。尽管直觉上可能认为换门与否的中奖概率均为50%，但数学分析表明，换门策略可将中奖概率从初始的1/3提升至2/3，为什么呢？

一个简单的想法是，首先假定参赛者第一把就选中了汽车（1/3），第二把选中汽车概率为0/1；再假定我第一把未选中汽车（2/3），第二把选中汽车概率则为1/1（我显然不会故意去选山羊），那么此时参赛者的预期中奖概率（）当然飙升了整整1/3。大部分人认为的一半对一半实际上只可能出现在主持人也不清楚具体奖品随机开门之时。

##### 伯克森悖论（Berkson’s Paradox）

是一种统计现象，揭示了在特定条件下可能出现的虚假关联。当两个原本独立的变量A和C均对结果B产生影响，而样本仅局限于B发生的个体时，可能会观察到A与C之间的负相关关系，即使总体中二者并无关联。这种现象被称为“对撞谬误”或选择偏差。



伯克森悖论

图[4.2](#fig:berkson)中展示了一例：众所周知，枪击和中风都会导致生命的消散，但是人的生命只有一次也不能复活，因而如果一定要死（也就是这两种死法），也只能够选择其中一种，显然，被枪打死的人不可能再中风死，在观察时若我们如果只看到死在这两种死法下的个体，则轻易可以得出中枪死与中风死有着负向相关的关系。

##### 辛普森悖论（Simpson’s Paradox）

是统计学中著名的聚合现象，指数据在整体层面呈现某种趋势，但按某一变量分层后，各子组的趋势方向与整体相反。



辛普森悖论

典型案例即为伯克利研究生院性别录取率差异，总的来看，男性的录取率高于女性（44.1%大于30.4%），但当我们把研究生院再划分为不同的院系后，大部分院系对于女性的录取率实际高于男性（见图[4.3](#fig:simpson)）。彼得·比克尔（Peter Bickel）认为这一现象乃处于女性更倾向于申请竞争压力较大的院系，在这种情况下，大量的陪跑案例当然拉低了总的录取率。该悖论揭示了忽视混杂变量可能导致的错误推论，强调分层分析与因果推断中考虑潜在混杂因素的必要性。

##### 罗德悖论（Lord’s Paradox）

指同一数据集通过不同统计方法分析可能得出矛盾结论的现象。这一悖论最早由统计学家弗雷德里克·罗德（Frederic Lord）在1967年提出，揭示了数据分析中一个关键问题：统计方法的选择如何影响研究结论，尤其是在涉及协变量调整时。



罗德悖论

在罗德的研究案例中，两位统计学家对同一组数据（学生入学体重X和一年后体重Y，图[4.4](#fig:lord)）进行分析，却得出截然相反的结论：

1. 简单均值比较：第一位统计学家直接比较男女学生的体重变化均值，发现两组均无显著变化，因此认为食堂对体重无影响。
2. 协方差分析（ANCOVA）：第二位统计学家学过控制入学体重（X）后，发现男女学生的体重变化存在显著差异（截距差为6.34磅），认为食堂对男女影响不同。

这一矛盾表明，是否调整协变量可能导致研究结论的差异，其背后的根本问题在于，我们无法得知不吃食堂这一反事实（counterfactual）条件下男女的体重情况（没有对照组），因而使得回归或者协方差分析等统计工具失效，并不能清楚地回答有关因果的问题。

通过因果图的形式，我们可以更直观地看到这些悖论的出现，此处可以跳到图[2.1](#fig:relation)复习。显然，这些变量关系在本章中又再次出现，由于已经引入了因果图的概念，我们可以对其进行案例化，如图[4.5](#fig:wahaha)：



财产分配的因果图

在某企业家引发的舆情事件中，我们尝试提炼出一个关于社会关系继承的因果机制模型。首先，从死亡方式切入，该企业家据称因病去世（例如中风），这一前提将我们讨论的问题限定在“自然死亡”或“疾病死亡”等非暴力性死亡情境之中。这一设定本身即排除了如枪击等“非自然死亡”情境所涉及的诸多干扰因素，从而有助于厘清自然死亡背景下的社会关系变动，当然我们在做大样本分析时不能排除非自然死亡的情况。

