第二篇

求索科学

第三章

精简和统一的基础概念体系

在本章中，我将定义一套精简和统一的基础概念体系作为本书的基础。

3.1 动机

不同学科经常使用名字相同的概念，但这些概念的内涵和外延并不一致。我在本书中尝试建立一套精简和统一的基础概念体系来描述不同学科中的问题。

不少基础概念及其它们之间的关系缺少明确的定义，例如对象，关系，事件等。在本书中，我尝试从头开始定义所有的概念，而且坚持用浅显易懂的语言。

我尽量减少基础概念的数目，目前只保留九个基础概念，包括对象、关系、事件、因果、求索、主体、命题、模型和公理系统。图 3.1 [1]展示了这九个基础概念之间的关系。从这个意义上来说，基础概念体系相当于公理系统的地基。在定义的过程中，我参考了关键文献 [1, 18, 294, 292, 339]。

我对一些基础概念进行了新的定义，例如因果，我从对象相互关系的角度对它进行了新的定义，这是本书的基础概念之一。

我定义了一些新的概念，例如因果独立性，因果条件独立性，这需要在将来发展新的数学模型和方法。

3.2 对象和关系

3.2.1 对象、关系和世界的定义

基本概念定义 1 (对象 (object))**.** *对象是拥有结构 (structure)、属性 (property)、机制 (mechanism) 和行为 (behaviour) 的一类实体 (a class of entities)。*

实体正是我在描述构建理论困境时提到的外生概念。我不打算定义实体，否则会陷入嵌套循环，这里的实体可以是组件、对象、嵌套定义的系统，甚至可以是衍生的属性、变量或者事件。

[1]康国新绘制了这个图。

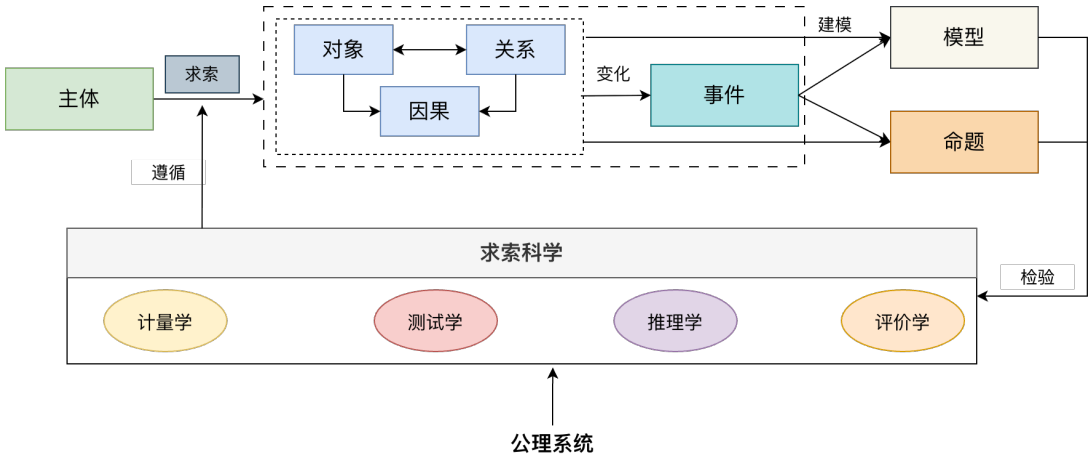

图 3.1: 对象、关系、事件、主体、求索和公理系统等九个基础概念之间的关系。

概念定义 1 (对象的结构). 对象的结构包括组件（*component*）及其组件之间的关系。

概念定义 2 (对象的属性（property）). 对象的属性是对象的内在特性（*intrinsic characteristics*）。对象的属性包括对象的结构属性以及行为属性。

概念定义 3 (机制（mechanism）). 机制是决定对象或者对象关系变化的根本方式，机制按照一定的规则将输入映射为输出。

机制分为内部机制（*internal mechanism*）和外部机制（*external mechanism*）。

概念定义 4 (内部机制). 内部机制是对象自身发生改变的根本方式。

概念定义 5 (外部机制). 外部机制是对象通过与外部世界交互发生改变的根本方式。

概念定义 6 (对象的状态). 对象的状态是指对象及其属性、结构、机制的快照。

一个典型的对象可以是自然科学、社会科学或者工程里的事物 (thing)、生命（life）、自然现象 (phenomenon)、抽象概念 (abstract concept)、过程 (process)、政策 (policy)、人造物（artifact）、虚构或者想象的事物。例如测试和测量过程可以视为一个对象。幽灵有可能不存在，但它也是一个对象。

在以上概念基础之上，关系可以定义如下：

基本概念定义 2 (关系（relationship）). 关系描述的是两个或更多实体之间的相互位置，可以用更高一层实体的属性、结构来描述或定义。

概念定义 7 (世界（world）). 世界是所有对象（*object*）及其关系的总和。

世界是最复杂的对象。

3.2.2　量和其他属性

概念定义 8 (计数 (counting))**.** 计数是为对象的特定属性赋予数值（后文简称赋值）的过程。

概念定义 9 (量 （quantity))**.** 量是对象可计数的属性（*countable property*)，包括结构量和行为量。

典型的对象的结构量有体积或质量。行为量有准确度和安全性等。

概念定义 10 (指标（metric）)**.** 指标本质上是量。

一个对象还具有其他不可计数的属性。心理学家斯坦利 · 史密斯 · 史蒂文斯 (Stanley Smith Stevens) 曾探讨这一主题，并将量的性质 (nature) 划分为四种：定类（nominal)、定序 （ordinal，基于顺序 [28])、 定距 （interval）和 定比 （ratio） [293]。定距和定比属于可计数属性。

英文原文中，斯坦利 · 史密斯 · 史蒂文斯使用了 “测量的水平或尺度” (levels or scales of measurement）这一术语。为了避免与本书的英文版本中使用的术语 “量表”（scales）或 “水平”（levels）的特定含义混淆，我使用 “性质”（nature）这一术语。

“定类” 性质代表最简单的测量形式，数字仅作为标签（label）或标识符（identifier）使用，用于建立等值关系（equality relation）”。“定序” 性质与定类性质不同：表现为 “对项（item）进行特定顺序的排序（ranking)”。定距性质 ： 具有 等距关系 的特征，即 (1) 零点的选择基于约定或便利；(2) 存在排序关系；(3) 当所有数值加减常数时，尺度（scale）具有不变性——数值间差异保持不变。“定比” 性质 ： 可支持 全部四种关系 ： 等值 (equality)、排序 (rank-ordering)、等距 (equality of intervals)、等比 (equality of ratios)。

属性还存在其他性质，这是一个值得研究的问题。

3.2.3　变化及对象行为

概念定义 11 (变化（change）)**.** 变化指对象、对象状态或者对象关系产生可观察、可测量、可测试或者可推理的差异。

对象的变化可以是对象属性、结构、机制和行为的差异，或者对象的消失 (come into being）和产生 (cease to being)。对象关系的变化可以用更高一层实体的属性或者结构来描述。

概念定义 12 (对象的行为（behaviour）)**.** 对象的行为是指对象及其状态被动或者主动地改变。

对象也可以是衍生的。

概念定义 13 (衍生对象 (derived object))**.** 衍生对象是指对象的属性、结构、机制及其变化，或者对象对其他对象产生的影响或者对象之间的关系及其变化。

从形式上来说，衍生对象是一个对象，但不一定具备完整的对象特性，例如拥有结构、属性、机制。后文提到对象的时候，在不同的语境下也可能包括衍生对象。

3.2.4　未知、已知、显式、隐藏和明确定义

概念定义 14 (已知 (known)，未知（unknown）对象)**.** 如果一个对象至少有一个属性或结构或机制或行为可以被求索，包括观察、测量、测试或者推理后通过检验，则称该对象是已知的，否则称其为未知的。

鬼魂或者幽灵可以称为一个未知对象的例子。

概念定义 15 (显式（manifest），隐藏（latent）结构或属性)**.** 如果对象的结构或者属性可直接观察、测量或者测试，称之为显式（*manifest*）结构或属性，如果不能直接观察、测量或者测试，则称之为隐藏（*latent*）[2]结构或属性。

取决于不同情况，对象的属性或者结构有时是显式的，有时则是隐藏的。

概念定义 16 (显式，隐藏机制)**.** 如果对象的机制可以复制（*replication*），则称之为显式机制，否则称之为隐藏机制。

概念定义 17 (显式，隐藏，部分显式（paritial manifest），部分隐藏对象)**.** 如果一个对象的全部属性、结构和行为可以观察、测量或者测试，全部机制可以复制，则称之为显式对象。如果一个对象的全部属性、结构和行为不可以观察、测量或者测试，全部机制不可以复制，称之为隐藏对象。介于显式对象和隐藏对象之间的对象称为部分显式对象或者部分隐藏对象。

概念定义 18 (明确定义（well-defined）的对象)**.** 如果一个对象的状态可以完全复制，则称之为该对象明确定义（*well-defined*）。

对象的划分有六个维度：未知和已知，真随机性和表面随机性，确定和不确定、明确定义（*well-defined*）和未明确定义；内生和外生；显式和隐藏。我将在本书的后续章节陆续定义部分尚未定义的概念。

关于对象的一些例子

一个生命的例子

一个人能测量他的体重。在这里，这个人既是对象又是主体；他的体重是他的量。人体内的新陈代谢是他的内在机制[a]。

[a]关于对象的例子由王晨曦提供，王磊博士进行了进一步校订。

一个自然现象的例子

一棵苹果树每年的苹果产量可以被测量得到。苹果树是对象，人是主体。苹果产量是苹果树的一个量，可以被人观察记录。植物的开花、授粉、结果是苹果树的内在机制。

[2]Latent 在一些专业领域翻译为潜或者潜在，例如潜变量（latent variable），我觉得不直观，全书统一翻译为隐藏。取决于能否直接观察、测量或者测试，相关概念包括显式属性、隐藏属性、显式量、隐藏量、显式变量、隐藏变量。