然而，在因果分析中一旦引入“是否死亡”作为控制变量，便可能无意间触发“**对撞偏差**”（collider bias）的问题：当一个变量（如“是否死亡”）同时受到两个原本无关联的因素（如“是否中风”与“是否枪击”）影响，且被纳入分析模型时，会在它们之间引入人为关联，从而扭曲对因果路径的识别。例如，死亡作为既成事实，是我们观察遗产分配问题的前提，而死亡方式作为路径条件，本应是划定分析边界的工具。但当我们将“是否死亡”作为控制变量时，原本独立的“是否中风”与“是否枪击”两个变量便在图式中联通，形成类似“ ”这样的虚假因果路径。这种结构性错误会削弱我们对“自然死亡背景下遗嘱确立行为”这一主路径的识别能力。

进一步而言，“确立遗嘱”通常通过影响“家庭关系”间接作用于“财产分配”，家庭关系因而可被视作一个具有机制意义的中介变量。该路径结构揭示出财产继承不仅是个人意愿的体现，也是亲属关系结构的延续与反映。然而，若研究者试图将所有可能的中介路径一并纳入模型进行调整（例如控制家庭互动频率、情感亲疏、文化观念等），则可能落入“**过度控制偏差**”（over-control bias）的陷阱。在这种情况下，“个人意愿”这一关键变量的真实效应，可能被其他解释变量的残差结构所吸收，从而妨碍我们对主因路径的识别与估计。

在死因探讨之外，“日常为人”（例如企业家的性格、道德声誉、人际互动方式等）不仅可能影响其家庭关系，也可能对其财产分配产生前因上的作用。因此，“日常为人”作为一个潜在混淆变量，若未被纳入分析，将可能引发如“辛普森悖论”这类由于聚合数据忽视分层关系而造成的推断错误，也叫“**混淆偏差**”（confusion bias）。

此外，不容忽视的是“法律制度”等宏观背景变量。虽然其通常具有高度稳定性，短期内不易观察其变化，但在长时段演化中，其对遗嘱制度与财产分配的规范作用不可忽略。尽管制度或文化等外在环境因子的量化困难重重，但已有研究通过创新方法成功应对这一挑战。

### 截断

面对上述的复杂因果关系，我们似乎难以把握，尤其是如何精准地提取出X对Y的影响。事实上，指望在统计学范畴内予以彻底解决——尤其是在社会科学领域中试图收集关于全体变量的完备数据——几乎是不可能的任务。正因如此，珀尔提出了“干预主义”思想，并以do算子为工具，试图跳出被动观察的框架，主动设定变量的状态，从而厘清因果路径。

然而，即便引入了干预操作，我们仍需回答一个基础性问题：在干预变量X的同时，是否还有其他路径会影响到Y？亦即，我们如何“截断”所有除 之外的混杂路径，只保留真正的因果链？为此，理解路径结构是至关重要的，以下两类共四种路径类型构成了因果图分析的基础。

##### 前门路径

前门路径指的是一种通过中介变量传导的路径，即存在一条 的路径，其中M是介于X与Y之间的中介变量。若前门路径独立于所有混杂路径，并且X对M与M对Y的因果关系均可识别，那么通过控制M这一中介变量，即可在未观测混杂变量Z的前提下识别 的因果效应。这便是所谓“前门准则”（front-door criterion）所依赖的结构条件。

在上述的例子当中，当我们认定生前为人对家庭关系并无影响的情况下，便可以通过估测是否死亡对家庭关系（ ）以及家庭关系对遗产分配（ ）得出一个确立遗嘱对遗产分配（ ）的因果效应来。

前门路径的核心优势在于：它不要求我们观测并控制所有影响 和 的共同原因，而是通过中介变量的“结构绕行”实现了因果推断。这也体现出结构因果模型（SCM）在非实验数据中的强大推理能力。

引入do算子之后，我们关注的已不再是观察性条件概率 ，而是介入性概率 。这是因为 表示我们人为地设定变量 的值，从而切断 与其上游变量（包括潜在混杂变量 ）之间的自然因果联系。这一操作是实现因果推断的关键步骤。

在满足前门准则的前提下， 可以通过如下公式进行识别：

这一表达式背后的逻辑如下：我们首先利用 捕捉 对中介变量 的因果影响，然后通过对所有可能的 求和来消除 与 之间的关联性，进而评估 对 的因果影响。由于 同时受到 的影响并影响 ，其在传导因果效应的过程中起到了桥梁作用。

继续回到前述的社会情境例子，在认为“为人”不影响“家庭关系”的前提下，如果我们能观测并建模“是否立遗嘱”对“家庭关系”的影响（），以及“家庭关系”对“遗产分配”的影响（），则即便存在无法观测的混杂变量（例如家族文化、遗产规模、潜在冲突历史等），但由于这些混杂变量不直接影响到家庭关系这一中介变量，故我们仍可以通过前门准则识别 对 的净因果效应，如图[4.6](#fig:wahaha2)所示。



前门路径示例图

##### 后门路径

相比之下，后门路径是一条从 开始，并通过其他变量“绕道”影响 的路径，但其首个箭头是指向 的，例如 。这样的路径通常并不代表真正的因果传导链条，而是一条混杂路径（confounding path），会引入 与 之间的虚假相关性，从而干扰我们对 因果效应的识别。

为了封闭此类路径，我们需要控制这些共同前因（common causes），即将变量 纳入模型作为控制变量，使得在条件化 之后， 与 之间的关联可以被解释为因果效应。如图[4.7](#fig:wahaha3) 示，在我们的实际社会例子中，“日常为人”与“法律制度”可能既影响立遗嘱的行为，也影响财产分配的结果，因此构成混杂变量，需予以控制。

珀尔提出的“后门准则”（back-door criterion）是识别这类因果效应的重要工具。具体而言，若存在一个变量集合 满足：

* 所有从 到 的后门路径都被 所阻断；
* 不包含任何 的后果（即不包括 的子节点，否则变为内生性问题）。

则因果效应 可通过如下方式识别：

在该表达式中， 是在观测数据中可以直接估计的条件概率，而 是混杂变量的边际分布。通过对所有 的加总，我们相当于模拟了一个“随机赋值” 的实验情境，从而排除了由于 引起的混杂偏误。

需要特别强调的是，这一公式的左侧为 ，表示我们关心的是在人为地干预 为某一特定值时， 的响应分布——这是因果推断的核心目标。右侧虽然只涉及观察性概率（即没有 ），但由于我们通过控制混杂变量 封闭了所有非因果路径，进而使得 可以作为 因果效应的替代估计。

换言之，这一加权求和过程相当于在每个 的条件下估计 对 的局部因果效应，并按 的边际分布 进行整合，从而还原出全局的干预效应 。通过这种方式，我们实现了从观察性数据中“模拟实验”，使得不具备随机化条件的数据也能够进行有效的因果识别。

后门准则的实质是：在能够观测并控制混杂变量的前提下，我们可以通过调整（adjustment）还原出一个近似实验的因果估计。而do算子的引入则使得这种估计从统计相关性上升为因果推断。



后门路径示例图

##### 开路径

所谓“开路径”是指一条变量之间的路径，其结构在没有被控制的前提下，允许信息在变量间流通，从而导致相关性。在因果图中，如果路径中的所有非冲节点（collider）未被控制，且没有冲节点的后代被控制，那么这条路径就是“开”的（也就是因果图中因能走向果）。这意味着X与Y之间可能因该路径而存在关联，尽管这种关联未必具有因果含义。因此，在识别因果关系时，必须将所有不希望开启的路径加以“**截断**”。