一个抽象概念的例子

一个明星可以测量他的社交媒体粉丝数。粉丝和明星都是抽象概念，既是对象又是主体，社交媒体粉丝数是明星的一个 量。粉丝和明星的关系是一种外在的影响机制。

一个政策的例子

申请一项福利政策的公民数可以被测量。福利政策是对象，公民是主体，申请福利政策的公民数可以是福利政策这个对象的量。福利政策规定的适用公民范围是福利政策这个对象的外在机制，它约束福利政策和公民之间的关系。

3.2.5　个体、系统、总体或者样本

概念定义 19 (个体（individual），系统（system）)**.** 一个 对象可以是 个体 或系统 。个体由若干组件（*component*）构成，而系统则是由相互作用或相互依赖的个体组成的协调实体（*coherent entity*） *[1, 18, 339]*。

系统具有可递归性（recursive）。最复杂的系统就是世界。在本文后续表述中，若无特别说明，“对象”一词将不区分个体与系统，也就是说，对象既可以是个体，也可以是系统。

对象拥有众多实例 （*instance*）。

概念定义 20 (总体（population），样本 (sample)，参数 (parameter)，统计量 (statistic))**.** 总体指对象实例的完整集合，样本 则是从总体中抽样的对象实例子集 *[292]*。参数是描述总体某个量（*quantity*）的数值， 统计量 则是描述样本某个量的数值。

3.3　事件

3.3.1　事件的定义

基本概念定义 3 (事件 (event)，现象（phenomenon）)**.** 事件是对象及其状态或者对象关系变化（*change*）的结果。对象变化可以是对象状态的变化、对象的消失或者新对象的产生。现象实质上是一组事件的集合。

事件也可以视为衍生对象。

经典的概率论教材通常将事件定义为随机实验（random trial）样本空间的一个子集，表示一次试验中可能发生的一个或多个结果 [159]。我将经典概率论教材中事件的定义修改为 “事件是对象或者对象关系变化的结果的样本空间中的一个子集”。Ω 是所有可能事件的总样本空间，$\mathcal{F}$ 是表示事件的 Ω 的子集集合。

3.3.2 变量和随机变量定义

概念定义 21 (变量（variable), 显式变量 (manifest variable), 隐藏变量 (latent variable)). 变量 (*variable*) *[326]* 是表达对象未知或者变化的量 *(quantity)* 的符号。对象可以直接测量或者测试的变量称为显式变量 *(manifest variable)*，否则称为隐藏变量 (*latent variable*)。

概念定义 22 (随机变量). 随机变量 X *[159]* 是一个可测函数 $X : \Omega \to \mathbb{R}$，它从一个概率空间 $(\Omega, \mathcal{F}, P)$ 扩展至 $\mathbb{R}$，并为每个 $\omega \in \Omega$ 分配一个数。在此公式中，P 表示概率。

随机变量 X 为每个样本点 $\omega \to \mathbb{R}$ 分配的是一个 实数 ，该数属于实数集 $\mathbb{R}$。它本质上是将一个随机事件的结果 “数量化”，以便进行数学分析。例如：在抛硬币实验中，若定义 $X(\omega) = 1$ 表示正面朝上，$X(\omega) = 0$ 表示反面朝上，则分配的数是 0 或 1。此时 $X(\omega)$ 是一个离散型随机变量。在测量某天降雨量的试验中，$X(\omega)$ 可能是某个非负实数 (如 12.5 毫米)，表示实际观测到的雨量，属于连续型随机变量。从数学结构上看，随机变量 X 是定义在样本空间 Ω 上的实值可测函数 (measurable function)，其作用是将抽象的样本点 ω 映射为具体的数，从而使得我们可以用概率论工具来研究这些数值出现的可能性。这种数量化处理是连接随机现象与数学分析的关键步骤 [159] 。

3.3.3 事件的概率定义

概念定义 23 (概率). 概率 P 是定义在事件集合 $\mathcal{F}$ 上的概率测度 (*measure* [3])。在概率空间 $(\Omega, \mathcal{F}, P))$ 中，P 的赋值有着严格的公理化定义 P 是一个从事件集合 $\mathcal{F}$ 到区间 *[0,1]* 的函数，必须满足以下三条柯尔莫哥洛夫 (*Andrey Kolmogorov*) 公理 *[159]*:

1. 非负性公理 (*Non-negativity*): $P(\omega) \geq 0$，对于任意事件 $\omega \in \mathcal{F}$。
2. 规范性公理 (*Normalization*): $P(\Omega) = 1$
3. 可列可加性公理 (*Countable Additivity*) [4]: 如果事件序列 $\omega_1, \omega_2, \omega_3, \ldots$ 两两互不相交 (*disjoint*) (即对于所有 $i \neq j$，有 $\omega_i \cap \omega_j = \emptyset$)，则这些事件并集的概率等于各事件概率之和。

$$\text{若} \omega_i \cap \omega_j = \emptyset \quad (\forall i \neq j), \text{则}: \; P\left(\bigcup_{i=1}^{\infty} \omega_i\right) = \sum_{i=1}^{\infty} P(\omega_i). \tag{3.1}$$

根据这些定义，概率可以视为衍生对象的量，这里的衍生对象是事件。柯尔莫哥洛夫 (Andrey Kolmogorov) 公理本质上是为概率这个衍生对象的量进行赋值的公理。

[3] 概率测度中的 “测度” 与 “测量” 在概念上有密切联系，但属于不同层次的术语，不能等同。“测度” (measure) 是数学中的抽象概念 ，用于给集合赋予一个表示 “大小” 的数值，如长度、面积、体积或概率。在概率论中，概率测度 (probability measure) 是一种特殊的测度，它将整个样本空间的 “大小” 归一化为 1，从而描述事件发生的可能性。这种 “测度” 是理论建构的核心工具，强调的是数学结构和公理化定义。“测量” (measurement) 在本书中有精准的定义。

[4] 尽管 Countable Additivity 的中文被翻译为可列可加性。需要指出的是在前面定义量 (quantity) 的时候，用的是 coutable property，可计数的属性。在本意上，这两个 countable 都是一个意思，实际上是表示可以被赋值 (计数)。

3.3.4 分布函数定义

概念定义 24 (分布函数)**.**
对于随机变量 X，其分布函数 *[159]* 是一个函数 $F_X : \mathbb{R} \to [0,1]$，定义为 $F_X(x) = P(X \leq x)$，该函数非递减、右连续，且满足 $\lim_{x\to-\infty} F_X(x) = 0$ 和 $\lim_{x\to+\infty} F_X(x) = 1$。

3.3.5 事件的独立性定义

概念定义 25 (事件的独立性（independence）)**.** 事件的独立性描述的指两个事件 ω_A 和 ω_B 之间的关系，如果它们满足 $P(\omega_A \cap \omega_B) = P(\omega_A) \cdot P(\omega_B)$，则称事件 ω_A 与 ω_B 相互独立 *[159]*，记为 $\omega_A \perp \omega_B$。

事件作为对象改变的结果，可以视为衍生对象。独立性描述的是衍生对象之间的关系，也间接地反映了对象之间的关系。

概念定义 26 (两个随机变量相互独立)**.** 如果两个随机变量 X 和 Y 的联合分布等于各自边缘分布的乘积，具体来说：若对于所有实数 x 和 y，都有

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y), \tag{3.2}$$

其中 $F_{X,Y}(x,y)$ 是 X 和 Y 的联合分布函数，$F_X(x)$ 和 $F_Y(y)$ 分别是 X 和 Y 的边缘分布函数，则称随机变量 X 和 Y 相互独立 *[159]*。

3.3.6 事件的条件概率定义

概念定义 27 (事件的条件概率（conditional probability）)**.** 事件的条件概率 *[159]* 的定义如下：
设 ω_A 和 ω_B 是两个事件，且 $P(\omega_B) > 0$，则在事件 ω_B 发生的条件下，事件 ω_A 发生的概率定义为：$P(\omega_A|\omega_B) = \frac{P(\omega_A \cap \omega_B)}{P(\omega_B)}$
其中：
$P(\omega_A|\omega_B)$：表示在事件 ω_B 已发生的前提下，事件 ω_A 发生的概率；
$P(\omega_A \cap \omega_B)$：表示事件 ω_A 与 ω_B 同时发生的联合概率；
$P(\omega_B)$：表示事件 ω_B 发生的概率，且必须大于 0，否则条件概率无定义。

3.3.7 事件的条件独立性定义

概念定义 28 (事件的条件独立性)**.** 设 ω_A、ω_B、ω_C 是定义在同一样本空间上的三个事件，且 $P(\omega_C) > 0$。若满足：$P(\omega_A \cap \omega_B|\omega_C) = P(\omega_A|\omega_C)\, P(\omega_B|\omega_C)$ 则称事件 ω_A 与 ω_B 在给定 ω_C 的条件下是条件独立的，记作 $\omega_A \perp \omega_B|\omega_C$ *[70]*。
等价地，也可以用条件概率的形式表达：$P(\omega_A|\omega_B, \omega_C) = P(\omega_A|\omega_C)$

条件独立性描述了事件之间的关系。条件独立性描述的是衍生对象之间的关系，也间接地反映了对象之间的关系。

3.3.8　事件的相关性定义

概念定义 29 (函数)**.** 一个函数 *[326]* 是集合 X（定义域）和 Y（值域）之间的一种关系（*relation*）$f : X \to Y$，它将每个 $x \in X$ 映射到有且仅有一个（*exactly one*）$f(x) \in Y$。

根据 *[294, 339]*，“函数通常表示为 f，是一种规则，它将一个集合 X 中的每个元素唯一地映射到另一个集合 Y 中的一个元素，该元素称为 $f(x)$。在此上下文中，“定义域，称为 X，指的是函数定义的所有可能值的集合 *[294]*”。另一方面，“函数的值域，称为 $f(x)$，由当 x 在定义域内变化时 $f(x)$ 的所有可能值组成 *[294]*”。