##### 死路径

与“开路径”相对，死路径指的是在控制某些变量后，使路径上出现冲节点或其后代，从而阻断信息流通的一条路径。典型的死路径形式是当路径中存在冲节点（如 ）且未控制W或其后代时，该路径自动封闭，变量之间不再相关。对撞路径实际上便是一条死路经，显然，“是否中枪”与“是否中风”的关联完全是由于死亡这一结果而出现，两者间实际上并无瓜葛。这一特性在因果识别中尤为关键，因为它说明并非所有路径都需要控制，有些路径由于结构上的封闭性天然为“死路径”，不构成混杂来源，控制其反而造成“对撞偏差”。

尽管珀尔的理论如此完备，想要据此寻找可行的前门路径和截断所有后门路径仍然不太可能，结构式的复杂也在此凸显无疑，当下的计量经济学似乎是，也仍然将是简约式的主场，通过设计，我们实现了在找不齐路径中变量的情况下仍然能控制偏差的壮举。

### 设计

回到林德医生的实验当中，其实验虽然未满足现代随机化的严格标准，但其核心逻辑——通过人为控制变量，比较处理与非处理之间的差异——却与现代实验设计高度一致。这种逻辑，正是后世“潜在结果模型”（Potential Outcomes Model）所要形式化表达的内容。

此处也是珀尔的第三层次——反事实，也即是人的想象力，对此做出重大贡献的乃是唐纳德·鲁宾（Donald Rubin），最早提出者为耶日·奈曼（Jerzy Neyman）。因此，这一理论框架通常被称为“奈曼—鲁宾因果模型”（Neyman-Rubin Causal Model, RCM）。其基本思想是：对于任一研究对象，在特定时间内只能处于一种处理状态（接受干预或未接受干预），因此我们永远只能观察其一个“潜在结果”，另一个结果则永远处于反事实状态中。这种因果分析中的结构性“数据缺失”，构成了所谓的“因果推断的根本问题”（Fundamental Problem of Causal Inference, FPCI）。

鲁宾因果模型的核心构件包括三个基本概念：干预（Treatment）、潜在结果（Potential Outcome）与干预效应（Treatment Effect）。干预指的是研究中所施加的某种处理或变化，通常是我们试图识别其因果影响的自变量；潜在结果则是指在不同干预状态下，一个研究对象可能产生的因变量取值；干预效应即为同一单位在两种状态下结果的差值，是我们最终希望估计的因果量。

根据作用层级的不同，干预效应可分为个体层面与总体层面。个体干预效应指单位 在接受干预与未接受干预状态下结果之差，形式上表示为：

由于每个单位在同一时间只能处于一种状态， 无法直接观测，这便导致了反事实推断的必要性。在总体层面，研究者通常通过对个体效应取期望来构造平均干预效应指标。其中，最基本的是平均处理效应（Average Treatment Effect, ATE）：

此外，还可针对特定子群体定义条件平均效应。例如，干预组的平均处理效应（Average Treatment Effect on the Treated, ATT）定义为：

控制组的平均处理效应（Average Treatment Effect on the Untreated, ATU）则为：

如霍兰德（Paul Holland）所指出，因果推断的根本难题在于，我们无法同时观察一个研究对象在两种处理状态下的结果。为了克服这一困难，研究者常常借助反事实估计的策略，最常见的方式是用控制组的结果均值 ，来估计干预组在未接受干预状态下的潜在结果 。这种方法的有效性则取决于“可忽略性假设”（Ignorability Assumption），即潜在结果与处理状态在统计意义上相互独立。

不同的估计目标对应不同形式的可忽略性假设。若估计 ATE，需要假定 ，即干预状态的赋值机制与潜在结果无关；若估计 ATT，仅需假定 ，即在控制状态下的潜在结果与是否接受干预无关；若估计 ATU，则需要假定 ，即在干预状态下的潜在结果不因是否接受干预而异。

在理想情境中，这一假设可通过随机分配来自然满足，因此随机对照实验（Randomized Controlled Trials, RCT）被视为因果推断的“黄金标准”。罗纳德·费舍尔（Ronald Fisher）首次明确提出并论证了随机分配的价值：通过将研究单元等概率地分配至干预组与控制组，使得两组在干预前具有相似的分布特征，从而确保组间的差异可以归因于干预本身。这一原则成为现代实验设计的核心支柱。

在鲁宾因果模型中，这种设计依赖两个关键假设：其一是“稳定单元处理值假设”（SUTVA），即个体之间的处理无交叉干扰；其二是“非混淆性分配机制”（unconfounded assignment），即组别分配不受潜在结果的影响。前者保证干预效应不会溢出至他人，后者则保证分配机制的独立性。只有在满足这两个前提下，实验设计才能为因果推断提供可信基础。

在社会科学中，这种理想的实验逻辑却并不容易实现。一方面，实验设计的伦理约束、成本问题与操作复杂性，往往限制了随机实验的实施；另一方面，个体特征的复杂性也使得控制变量难以穷尽，尤其是不可观测因素对结果的干扰始终构成威胁。因此，社会科学者在实践中发展出多种实验设计样式，以应对复杂因果关系并提高效度。

* 被试内设计（Within-Subject Design）：即每个受试者依次接受多个处理。该设计优势在于控制了个体间差异，提高了实验效率；但也可能引发“顺序效应”，如前测、疲劳、练习等非处理因素对结果的影响。
* 被试间设计（Between-Subject Design）：将被试随机分配至互斥的处理组中。该设计较为常见，尤其适用于干预具有不可逆性或需避免互相影响的场合。其进一步发展包括：
  1. 前测—后测控制组设计，用于比较处理前后的变化。
  2. 仅后测设计，适用于无法实施前测的场景。
  3. 所罗门四组设计，则综合两者以控制前测效应的混入。

无论何种设计，其目的都在于提升内部效度，即确保观测到的因果关系确实来自处理本身。然而，若实验过于理想化，其在真实世界中的适用性便会受限，从而降低外部效度。社会科学中的许多研究尝试在二者之间求取平衡——在确保干预效应估计准确的同时，也希望研究结论具有一定程度的概括力。

例如，实地实验（Field Experiment）作为一种介于控制实验与观察性研究之间的设计方式，在自然情境中施加干预，试图在控制干预变量的同时保留社会现实的复杂性。这类设计既考虑了实验的可行性，又兼顾了结论的推广性，因而在发展经济学、政治学与教育研究中获得广泛应用。

尽管实验的结果可能较为稳健，但不同于自然科学，社会科学学者更多还是依赖观察性的数据展开研究，试图戴着研究伦理的镣铐跳舞，从自然实验（Natural Experiments）发展到准实验（Quasi-Experiments）也就不足为奇。

## 从统计推断迈向因果推断

理解因果关系是社会科学的核心任务之一，传统统计推断方法，尽管在描述变量间的相关性方面表现出色，却在揭示变量之间的因果机制时面临着根本性局限。现代计量经济学的发展，正是试图填补这一从统计相关到因果推断之间的理论与方法鸿沟。计量方法的基本目标，是在有限的数据约束下，通过形式化建模揭示变量间的因果联系，并对特定政策或干预的效果做出可靠的估计。在具体实践中，计量研究常常依赖观察性数据（observational data），而非实验性数据（experimental data）。观察性数据的最大特征在于，研究者无法控制处理变量的赋值机制，因此极易受到各种混淆因素的干扰。在这种情形下，内生性问题成为因果推断中最为关键的挑战之一。内生性可能源于多个方面：一是遗漏变量，即模型中未控制的重要变量与处理变量和结果变量同时相关；二是测量误差，尤其当关键解释变量存在系统性偏误时，会导致估计量偏离真实值；三是联立性问题，即因果方向本身具有双向性，被解释变量反过来也影响解释变量。