概念定义 30 (自变量（independent variable）)**.** 自变量是“一个符号，它包含函数定义域内的任意一个数 *[294]*”。因变量（*dependent variable*）也是一个符号，用于“表示函数值域中的一个数 *[294]*”。

概念定义 31 (混杂（confounding）)**.** 混杂指在研究某个自变量（原因变量）对因变量（结果变量）的影响时，当第三个变量存在与否时，因变量的观察值会发生改变，在这种情况下通常称这个第三变量为 混杂因素（*confounder*）或 混杂变量。

概念定义 32 (事件的相关性（correlation）)**.** 设 ω_X 和 ω_Y 是两个事件，随机变量 X 和 Y 分别是可测函数 $X : \Omega \to \mathbb{R}$，$Y : \Omega \to \mathbb{R}$；可测函数从一个概率空间 $(\Omega, \mathcal{F}, P)$ 扩展至 $\mathbb{R}$，并为事件的每个样本点 $\omega \in \Omega$ 分配一个数。事件 ω_X 和 ω_Y 的相关性可以用随机变量 X 和 Y 之间的相互关系来衡量，用皮尔逊相关系数 r 表示 *[225]*。其计算公式为：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3.3}$$

其中 $\bar{x}$ 和 $\bar{y}$ 分别为 X 和 Y 的样本均值。

混杂和相关性分别描述自变量和因变量以及事件之间的关系，也就是衍生对象之间的关系，间接地反映了对象之间的关系。

3.4　主体

概念定义 33 (被动对象 (passive object) 或者非生命 (non-life)，生命 (life))**.** 被动对象或者非生命是只能依靠外部机制发生变化的对象。生命是能受内部机制影响发生变化的对象。

概念定义 34 (自由意志 (Free will))**.** 自由意志是指做出自由且有意愿选择的潜能（*capacity*）与实际能力（*capability*）。

自由意志存在不同的程度。

猜想 2 (自由意志第一猜想)**.** 从对象行为的求索角度来看，自由意志介于表面随机性与真随机性之间。

概念定义 35 (心智(Mind))**.** 心智是指能够超越感官经验——也就是实证经验(*Empirical experience*)——的潜力与实际能力。

基本概念定义 4 (求索 (interrogation))**.** *求索是理解对象及对象关系的潜力（capacity）、实际能力（capability）和过程（process）。求索是一种心智能力。*

在评价学的英文著作里，我用的是 interrogation，英文相近的词语有 Inquiry 或 enquiry [5]，中文可以翻译为“探寻”，本书翻译为“求索”，取自屈原名句“路漫漫其修远兮，吾将上下而求索”。在第 二部分，我将重点讨论理解对象的能力、方法和过程。对象之间的相互影响是对象关系的一种，我将在第 三部分进行讨论。

概念定义 36 (自动对象（automatic object），智能生命（intelligent life）)**.** *自动对象是指能够进行求索但缺乏自由意志的对象。智能生命则是指具有心智能力并拥有自由意志的生命。*

猜想 3 (智能生命猜想)**.** *心智能力与自由意志是智能生命区别于普通生命的本质机制。*

基本概念定义 5 (主体（subject）)**.** *主体要么是一个自动对象，要么是一个智能生命。*

主体自身也可以是求索对象，主体也可以求索其自身，例如一个人求索他自己。

概念定义 37 (人造物(artifact))**.** *人造物指的是来自主体猜想 (conjecture)、设计(design)和制造（fabrication）的对象。*

概念定义 38 (利益相关方（stakeholder）)**.** *利益相关方是指对求索对象负有责任或持有利益的智能生命，或由智能生命组成的系统。*

求索过程本身也可以是一个对象，因此求索过程也可以成为被求索的对象，可以称之为元求索（*meta-interrogation*）。

一次求索包括求索对象 (interrogated object)、求索条件 (interrogation condition)、求索主体 (interrogation subject)、求索方法学 (interrogation Methodology)、求索工具 (interrogation Instrument)、求索过程 (interrogation Process) 以及求索结果 (interrogation Outcome)。

概念定义 39 (求索系统（interrogateion system），求索对象（interrgoated object），求索条件（interrogation condition），求索主体（interrogation subject），求索方法学, 求索工具, 求索过程)**.** *求索系统是参与求索的所有对象及其关系的总和。被求索的对象或者衍生对象是求索对象。求索条件是对求索对象进行求索的上下文（settings），是参与求索的系统中除了求索对象以外的部分。求索主体（interrogation subject）是具体从事求索的对象，主体在将在后文正式定义。求索方法学是求索基础协议（protocol ）。求索工具是实现求索方法的工具。求索过程是参与求索的系统所有对象产生的事件的总和。*

求索对象可以是某个具体的对象或其结构、属性和机制、或者对象与其他对象之间的关系以及其他衍生对象。

概念定义 40 (求索结果（interrogation outcome）)**.** *求索结果可以是对象的结构、属性、机制和行为的观察、测量或测试结果，或者是对象结构、属性、机制、行为或者对象本身的命题或模型，或者是对象关系的命题和模型。求索结果同时受求索对象、求索条件、求索主体、求索方法学、求索工具以及求索过程的影响。*

[5]在参考文献 [224], Interrogation 一词被广泛使用，例如”The whole art and practice of scientific experimentation is comprised in the skillful interrogation of Nature”。

概念定义 41 (数据（data）)**.** 数据则是求索主体在不同求索条件下获得的求索对象的原始求索结果或衍生结果。

求索有不同的复杂度。观察（*observation*）和实验（*experiment*）都是一种求索。

概念定义 42 (实验（experiment），观察（observation）和随机实验（random experiment）)**.** 实验[6]可以改变求索条件或求索对象。观察无法改变求索对象或求索条件。随机实验同时改变求索对象与求索条件，并且依赖于真随机化机制将求索对象分配给求索条件。

测量、测试、推理和评价是四种基础的求索。本书将分别在第 五章、第 六章、第 七章和第三部分讨论它们。测量、测试、推理和评价是根据求索的功能（function）和角色（role）划分的，而观察与实验则是根据求索对象与求索条件之间的关系（interrelationship）来划分的。

一些求索的例子

一个测量的例子

一个人能测量他的体重。这个人既是对象又是主体；他的体重是量。

一个测试的例子

一个人可以测试他的视力。这个人既是对象又是主体；他的视力是量。标准视力表是检验标准。

一个推理的例子

大前提（普遍的）：所有人都会死。小前提（特殊的）：爱丽丝是人。结论：因此，爱丽丝会死。

3.5　因果：原因和影响

概念定义 43 (对象内部机制的影响)**.** 考虑仅存在单个对象 A 的理想环境（*setting*），不存在其他对象，或者其他对象的影响可以被忽略；对象 A 自身发生的可观察、可测量、可测试的差异，称为对象内部机制对其产生的影响。

内部机制的影响可以用第 1.3节定义的存在算子表示为 $A \mid ext(A)$。这里的 $ext()$ 是一个算子，表示相对于一个对象不存在时，该对象的存在产生的影响，是一个物理操作。

基本概念定义 6 (原因对象（UO），影响对象（AO），影响)**.** 考虑由对象 A 和 B 组成的理想环境，A 或 B 不受内部机制的影响，且不存在其他对象，或者其他对象的影响可以被忽略；相对于 A 不存在时，A 的存在导致 B 中可测量、可测试或可观察到差异。A

[6]在中文，实验和试验的含义有所区分，“实验” 更强调 探索性与科学性 ，常用于科研、教学或理论验证场景。“试验” 更侧重 实用性与结果导向 ，常用于工程、产品开发或性能检测。《现代汉语词典》指出：“实验” 重在原理验证，“试验” 重在性能考察。由于含义接近，本书不作区分，取更广泛的用法：实验。

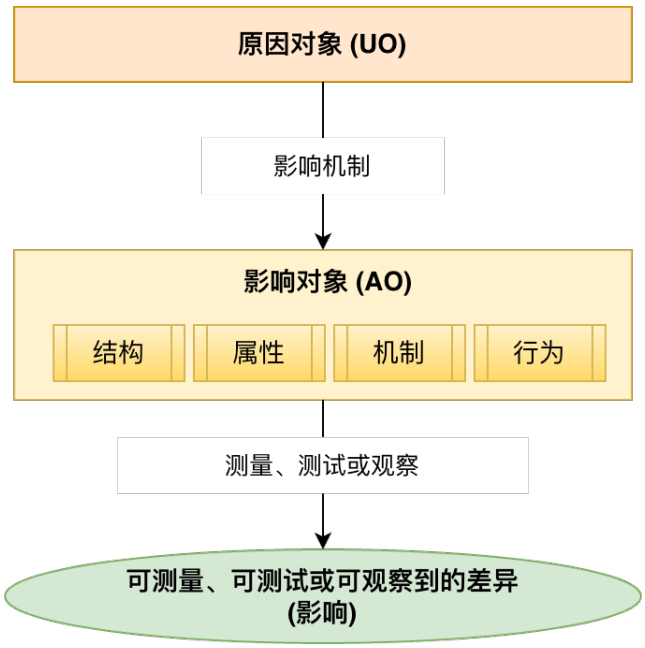

图 3.2: 原因对象、影响对象、影响机制和影响之间的关系。

是原因对象 *(cause object, UO)*，简称原因或者因 *(cause)*，B 是被影响对象（*affected object)*, 简称 AO 或影响对象，B 中可观察、可测量、可测试的差异称为 A 对 B 的影响（*effect*）。

影响是一种变化。A 对 B 的影响可以是 B 结构、属性、机制或者行为的变化，或者 B 的消失，或者新对象 C 的产生。可以用第 1.3节定义的存在算子表示为 $B \mid ext(A)$。

一个对象 A 可以影响多个对象，可以简单地记为 $\mid \mathrm{ext(A)}$。

概念定义 44 (影响机制)**.** 影响机制是原因对象对影响对象施加影响的方式。影响机制是一种外部机制。

原因对象对影响对象的影响可以是持续的，也可以是间断地施加不同的影响。当提到 对象 A 和 B 之间的相互影响 时，指的是 A 对 B 的影响和 B 对 A 的影响。图 3.2 展示了原因对象、影响对象、影响和影响机制之间的关系。