在进行因果分析前，研究者往往需要先构建一个适当的统计模型来刻画变量之间的表面关系。最常用的工具是线性回归模型，尤其是最小二乘法（Ordinary Least Squares, OLS）。OLS的估计建立在一系列关键假设之上，其中最为核心的是零条件均值假设，即解释变量与误差项不相关（又称无内生性），以及误差项具有同方差性（即误差的方差在解释变量的各个取值下是常数）。一旦这些假设得到满足，回归系数就可以被解释为在其他变量控制不变的情况下，自变量变化一个单位所导致的因变量的平均变化值。在线性框架下，统计推断工具如t检验和F检验可以用于判断系数的显著性，但模型的可靠性还需经过多方面的诊断。常见的诊断方式包括残差分析，用以检验异方差、自相关和正态性等；在面对异方差时，可采用稳健标准误进行修正；此外，当解释变量之间存在高度相关性时，共线性问题会削弱估计结果的稳定性，此时需借助方差膨胀因子（VIF）进行评估。

当数据在时间与个体两个维度上同时存在时，面板数据模型便成为更有力的工具。面板模型允许研究者在控制个体不随时间变化的固定特质的同时，利用时间序列变化捕捉因果效应。固定效应模型（Fixed Effects, FE）通过引入个体固定项来控制这些不可观测的稳定特征，而随机效应模型（Random Effects, RE）则假定个体差异是随机的，并与解释变量不相关。一阶差分模型（First Difference, FD）通过对变量做差，消除固定不可观测项，在某些数据结构中也具有优势。在选择固定效应还是随机效应模型时，豪斯曼检验（Hausman Test）提供了一个标准化的判断依据，以检验个体效应是否与解释变量相关。

处置效应（Treatment Effect）是因果推断中一个核心概念，其基本思想是比较单位在接受与不接受某一干预时的结果差异。除了平均处理效应（Average Treatment Effect, ATE），研究者还常常关注特定人群的处理效应，如干预组的平均处理效应（Average Treatment Effect on the Treated, ATT）以及条件平均处理效应（Conditional Average Treatment Effect, CATE），以便更细致地理解干预在不同子群体上的异质性影响。

在缺乏实验条件的现实中，匹配方法成为解决可比性问题的重要工具。其中最为广泛应用的是倾向得分匹配法（Propensity Score Matching, PSM），该方法首先估计个体接受干预的概率（即倾向得分），然后在倾向得分相近的处理组与对照组之间配对，以构造出具有协变量平衡的对照组。匹配策略包括最近邻匹配、半径匹配与核匹配等多种形式，它们在匹配精度与样本利用率之间各有权衡。通过匹配，研究者可以降低由于协变量不平衡所导致的估计偏误，从而在非实验条件下近似实现条件可忽略性假设。

对于更为复杂的内生性问题，仅仅通过协变量匹配往往不足以解决。此时，工具变量法（Instrumental Variables, IV）提供了一个替代路径。工具变量是一个与解释变量高度相关，但与被解释变量仅通过该解释变量产生影响的外生变量。满足相关性与排他性两大条件的工具变量，能够为解释变量中的内生部分提供“自然实验”的来源。两阶段最小二乘法（Two-Stage Least Squares, 2SLS）是IV估计的标准技术路径，第一阶段用于预测内生变量的拟合值，第二阶段再将其带入回归进行因果估计。然而，工具变量若与被解释变量的关系过弱，将导致“弱工具变量问题”，使得估计不但不一致，而且在有限样本中方差极大，因此对工具变量的强度检验也成为实务中的必要步骤。

当干预具有明确的时间切换点时，双重差分法（Difference-in-Differences, DiI）是一种极具解释力的准实验设计。该方法通过比较干预组与对照组在干预前后的差值，进而构造出一个“差中的差”来识别干预效应。其核心在于“平行趋势假设”，即在没有干预的情形下，两组单位应具有类似的时间趋势。这一假设虽不可直接检验，但可以通过图表直观呈现，并辅以前期趋势检验作为间接支持。DID方法尤其适用于政策评估与制度变革等自然实验情境，是当前实证社会科学中的主流技术之一。