3.5.1　混杂对象定义

概念定义 45 (混杂（confounding）对象)**.** 研究对象 A 对对象 C 的影响 $C \mid ext(A)$，当对象 B 和 A 对 C 的影响纠缠在一起无法隔离时，则称对象 B 是相对于对象 A 和 C 的混杂（*confounding*）对象，简称混杂对象。

3.5.2　确定和随机性定义

概念定义 46 (确定的 (determined) 求索对象)**.** 一个求索对象在一个求索条件的一个具体配置下有且仅有一个符合事实的求索结果，也就是真求索结果，则称该求索对象为确定的。测量的真值就是一个真求索结果。

概念定义 47 (求索结果的真值)**.** 如果一个求索结果是确定的，真值是在理想求索条件下获得的求索结果，它是求索的隐藏量。

概念定义 48 (对象的属性具有真随机性)**.** 如果一个对象的属性是可观察、可测量或可测试的（显式的），但它的结果是不确定的，则称该对象的属性具有真随机性（*true randomness)*。

请注意，隐藏属性也不能通过测量或者测试确定。

概念定义 49 (对象的结构具有真随机性)**.** 如果一个对象的结构是可观察、可测量的（显式的），但结构是不确定的，则称该对象的结构具有真随机性（*true randomness)*。

概念定义 50 (机制具有真随机性)**.** 如果一个机制在给定的输入情况下，其输出是不确定的，则称该机制具有真随机性（*true randomness)*。

概念定义 51 (对象的行为具有真随机性)**.** 如果一个对象的行为是可观察、可测试的（显式的），但其结果是不确定的，则称该对象的行为具有真随机性（*true randomness)*。

概念定义 52 (对象不具有真随机性)**.** 如果一个对象的属性、结构、机制和行为不具有真随机性，则称对象不具有真随机性，是确定的。

概念定义 53 (对象的属性具有表面随机性)**.** 如果一个对象的属性是确定的，但其求索结果仍然受未知的混杂对象或者已知对象的混杂因素、求索条件、求索过程、求索方法学、求索工具或者求索主体的影响。在这种情况，它的求索结果是无法完全确定的，则称该对象的属性具有表面随机性（*pseudo randomness*）。

概念定义 54 (对象的结构具有表面随机性)**.** 如果一个对象的结构是确定的，但其求索结果仍然受未知的混杂对象或者已知对象的混杂因素、求索条件、求索过程、求索方法学、求索工具或者求索主体的影响。在这种情况，它的求索结果是无法完全确定的，则称为该对象的结构具有表面随机性（*pseudo randomness*）。

概念定义 55 (机制具有表面随机性)**.** 如果一个机制在给定的输入情况下，其输出是确定的，但其求索结果仍然受未知的混杂对象或者已知对象的混杂因素、求索条件、求索过程、求索方法学、求索工具或者求索主体的影响。在这种情况，它的求索结果是无法完全确定的，则称该机制具有表面随机性（*pseudo randomness*）。

概念定义 56 (对象行为具有表面随机性)**.** 如果一个对象的行为是确定的，但其求索结果仍然受未知的混杂对象或者已知对象的混杂因素、求索条件、求索过程、求索方法学、求索工具或者求索主体的影响。在这种情况，它的求索结果是无法完全确定的，则称为该对象的行为具有表面随机性（*pseudo randomness*）。

3.5.3　因果律公理

公理 1 (因果律公理（Axiom of cause and effect），确定的影响公理（axiom of dermined effect）)**.** 如果对象及对象之间的影响机制不具有真随机性，则对象之间的影响是确定的。

无论是在什么条件下，除非原因对象、影响对象和前者对后者的影响机制具有真随机性，否则原因对象对影响对象的影响是确定的，具有真值。

但这并不意味着这个影响可以直接被测量或者测试。例如，影响对象不是一个显式对象，量的变化不能被直接测量或者测试；或者有多个混杂对象的影响无法隔离，使得总体影响（overall effect）无法与关注的影响分离等。

并非所有对象的影响是确定的。例如，幽灵有可能不存在，但它也是一个对象，而且能够对其他对象产生影响，有时候能吓死人，但其影响是不确定的，具有不可计算性。

概念定义 57 (对象具有不可计算性)**.** 对象具有不可计算性是指不能建立对象的模型（简称建模），尤其是不能建立数学模型。

一些关于原因、AO、影响以及影响机制的例子。

一个物理学例子

伊萨克 · 牛顿的苹果树：地球是原因，苹果是 AO，重力（万有引力）是影响，万有引力定律是影响机制。

一个化学例子

拉瓦锡的燃烧实验：氧气是原因，可燃物是 AO，燃烧是影响，燃烧的化学方程式是影响机制。

一个生物学例子

格雷戈尔 · 约翰 · 孟德尔的豌豆：亲本基因是原因，子代基因是 AO，子代基因/性细胞的产生是影响，基因的分离定律和自由组合定律是影响机制。

一个计算机科学例子

CPU 评价：CPU 是原因，计算机系统是 AO，CPU 对计算机系统上测量得到的整体性能指标值的影响是影响，CPU 和计算机系统中其他不可或缺组件的协作运行机制是影响机制。

一个人工智能例子

评价一个人工智能算法：人工智能算法是原因，AO 包括标好事实和客观真相标签的数据集、算法实现、操作系统、处理器、内存。人工智能算法对整体算法准确率指标的影响是影响，算法拟合输入-输出关系的能力是影响机制。

一个医学例子

药物评价：药物是原因，患者是 AO，患者身上测量或测试得到的差异是影响，一些已知的或未知的生理机制是影响机制。

一个社会科学例子

政策评价：政策是原因，政策实施过程中的不同参与者是 AO，政策对整体评价结果比如说支持率的变化是 影响，影响机制有多种表现形式，如政治、社会或心理上的表现。

概念定义 58 (计算复杂性（computational complexity）)**.** 计算复杂性研究在给定数学模型下求解或验证问题所需的计算资源，常见资源包括时间和空间资源。

复杂性类（*complexity class*）把资源需求具有相似性质的问题组织为问题集合 *[282]*：*P* 类（*polynomial time class*）问题是其数学模型可由确定性机器在多项式时间内完成求解的问题；*NP* 类（*nondeterministic polynomial time class*）问题是在给定候选解的条件下，其数学模型可由确定性机器在多项式时间内完成验证的问题；*co-NP* 类（*complement of NP class*）刻画 *NP* 问题的补问题；*NP-hard* 类（*NP-hard class*）刻画至少与 *NP* 中最难问题一样难的问题；*NP-complete* 类（*NP-complete class*）刻画同时属于 *NP* 和 *NP-hard* 的问题。

这里验证的含义等同于第 六章定义的检验。

3.5.4　事件的原因独立性定义

概念定义 59 (原因独立性（cause independence）)**.** 事件 ω_1 和 ω_2 分别是对象或者对象关系变化的结果。记影响事件 ω_1 和 ω_2 的原因对象集合分别为 $\mathbb{A}_1$ 和 $\mathbb{A}_2$。如果事件 ω_1 和 ω_2 具有原因独立性，则以下公式成立：

$$\omega_1 \perp\!\!\!\perp \omega_2 \to \mathbb{A}_1 \cap \mathbb{A}_2 = \varnothing \tag{3.4}$$

事件的原因独立性不同于 3.3.1节定义的事件独立性概念。因此，我记事件的原因独立性符号为 $\perp\!\!\!\perp$，区分事件的独立性符号 $\perp$。

在概率论的初级教材里，通常认为一个人连续两次抛硬币或者掷骰子是独立的事件。然而，一个人连续掷筛子或者抛硬币有着几乎相同的原因对象集合，从原因对立性的角度来说，它们不是独立的，抛硬币或者掷骰子不是随机事件，它们表现的只是表面随机性。其结果之所以不能完全确定是因为作用的原因对象数目多，相互影响复杂。

另外一个考察原因独立性的例子就是交通。举个例子来说，我住在交通拥塞的北京，我和我的一个同事，两个人在不同的时间从不同的地点（各自的家）出发，开车去单位。这两个事件之间（开车花费的时间是结果之一）是独立的吗？从原因独立性的角度来说，这两个事件显然不是原因独立的，因为这两个事件是由大量对象影响的结果，而且可以确定的是部分对象之间有交集。

3.5.5 事件的条件原因独立性定义

概念定义 60 (条件原因独立性)**.** 事件 ω_1，ω_2 和 ω_3 分别是对象或者对象关系变化的结果。记影响 ω_1，ω_2 和 ω_3 的原因对象集合分别 $\mathbb{A}_1$，$\mathbb{A}_2$ 和 $\mathbb{A}_3$，如果事件 ω_A 与 ω_B 在给定 ω_C 的条件下是条件原因独立的，则以下公式成立：

$$\omega_A \perp\!\!\!\perp \omega_B | \omega_C \to (\mathbb{A}_1 \setminus \mathbb{A}_3) \cap (\mathbb{A}_2 \setminus \mathbb{A}_3) = \varnothing \tag{3.5}$$

符号 $\setminus$ 表示 差集 运算，$(\mathbb{A}_1 \setminus \mathbb{A}_3)$ 即从集合 A_1 中移除所有属于集合 A_3 的元素。事件 ω_A 与 ω_B 在给定 ω_C 的条件下是条件原因独立的。条件原因独立性的符号为 $\perp\!\!\!\perp$ |。

3.6 命题

基本概念定义 7 (命题 (proposition))**.** 命题 是关于对象或者对象关系可测试 (*testable*) 的陈述。

对象的命题通常是关于对象或其结构、属性、机制和行为的陈述。我将在第 六章正式定义“可测试”的具体含义。

概念定义 61 (前提 (premise), 假说(assumption), 假设(hypothesis))**.** 前提 (*premise*) 是作为逻辑推导起点的命题 *[55]*。假说(*assumption*)是被接受为事实(*fact*)或真理(*truth*)而无需证明的命题, 是在特定逻辑论证中被采用的 临时性前提 (*provisional premise*) *[55]*。假设 (*hypothesis*) 是关于对象的命题 *[189]*。

假设 (hypothesis) 与假说 (assumption) 之间的区别如下。
其中 P 是一组前提 (premises) :

- 假说是 P 中的元素: $P = \{A_1, A_2, \ldots, A_n\}$.
- 假设是从 P 中推导出的可测试的陈述: $H \subseteq \mathcal{P}(P)$, 读作“$P$ 为一个集合，H 是 P 的幂集的一个子集”。
- H有效性成立，当且仅当: $(\bigwedge_{i=1}^{n} A_i) \to H$, 读作“$H$ 有效性成立，当且仅当从 $i = 1$ 到 n 所有 A_i 的合取蕴涵 H.”