当处理变量的分配存在某种阈值机制时，断点回归设计（Regression Discontinuity Design, RDD）提供了一种接近随机化的识别策略。若处理的赋值完全由某一变量是否超过某一界限所决定，则构成“锐断点设计”（Sharp RDD）；若该界限只部分决定处理状态，则为“模糊断点设计”（Fuzzy RDD）。RD设计基于连续性假设，即阈值两侧单位在未处理状态下的潜在结果应是连续的，从而将阈值附近的单位视为在期望上可比。估计的关键在于选择合适的带宽范围与拟合方法，常用图表辅助判断结果的可信度，是实证研究中一种严谨但要求较高的识别工具。

后文将对每一段进行细致地展开，并配有Stata代码案例，基本格式如下：

\* Stata 输入代码

sysuse auto, clear // 加载系统自带数据集  
 summarize // 描述性统计  
 graph bar weight price, over(foreign) // 绘制分组条形图

\* Stata 输出文字

(1978 automobile data)  
  
 Variable | Obs Mean Std. dev. Min Max  
 -------------+---------------------------------------------------------  
 make | 0  
 price | 74 6165.257 2949.496 3291 15906  
 mpg | 74 21.2973 5.785503 12 41  
 rep78 | 69 3.405797 .9899323 1 5  
 headroom | 74 2.993243 .8459948 1.5 5  
 -------------+---------------------------------------------------------  
 trunk | 74 13.75676 4.277404 5 23  
 weight | 74 3019.459 777.1936 1760 4840  
 length | 74 187.9324 22.26634 142 233  
 turn | 74 39.64865 4.399354 31 51  
 displacement | 74 197.2973 91.83722 79 425  
 -------------+---------------------------------------------------------  
 gear\_ratio | 74 3.014865 .4562871 2.19 3.89  
 foreign | 74 .2972973 .4601885 0 1

\* Stata 输出图像

![image](data:application/pdf;base64,)

99 邱嘉平，《因果推断实用计量方法》，上海财经大学出版社

赵西亮，《基本有用的计量经济学》，北京大学出版社

珀尔  等，《为什么》，中信出版集团

安格里斯特  等，《基本无害的计量经济学》，格致出版社

游宇  等.一切为了理论——理论化与整合性的案例选择策略[J].世界经济与政治,2023,(12)

许琪.因果推断五十年：成就、挑战与应对[J].学术月刊,2024,56(11)

燕继荣  等，《发展政治学学科地图》，北京大学出版社

# 统计推断的理解运用

## 多元回归

### 理解线性回归

### 小样本OLS

从小样本开始学习“最小二乘法”较为合适，首先是三大假定：

##### 线性假定

古典线性回归模型（Classical Linear Regression Model，CLRM）可以表示为：

其中，表示样本容量，代表第个观测值的第个解释变量（），而则是待估参数（回归系数）。特别地，若模型包含常数项，通常令。

线性假定的核心含义在于每个解释变量对被解释变量的边际效应为常数。值得注意的是，虽然称为"线性"假定，但模型中可以包含高次项或交互项来处理非线性关系，这仍然被视为满足线性假定。

为了更简洁地表达模型，我们定义以下矩阵：

* 表示扰动项

单个观测方程可以表示为：

将所有观测方程叠放，我们得到：

其中：

* 是的设计矩阵
* 是扰动项向量

##### 严格外生性假定

严格外生性假定可以表示为：

这一假定的含义是，给定数据矩阵，扰动项的条件期望为零。这意味着均值独立于所有解释变量的观测数据。由此可以推导出两个重要推论：首先，扰动项的无条件期望也为零，即；其次，解释变量与扰动项正交，即。