书法体大写字母 $\mathcal{P}$ 表示一个集合，其中包含多个 (有限或无限) 命题。$\mathcal{P}$ 通常是 P 的 幂集 (power set)。给定集合 P 的幂集是指一个由 P 的所有可能子集 (包括空集和 A 本身) 组成的新集合，记为 $\mathcal{P}(P)$ 或 2^P 或 $\mathcal{P}$。

3.7 模型

基本概念定义 8 (模型 (model)，建模（modeling）)**.** 模型 是对象或者对象关系的简化表示 *(streamlined representation) [334, 339]*。建立模型的过程称为建模。

可以对对象本身、对象结构、对象机制、对象行为或者对象的关系进行建模。

模型可以表现为物理实体、数学结构或其他形式（例如视觉形式）的构造（construct）。

与对象的命题相比，对象的模型通常更为复杂，能表达对象自身、对象的结构、机制、行为或者对象之间的关系。

概念定义 62 (数学模型)**.** 数学模型 是对象或者对象关系的数学简化表示（*representation*），通常用函数或方程来表达。

概念定义 63 (解释（explanation）)**.** 解释（*explanation*）本质上是命题或者模型。当它为动词时，是提出命题或者模型。

概念定义 64 (物理执行，虚拟执行)**.** 物理执行指在物理世界中驱动对象的行为。虚拟执行指无法在物理世界中驱动对象的行为，只能在数学模型上进行操作。

3.8 公理系统

概念定义 65 (事实（fact），真理（truth），知识)**.** 一个关于对象或者对象关系的命题或模型如果可由对象以外的主体检验（*test*）为真，这个命题或者模型即事实（*fact*）或真理（*truth*）。知识 是所有对象或者对象关系的事实或真理的集合。

猜想 4 (知识来源猜想)**.** 知识有且仅有两个来源：一是智能生命的感官经验；二是智能生命的心智能力，均需要通过检验。

在本书中，我将 *Ground Truth* 翻译为基准事实，它只是一种习惯性说法，本质上就是事实。

基本概念定义 9. 公理系统（*axiom system*）是一组自洽的假说（*self-contained assumptions*）集合，基于这组假说可以构建特定领域或超出特定领域的知识、逻辑系统或者形式系统。本节中简称的系统指知识、逻辑系统或者形式系统。

基于公理构建系统的一种方式是证明。对于某个命题 A，有关 A 的证明代表一个命题序列，该序列从公理开始，以 A 结束，并遵守推理规则。公式 $B \vdash A$ 表示从 B 出发至少存在一个对 A 的证明。

本书建立在公理系统之上。后两小节将从不同角度解释什么是公理系统。

3.8.1 古典视角

从古典视角来看，公理是不证自明的假说（assumption）。古典视角植根于欧几里得几何 [114] 和理性主义哲学（rationalist philosophy），例如笛卡尔的哲学 [73]。古典视角下的公理的核心观点有：

- *不证自明*: 公理是系统中预先接受的基础假说集合，不需要由系统中的其他命题推导得到。$\forall A \in \mathcal{A}_{\text{classical}}, \text{Proof}(A) = A$，这里 $\mathcal{A}_{\text{classical}}$ 表示古典视角下的基础假说集合。
- *普遍真理*: 它们被认为是普遍、无可置疑的。公理是先验的 [7]和普遍有效的：

$$\vdash A \quad \text{for all } A \in \mathcal{A}_{\text{classical}}. \tag{3.6}$$

一个典型的公理系统例子

欧几里得第一个公理：“任意两点之间可以连成一条直线段。”

3.8.2 现代视角

从现代视角看，公理并不必然是“不证自明”或绝对的真理，而是具有一致性和相对性的基础假说集合 [123, 122]。现代视角下的公理的核心观点有：

- *基础假说（assumption）*：基础假说是构建不同的系统而选择的起始点。
- *公理具有一致性*: 从集合 $\mathcal{A}$ 中的任何基础假说出发，都无法推导出矛盾。
- *公理具有相对性*: 构建不同系统时，选择的起始点可以不同。在一个系统里“不证自明”，在另一个系统里可能不成立。

一个典型的例子。

- 欧几里得第五公理（平行公理）：“通过不在直线上的点，有且只有一条平行线。”
- 双曲几何公理：“存在无限多条平行线。”

尽管两者相互矛盾，但这两个系统内部都是自洽的，这表明公理是 约定（conventions），而非普遍真理（universal truths） [114, 122]。

现代数学将公理重新定义为 [123, 262]：

1. *形式化基础*： $\mathcal{A}_{\text{modern}} := \{A_1, \ldots, A_i, \ldots A_n\}$ 是现代视角下的基础假说集合。$\mathcal{A}_{\text{modern}}$ 需满足一致性，即不存在矛盾：$\nexists S \, (\mathcal{A}_{\text{modern}} \vdash S \wedge \neg S)$.

[7]先验的英文是 a priori，指不依赖经验、先于经验而成立的知识 。与 a priori 对应的是 a posteriori（后验的），指必须通过经验、观察或实验才能确认的知识。

2. 公理选择的相对性：公理是约定，可以有不同的选择，而非真理。例如，欧几里得与非欧几里得几何不同的公理选择。

3.8.3　关键不同

经典观点	现代观点
公理 ⊆ 普遍真理	公理 ⊆ 约定
不证自明	依赖于系统本身（System-dependent）

表 3.1: 有关公理的经典观点与现代观点

表 3.1总结了有关公理的经典观点与现代观点的关键差异。

3.9　总结

本章定义了一个精简和统一的基础概念系统，包含对象、关系、原因和影响、事件、求索、主体、命题、模型和公理系统，它是第 二部分定义的求索的科学的基础概念。

第四章

三个世界假说

本章中，我将详细阐述三个世界假说（*Assumption of Three Worlds*）。

基于直觉，我认为真随机性（true randomness）、因果律（表面随机性）以及自由意志（free will）（介于真随机性和表面随机性之间）是三种支配世界变化的根本机制（mechanism）。我在本章将其总结为假说。这三种机制需要用不同的方法建模。

4.1 假说

基本假说 1 (三个世界假说 (Assumption of Three Worlds))**.** *存在三个世界：微观对象世界、被动对象世界 和 生命世界，它们由一致但不同的规律（law）支配。*

概念定义 66 (规律)**.** *规律是验证过的模型。*

4.2 微观对象世界

概念定义 67 (微观对象世界 (Micro object world))**.** *微观对象世界由 微观对象组成，微观对象的尺度在原子和亚原子粒子级别。微观对象世界受量子力学原理支配 [31, 195]* [1]。

微观对象具有四个通常被称为 量子效应 的特性，这些特性从根本上不同于其他对象的特性。

第一个特性是 量子化，也就是说，微观对象的能量、动量、角动量及其他可计数属性并非连续，而是以离散的形式存在，称为“ 量子（quanta）”。

第二个特性是波粒二象性。微观对象，如电子和光子，既表现出粒子性，例如具有确定的位置、会发生碰撞，也表现出波动性，例如会发生干涉和衍射。

第三个特性是量子叠加态。一个微观对象，或称为量子系统，如电子或光子，在被测量前可以同时处于多个可能状态的组合中。测量后，它会“坍缩”为这些可能状态中的某一个。

一个著名的例子是薛定谔的猫（Schrödinger's cat） [269]。在盒子被打开并观察之前，“薛定谔的猫”可以同时处于既生又死的状态。

[1] 王晨曦调研了量子力学的原理。

第四个特性是量子纠缠。两个或更多微观对象，如粒子，以某种方式纠缠（entangled）在一起，测量其中一个粒子会瞬间确定其他粒子的状态，无论它们相距多远。

一些微观粒子“奇怪”性质的例子 Some Examples on the “Strange” Properties of Microscopic Objects

电子在能级间跃迁 Electrons Jump Between Energy Levels

原子中的电子只能占据特定的能级。它们通过吸收或发射一个光量子（光子）在这些能级之间跃迁 [89]。 [a]

[a]一些微观粒子“奇怪”性质的例子由王晨曦提供，王磊博士进行了进一步校订。

双缝干涉实验 Double-Slit Experiment

在双缝干涉实验中 [45]，一个射向双缝的电子表现出波的特性，并在屏幕上形成干涉图样，仿佛它同时穿过了两条缝。然而，它又像粒子一样击中屏幕上的一个点。

一对量子纠缠的光子 Photons in an Entangled Pair

当测量并确定一对量子纠缠的光子中一个光子的偏振方向时 [35]，另一个光子的偏振方向会立即以非局域和瞬时的方式关联起来 [332]。

量子隧穿效应 Quantum Tunneling Effect

在半导体器件中，电子可以穿过高于自身能量的势垒，这种现象称为量子隧穿效应 [328]。即使电子的能量低于势垒的高度，它仍然有一定的概率出现在势垒的另一侧。

4.3　被动对象世界

概念定义 68（被动对象世界（passive object world））. *被动对象（passive object）是正常尺度下的对象，在没有外部机制的作用下，被动对象不会发生变化。被动对象世界（passive object world）由被动对象组成，受因果律支配。*

我已经在第三章定义了什么是原因和影响，并提出了因果律公理。简单来说，我定义的因果律给出：在什么条件下，影响是确定的。

微观对象具有真随机性，也就是说微观对象的量值 (quantity value) 本身是不确定的。而被动对象是确定的。具体来说，被动对象的属性、结构、机制和行为都是确定的。被动对象表现出的随机性是一种表面随机性，受因果律支配。一个被动对象的求索结果，例如测量值，之所以表现出随机性，是因为它的量受多种因素影响，如第 3.5.3节阐述的。

在概率论的经典教材里，通常认为抛某一枚硬币是一个随机实验。从求索的角度来理

解，求索对象是硬币，求索主体是抛硬币的人，求索条件包括抛硬币的人对硬币施加的不同的力、空气、风、地面等。每次求索的结果是对象的一个事件：硬币掉落在地上时，是正面朝上还是正面朝下。

以抛掷硬币为例，它受确定的力学规律影响，结果是确定的。之所以表现出随机性，是因为复杂的求索条件无法精准复制（replicate）。从求索的科学的角度看，这属于一种表面随机性。

4.4　生命世界

概念定义 69 (生命世界（life world）)**.** 生命世界（*life world*）由普通生命和智能生命组成。普通生命不仅被动接收原因对象的影响，而且会在内部机制作用下发生变化。相对于普通生命，智能生命还具有自由意志，不仅被动接收原因对象的影响，在内部机制作用下发生变化，而且还能做出自由且有目的的选择。

自由意志的“自由”不同于微观对象的真随机性和被动对象的表面随机性，它的本质特性是有意图的行为。见自由意志猜想。

自由意志的典型例子　包括动物和人类日常生活中的决定，人类社会中的爱、写作、科学研究、艺术、投票、商品购买以及大学或项目申请。

4.5　关于根本机制的讨论

我在引言里提到，真随机性、因果律和自由意志是主导世界的三大根本机制。它们需要以不同的方式建模。

物竞天择是在四种机制同时作用的过程和结果。例如，微观对象的真随机性会导致生命的基因发生变异，而在更大的尺度上，被动对象因为多种因素的复杂影响，也可能在不同的条件下发生不同的变化。普通生命会因为内部机制发生变化，而智能生命拥有自由意志，也会做出有意的改变。这些机制都会导致变化，而生命在微观物质层次上的结构或者机制的改变正是进化的基础。

我将其总结为以下猜想：

猜想 5 (自然界进化猜想)**.** 自然界的进化同时由微观对象世界的真随机性、被动对象世界的因果律（表面随机性）以及生命世界的内部机制和自由意志决定。

4.6　本章总结

本章阐述了三个世界假说，提出支配世界的三大根本机制：真随机性、因果律（表面随机性）和自由意志（介于表面随机性和真随机性之间）。

第五章

计量学：量的定义、实现和赋值的科学

本章阐述了我对计量学的新解释，包括基本概念、问题定义、基本假说、方法学和案例研究。王磊和我实现了这些想法，王磊贡献了第5.8节、第5.9节和第5.10节。范帆达和王晨曦也对第5.10节有贡献。

在第四章，三个世界假说指出微观对象世界的量具有真随机性，本章暂时不处理真随机量的测量。

5.1 什么是计量学？

计量学概念定义 1 (计量学)**.** 计量学是量的定义、实现和赋值的科学。具体来说，是对象的结构量和行为量的定义、实现和赋值的科学。

传统的计量学定义是测量的科学及其应用 [28]，主要涉及对象的结构量。我在第 六章，将计量学拓展到对象的行为量，将其称为测试学。

计量学主要包括三个核心议题：量的定义、量的实现和量的赋值。

计量学概念定义 2 (量的定义)**.** 量的定义本质上建立量这个衍生对象的模型，它的有效性需要被检验。

根据第 三章的定义，量可以视为衍生对象。

计量学概念定义 3 (量的实现)**.** 量的实现是根据量的定义（衍生对象的模型）来实现测量标准。测量标准本质上是一个系统。

计量学概念定义 4 (量的赋值)**.** 量的赋值，也就是具体的测量，是依据建立的测量标准或者测量仪器，为测量对象的量赋值的过程。

5.2 概念定义

计量学概念定义 5 (参考 (reference))**.** 参考是一个约定 *(convention)*、标准 *(standardization)* 或公理 (*axiom*)。

我在第3.8.2节中讨论了现代意义上的公理系统也是一个约定。标准和公理都可以认为是一个参考和约定，但可以根据第 五部分的评价学方法来揭示不同的标准、公理和约定对测量结果带来的不同影响，从而设计或者优化标准、公理和约定。

计量学概念定义 6 (测量单位或者计量单位（unit of measurement）)**.** 测量单位或者计量单位是量的定义，是借助参考对象（*reference object*）定义的单位量（*unit quantity*），也是量这个衍生对象的模型。

在本书的大多数场景下，我不区分计量和测量，计量和测量都是 measurement 的中文翻译。

计量学概念定义 7 (测量标准 (measurement standard))**.** 测量标准是量的实现，也是对测量单位定义的实现（*realization*），它是一个参考，不同的实现存在准确度和开销的差异 *[28]*。反过来，量的定义可以视为测量标准的模型，也就是测量标准的简化表示。

计量学概念定义 8 (测量仪器 (measuring instrument))**.** 测量仪器是参考测量标准实现的装置，用于量的赋值。它既能以独立的物理实体存在，亦可与其他辅助设备（如延长线缆、外部传感器）组合使用 *[28]*。

计量学概念定义 9 (被测量 (measurand))**.** 被测量对象（简称测量对象）待测量的量称为被测量。

计量学概念定义 10 (测量)**.** 测量是量的赋值过程，它在一个具体求索条件下，比较被测量与测量标准规定的参考对象的单位量（*unit quantity of a reference object*），为被测量赋值。

计量学概念定义 11 (可测量性（measurable）或（measurability）)**.** 对象的可测量性是指对象的量能够与参考对象的相应量进行比较，简称可测量性。

计量学概念定义 12 (测量系统)**.** 测量系统是参与测量过程的对象与衍生对象组成的系统，包括量的定义、测量标准、测量仪器、测量主体、测量方法学和测量过程等。

在计量学中，测量通常被定义为“客观获取赋予给某个量的一个或者多个值 [28]。”在社会科学领域，测量的另一个广为流传的定义是“根据某种规则（rule），将数字（numeral）赋予给对象（object）或事件（event）” [293]。这一定义可追溯至 1946 年。我更倾向于我自己的定义，因为它揭示了测量的本质。

5.3　问题定义

我将测量问题表述如下：

问题定义 1 (测量问题定义)**.** 给定一个量（*quantity*），如何借助参考对象定义量的模型，即测量单位？如何在不同的准确度（*accuracy*）和开销（*overhead*）约束下，通过测量标准的形式实现测量单位？如何依据测量标准定义的参考，为测量对象的量赋值。

具体来说，测量问题的输入是对象的给定量；输出包括测量单位的定义以及该测量单位的实现；约束条件则是这些定义和实现所具有的准确度与开销。

在第 5.10节，我将定义测量的准确度和开销。

5.4　基本假说

测量系统的建立基于三个基本假说（assumption）。

计量学假说 1 (量的真值假说)**.** 一个对象的可计数属性，也就是量，是确定的，存在一个真值（*true value*）。

这与第 3.5.3节阐述的因果律公理是一致的。我在第 三章已经定义了什么是确定的。尽管对象的量是确定的，具有真值，因为测量系统的影响，它表现为具有表面随机性。从对象的性质角度来看，量的真值不可以被测量，是任何测量方法趋近的目标，只能通过推理得到，是一个隐藏量。

本书不讨论量具有真随机性的测量问题。它的真值可能是一个分布。

国际计量局（BIPM）组织发布的《测量不确定度表示指南》(Guide to the Expression of Uncertainty in Measurement，简称 GUM） [140] 将真值定义为 “与给定的量的定义相一致的量值”，并指出它是通过完美测量所能获得的量值。柯克普（Kirkup）和弗伦克尔（Frenkel）对真值进一步解释为：真值是对 “完全定义的被测量”（completely specified measurand），在 “完全指定的环境” 中使用 “理想仪器” 测量时，所能获得的数值 [156]。

计量学假说 2 (可利用属性假说)**.** 一个参考对象具有一个可利用的属性（*property*），它可以作为定义测量单位的基础。

参考对象源自主体的定义，是一种约定（convention）或公理。

计量学假说 3 (可获得性假说)**.** 对于任何对象的给定量，主体都能够实现借助参考对象定义的单位量，这个实现可以在不同时间、空间下被不同主体获得。

这三个假说本质上构成了一个测量系统的公理体系。

5.5　测量的基本作用

测量在主体的求索过程中发挥着重要作用。它为不同对象的同一量赋值，提供了一个普遍可获得的比较参考（comparison reference），同时这些值是建立对象或者对象关系的命题（proposition）或模型 (model)的基础，而这些命题或者模型可以成为推理（reasoning）的输入，从而生成新的命题或者模型。

5.6　方法学

方法学的核心在于量的定义、实现和赋值。图 5.1 [1] 展示了计量学的方法学。正如图 5.1 所示，在求索科学的体系下，量、测量单位（量的定义）、测量标准（量的实现）以及测量仪器构成了计量学的核心。