##### 不存在严格多重共线性假定

该假定要求数据矩阵满列秩：

如果这一条件不满足，则参数不可识别。在实际应用中，除非设置过多虚拟变量且包含常数项，否则严格多重共线性的情况较为少见。

##### 球型扰动项

球型扰动项假定可以表示为：

这一假定包含两个重要性质：条件同方差性（）和无自相关性（对于）。这些性质保证了OLS估计量的有效性。

##### 残差平方和最小化

对于参数的任意假想值，我们定义残差向量，最小化目标函数为：

展开后得到：

##### 向量微分与正规方程

利用向量微分规则，我们对求导并令其为零：

由此得到正规方程组：

解得OLS估计量：

##### 残差性质与方差估计

残差向量具有与解释变量正交的重要性质：。拟合值与残差正交，构成了的分解。

扰动项方差的无偏估计为：

其中为自由度。回归方程的标准误差则为。

##### 投影矩阵与消灭矩阵

OLS估计可以通过投影矩阵和消灭矩阵来解释。拟合值和残差可以分别表示为：

这些矩阵具有以下重要性质：

* （投影矩阵作用于设计矩阵保持不变）
* （残差与拟合值正交）
* （消灭矩阵作用于设计矩阵得零）
* 和都是对称且幂等的矩阵

##### 平方和分解

当模型包含常数项时，可以建立平方和分解关系：

其中TSS表示总离差平方和，ESS表示回归平方和，RSS表示残差平方和。

基于此分解，我们定义拟合优度：

它等于实际值与拟合值相关系数的平方。为了惩罚过度参数化，我们还定义了校正拟合优度。

##### 无常数项情况

当模型不含常数项时，传统的平方和分解不成立，但可以定义非中心：

这一指标在软件输出中更为常见。

##### 基本性质

OLS估计量具有以下重要性质：

1. **线性性**：是的线性组合
2. **无偏性**：在严格外生性下，
3. **方差表达式**：

##### 高斯-马尔可夫定理

在古典假定下，OLS估计量是最佳线性无偏估计量(BLUE)，即在所有线性无偏估计量中具有最小方差。这一结论的证明通过比较任意线性无偏估计量与OLS估计量的方差完成。

##### 方差估计

是的无偏估计，这一结论基于迹运算的性质：

因此，的无偏估计为。

##### 单个系数的检验

在正态性假定下，我们可以构造t统计量检验单个系数：

检验过程包括计算统计量、确定临界值和比较p值。置信区间的构建也基于t分布。

##### 线性假设的检验

对于多个线性约束，我们使用F检验：

F统计量也可以通过约束与无约束模型的残差平方和差异来表示。

##### 点预测与区间预测

给定新的解释变量值，点预测为：

预测误差的方差包含两部分：估计误差和扰动项方差。基于此，我们可以构建预测区间：

### 理解方差

1. 同方差性假设与其在OLS中的作用
2. 异方差性：定义、产生原因与后果
3. 异方差的检验方法（如Breusch–Pagan、White检验）
4. 异方差稳健标准误（heteroskedasticity-robust SE）
5. 聚类稳健标准误（clustered SE）及其适用情境

### 非OLS回归与时空计量经济学

1. 离散与受限因变量
2. 时间序列分析
3. 空间计量经济学

## do算子与处置效应

### 虚拟变量

1. 虚拟变量的定义与生成方法
2. 类别变量在回归中的处理
3. 基准组选择与虚拟变量陷阱
4. 交互项的引入与解释（特别是在异质效应分析中）

### 处置效应与偏误

1. 处置效应的定义（平均处理效应 ATE、ATT 等）
2. 反事实框架（potential outcomes）
3. 选择偏误问题及其来源（自选择、遗漏变量等）
4. 识别策略的必要性与作用

## 两种策略

### 匹配法

1. 匹配法的基本思想与目的
2. 匹配方法种类（最近邻匹配、半径匹配、核匹配等）
3. 倾向得分（propensity score）的估计与匹配
4. 匹配后平衡性检验
5. 匹配法的优点与局限性

### 回归法

1. 回归模型在估计处置效应中的作用
2. 包含控制变量的线性回归模型
3. 交互项与异质性处置效应
4. 回归调整与回归偏误
5. 与匹配法的比较：何时选用何种方法