[1]在我绘制的草图基础上，王磊绘制了图 a. 求索科学下的计量学；王磊独立绘制了图 b. 传统计量学定义。

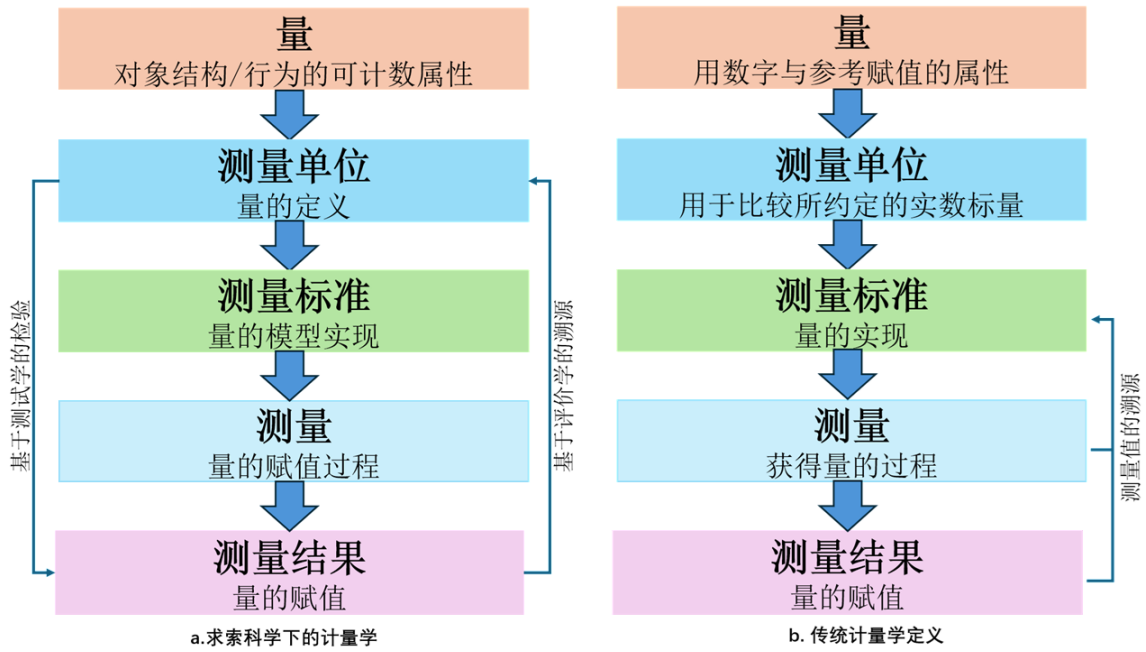

图 5.1: 计量学概念框架简明对比。传统的计量学定义参考 [28, 146] 绘制。因为空间的限制，求索科学下的计量学没有添加推理学，它可以为计量学提供更严谨的推理体系。

测量单位在测量中起着至关重要的作用 [146]。从理论上来说，测量单位是量的定义，也是测量标准的模型。从测量实践来说，测量单位是约定、定义并采用的单位量，用于表达和比较同类量。

测量标准是量定义的实现 [146]，它是用于比较同类量的参考标准。测量标准的作用在于对测量进行标准化，从而允许测量之间的比较，并确保不同的测量系统能够交流准确和可比较的量值。

测量单位和测量标准是通过国际协议确立的，这形成了测量在科学、工业和日常应用中的统一性（uniformity）和准确度（accuracy）的基础。不同的测量标准的层级结构（hierarchy）呈现从低到高的进阶关系，其准确性不断提高，而成本也随之增加。这个层级从国家标准开始，并延伸至国际标准。

计量学概念定义 13 (测量的可溯源性(measurement traceability))**.** 测量的可溯源性 *[28]* 是测量利益相关方为了保证测量的准确度对测量过程的一种要求，它要求测量结果与量的定义（测量单位）和测量标准通过有据可查、不间断的校准链建立关系，具体体现为测量仪器或者测量标准中每一级都应使用更高一级（具有更高准确度）进行校准，以帮助识别和减小系统偏差。

给量赋值的过程（测量），使用既可重复（repeatable）又可再现（reproducible）的测量方法学至关重要，这能够确保测量结果的一致性（consistency）与可靠性（reliability）。在实践中，测量借助具体的测量仪器完成。

5.7　求索科学体系下的计量学和传统计量学的差异

如图 5.1所示，求索科学体系下的计量学和传统计量学存在明显差异。传统计量学是一门研究基于实数标量对“用数字与参考值赋值的属性”进行测量的科学 [28]；而求索科学体系下的计量学则基于对象的概念，将计量学扩展为针对“对象结构和对象行为属性”进行建模的科学。同时，求索科学体系使得计量学更具有科学性，具体体现为：可以通过测试学检验测量学里的模型，例如量的定义，和推测结果的有效性，通过推理学建立计量学知识体系的严谨性；通过评价学实现测量结果相对于测量单位的可溯源性。

5.8　基本量的定义

如前所述，量是对象的可计数属性。基于参考对象的属性可以定义测量单位。例如，长度是一个基本量，其测量单位–米–目前是根据真空中的光速来定义的。在这种情况下，光起着参考对象的作用，其属性——恒定速度——被用来提供一个准确且可普遍复制的测量标准。

国际计量体系包含七个基本量：时间（time）、长度（length）、质量（mass）、电流（electric current）、热力学温度（thermodynamic temperature）、物质的量（amount of substance）和发光强度（luminous intensity）[28]。这些量构成了所有物理测量的基础，所有的物理单位和测量都是基于这些基本量来定义的。

七个基本量构成了国际单位制（International System of Units，SI）的基础，并为所有物理测量提供了根基。国际单位制通过固定七个基本（defining）物理常数（physical constants）的数值（numerical values）来定义，具体如下：

基本量定义 1 (秒 (second) (s))**.** *秒是国际单位制中的时间单位，通过取铯原子频率 $\Delta\nu_{Cs}$ 的固定数值来定义。铯-133 原子未受干扰的基态超精细跃迁频率为 $\Delta\nu_{Cs} = 9\ 192\ 631\ 770\ Hz$，这定义了一秒的时长 [72]。在这种情况下，铯-133 原子是参考对象，其超精细跃迁的精确频率被用来作为参考属性。*

基本量定义 2 (米 (meter) (m))**.** *米是国际单位制中的长度单位，通过取真空中光速的固定数值 $c = 299\ 792\ 458\ m/s$ 来定义，即 1 米是光在真空中 1/299 792 458 秒内所传播的距离 [72]。在这种情况下，参考对象是真空中的光，利用其速度固定不变的属性。*

基本量定义 3 (千克 (kilogram) (kg))**.** *千克是国际单位制中的质量单位，通过取普朗克常数的固定数值 $h = 6.626\ 070\ 15 \times 10^{-34}\ kg \cdot m^2 \cdot s^{-1}$ 来定义 [72]。在这种情况下，普朗克常数 h 为千克的定义提供了基本物理参考。基布尔天平（Kibble balance）是实现千克的一类测量标准；它通过电磁功率与机械功率的等价关系，将质量测量与普朗克常数联系起来，从而实现千克的测量。*

基本量定义 4 (安培 (ampere) (A))**.** *安培是国际单位制中电流的单位，通过取基本电荷的固定数值为 $e = 1.602\ 176\ 634 \times 10^{-19}$ 库仑 来定义，其中 1 库仑 = 1 安秒 [72]。在这种情况下，参考对象是一个质子或电子，参考属性是单个质子或电子所携带的基本电荷是绝对不变且恒定的。*

基本量定义 5 (开尔文 (kelvin) (K))**.** 开尔文是国际单位制中热力学温度的单位，通过取玻尔兹曼常数的固定数值为 $k = 1.380\,649 \times 10^{-23}\ kg \cdot m^2 s^{-2} K^{-1}$ 来定义 *[72]*。在这种情况下，参考对象是理想气体或其粒子随机热运动的热力学系统。所利用的属性是系统中粒子的平均动能与热力学温度成正比。

基本量定义 6 (摩尔 (mole) (mol))**.** 摩尔是国际单位制中物质的量的单位，通过取阿伏伽德罗常数 (*Avogadro constant*) 的固定数值为 $N_A = 6.022,140\ 76 \times 10^{23}\ mol^{-1}$ 来定义，这样 *1* 摩尔包含确切的 $6.022\ 140\ 76 \times 10^{23}$ 个指定的基本实体 (*elementary entities*) *[72]*。在这种情况下，参考对象是一组基本实体，这些基本实体可以是原子、分子、离子或其他指定的粒子。所利用的属性是一摩尔包含的基本实体具有确切数量 ($6.022\ 140\ 76 \times 10^{23}$)。

基本量定义 7 (坎德拉 (candela) (cd))**.** 坎德拉是国际单位制中光强度的单位，通过取频率为 540×10^{12} 赫兹的单色辐射的光效 K_{cd} 的固定数值来定义。这一定义意味着频率为 $\nu = 540 \times 10^{12}$ 赫兹的单色辐射，存在精确的关系 $K_{\mathrm{cd}} = 683\ cd \cdot sr \cdot kg^{-1} \cdot m^{-2} \cdot s^3$，其中球面度 sr 是国际单位制中立体角的单位。将此关系倒过来，可得出坎德拉的精确表达式：$1\ \text{坎德拉} = (K_{\mathrm{cd}}/683)\ kg \cdot m^2 \cdot s^{-3} \cdot sr^{-1}$ *[72]*。在这种情况下，参考对象是一个完全单色的光源，具体来说是一个激光器或其他能发射频率为 540×10^{12} 赫兹辐射的发光装置，这对应于绿色光。所利用的属性是这种单色光源辐射的光效，它描述了单位能量下光的感知亮度或强度。通过固定这个值，坎德拉被定义为这种辐射频率的光源所发出的光强度。

5.9　案例研究：米的定义与实现

本节以米为例来说明基本量的定义和实现。

5.9.1　关于“米”的历史定义

米的初始定义 (1793 年)　1793 年，米被定义为地球子午线四分之一的千万分之一 [228]。地球是理想的参考对象。其特性——子午线四分之一——被用来定义测量单位。

在此定义之前，欧洲使用了多种不同的单位（如英尺、英寸和里格），这些单位在不同地区差异很大。为了统一测量单位，法国科学院提议将米定义为从北极到赤道沿巴黎子午线距离的千万分之一。这一定义是首次尝试基于地球不变的自然属性来定义一个单位，这个自然属性任何国家都可以通过测量获得。尽管后来的定义提高了准确性，但最初将米与地球本身联系起来的概念，仍然是计量学思想的核心所在。

为了确定这一长度，天文学家让 · 巴蒂斯特 · 约瑟夫 · 德朗布尔 (Jean-Baptiste Joseph Delambre) 和皮埃尔 · 梅香 (Pierre Méchain) 在 1792 年至 1799 年间进行了规模宏大的大地测量，测量了从敦刻尔克到巴塞罗那的子午线弧长 [27]。这些数据构成了计算整个象限的基础 [2]。在测量过程中，1793 年根据初步结果制定了临时米制。然而，当年正式颁布的法定定义是基于测量的最终结果。

[2]完整子午线可以看作经过南北两极并环绕地球一周的闭合曲线。从赤道到极点的子午线弧占完整子午线周长的四分之一，因此称为子午线象限。敦刻尔克至巴塞罗那子午线弧只是其中的一段，该测量结果被用于推算完整子午线象限的长度。

国际标准“米”的原型（1875 年） 1875 年，国际计量大会（CGPM）重新定义了米，采用了一根铂铱合金棒（含 90% 铂和 10% 铱）作为参考对象 [214]。这根棒被称为国际米原器（IPM），也称为编号为 27 的棒，存放在国际计量局（BIPM）。

参考对象铂铱合金棒（含 90% 铂和 10% 铱）。其耐久性和低热膨胀率（8.7 微米/米 · 摄氏度）的特性被用来定义测量单位。测量时需在 0 摄氏度下进行。该棒设计为 X 形截面，总长 102 厘米，两根刻线的中点之间为 1 米。30 根复制品被分发给成员国用于校准，其中编号为 6 的棒被送到了美国。尽管这种物理标准对于 19 世纪的计量学来说具有革命性意义，但它存在风险，比如可能受损、容易经受微小磨损以及难以获取等问题。

基于氪-86 的定义（1960） 1960 年，国际计量大会（CGPM）利用氪-86（^{86}Kr）的发射光谱重新定义了米，这标志着首次基于自然常数的定义测量单位 [64]。*参考对象是氪-86*（^{86}Kr），*其发射光谱被用于定义计量单位*。新的定义规定：

> 1 m = 橙红色谱线波长的 1 650 763.73 倍
> 该谱线由^{86}Kr 同位素在能级$2p$ 与$5d$ 之间跃迁时发射，
> 其中$2p$ 和$5d$ 表示特定的原子能级。

该定义的精度达到了 ±4 亿分之一（ppb），其中 1 ppb 表示十亿分之一，即相对不确定度为 10^{-9}，超越了铂铱合金米尺的局限性。它采用了一种充有纯 ^{86}Kr 蒸气的放电灯，通过电能激发来发射参考波长（真空中的 605.780210 纳米）。新标准使全球实验室能够独立复现米制（米的定义），而无需比对实物标准。

转向基于光速的定义（1983 年） 1960 年代，基于氪-86 的米制定义展现了国际单位制（SI）系统对科学进展的适应能力。它为 1983 年基于光速重新定义米奠定了基础 [308]。尽管氪-86 的定义已被替代，但它在建立计量学原理方面是一个重要的一步，特别是利用原子现象作为计量的基础。

5.9.2　“米”的最新定义

1983 年，米基于光速来定义，真空中的光速（c）以米 × 每秒为单位表示时，将光速的数值固定为 299 792 458 [308]。参考对象是真空中的光，其速度恒定这一属性被用于定义计量单位。这一定义表明：1 m = $c \times (1/299\ 792\ 458)$ s，其中真空中的光速被定义为 $c = 299\ 792\ 458$m/s（精确值（exact））。

5.9.3　米测量标准的最新实现

目前米的定义是基于真空中的固定的光速确定的，其测量标准的实现则必须能够溯源至原子时间标准（atomic time standards）[72]。在各种实现方法中，国际计量局（BIPM）正式推荐了一种实现米的基准（primary standard）[3]：碘稳频氦氖（iodine-stabilized helium-neon）激光器。

- *原理*（*Principle*）：波长锁定于 $^{127}I_2$ 的 R(127) 跃迁，波长为 632.991 nm

[3] 中文计量学通常将 primary standard 翻译为基准，它与本书定义的比较基准的含义一致。

- 组件（*Components*）：
  - 氦氖激光器 (633 nm)
  - 温度可控的碘电池（±0.01 摄氏度）
  - 法布里-珀罗腔（精细度 > 100）

以下概述了基准的复制要求，包括真空、热控制、振动和可追溯性等方面的条件。

- 真空（*Vacuum*）：对于基准，不超过 1 微巴 [4]
- 热控（*Thermal Control*）：±0.1 摄氏度（实验室），±1 摄氏度（工业）
- 振动（*Vibration*）：均方根值小于 10 纳米每秒平方
- 可溯源（*Traceability*）：所有二级测量设备必须具有可追溯至国家标准的有效校准证书

5.10　求索的基本量

本节将对求索本身视为一个对象，这里的求索可以是测量、测试、推理或评价。尽管求索对象或求索条件各有不同，但它们有相同的结构：那就是在一个求索条件下获得求索对象的求索结果。本节定义求索的基本量。

本文记第 i 次求索结果为 $\hat{\theta}_i$，求索结果的真值为 θ。

5.10.1　反映求索结果之间关系的量

求索基本量定义 1 (一致性(*consistency*))**.** 设 n 次重复求索结果为 $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$，则一致性可由样本标准差 (*sample standard deviation*) s 或者变异系数 (*coefficient of variation*) CV 量化，s 或者 CV 越小，一致性越高 *[28, 3]*。

- 样本方差（*sample variance*）：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}i - \bar{\hat{\theta}})^2 \tag{5.1}$$

  其中 $\bar{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$ 为样本均值。

- 样本标准差（*sample standard deviation*）：

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}i - \bar{\hat{\theta}})^2} \tag{5.2}$$

  其中 $\bar{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$ 为样本均值。

[4] 1 微巴（bar）就是百万分之一巴，是一个衡量压强的单位，特别适合描述微小的压力变化。1 微巴 = 0.1 帕斯卡 (Pa)

- 变异系数（*coefficient of variation*）（$\bar{\hat{\theta}} \neq 0$）：

$$\mathrm{CV} = \frac{s}{|\bar{\hat{\theta}}|} \tag{5.3}$$

求索基本量定义 2（精确性（precision）或可重复性（repeatability））. 精确性或可重复性指在相同求索条件下重复求索时，求索结果之间的一致性，反映随机误差的大小。一致性越高，则精准性或者可重复性越高 *[28]*。

精准性或者可重复性是求索的显式量。

求索基本量定义 3（可再现性（reproducibility））. 可再现性指在不同条件下（不同主体、不同工具、不同地点），求索结果之间的一致性 *[28]*。可再现性同样以样本标准差量化，但其数据来自不同求索条件下的求索。

求索基本量定义 4（可靠性（reliability））. 可靠性（*reliability*）关注求索过程在长期运行或重复执行中保持结果一致性的能力。设同一求索对象在两次独立求索中分别产生结果序列 $\{x_i\}$ 和 $\{y_i\}$ $(i = 1, \ldots, n)$，则可靠性可由两次求索结果的相关性来量化 *[225, 192]*：

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.4}$$

r_{xy} 的取值范围为 $[-1, 1]$，越接近 *1*，表示两次求索结果的一致性越高，可靠性越好 *[225, 192]*。

求索基本量定义 5（敏感性（sensitivity））. 敏感性刻画求索对象的求索结果对求索条件变化的响应程度。若同一求索对象在不同求索条件下的结果变化较大，则求索结果对该求索条件更敏感；若变化较小，则结果更稳定或更稳健。在评价学场景中，一个常见用法是评价结果对评价条件（*EC*）变化的敏感性 *[265]*。

求索基本量定义 6（健壮性（robustness））. 健壮性度量保持求索结果稳定的能力，可由崩溃点（*breakdown point*）量化：崩溃点是使得统计量严重失真、趋于极端失效的最小污染数据占比。例如，样本均值的崩溃点为 $1/n$（仅需一个极端值即可使其任意偏移），而样本中位数的崩溃点为 $\lfloor (n+1)/2 \rfloor / n \approx 50\%$（需要超过半数的数据失效），因此中位数比均值更具健壮性 *[169, 244]*。

5.10.2　相对真值的量

求索基本量定义 7（偏差（bias））. 偏差指求索结果的期望值与真值之间的系统性偏离。偏差是一种系统误差 *[174, 46]*。

$$\mathrm{bias} = E[\hat{\theta}] - \theta \tag{5.5}$$

若偏差为零，则称该求索结果为" 无偏的"。偏差也是求索的隐藏量。

求索基本量定义 8（误差（error））. 误差指单次求索结果与真值之间的差异 *[28]*。

$$\text{error} = \hat{\theta} - \theta \tag{5.6}$$

误差包含偏差（系统误差）和随机误差两部分。误差也是求索的隐藏量。

求索基本量定义 9 (准确性或准确度（accuracy）)**.** *准确度或准确性*[5]*指单次求索结果与真值的接近程度，反映系统误差的大小 [28]。*

准确度定义如下：

- **绝对误差（absolute error）**：
$$\varepsilon_{\text{abs}} = |\hat{\theta} - \theta|$$
- **相对误差（relative error）**（$\theta \neq 0$）：
$$\varepsilon_{\text{rel}} = \frac{|\hat{\theta} - \theta|}{|\theta|}$$

绝对误差直接刻画偏离程度，相对误差则消除了量纲影响，便于不同量级的比较。基于相对误差的准确度表示为 $(1 - \varepsilon_{\text{rel}})$ [28]。准确度也是求索的隐藏量。

一个求索过程可能准确但不精确，也可能精确但不准确。理想的求索应同时具备高准确度与高精确性。

求索基本量定义 10 (置信度（confidence level）)**.** *置信度表示多次重复求索时，每次在获得求索结果的样本上构建的置信区间包含求索结果总体参数真值的概率 [206, 46]。*

设求索结果的标准误差（standard error）为 $\text{SE}(\hat{\theta})$。

$$\text{SE}(\bar{\hat{\theta}}) = \frac{s}{\sqrt{n}} \tag{5.7}$$

其中 s *为样本的标准差。当* $\hat{\theta}$ *的抽样分布服从正态分布或可由正态分布近似时，在置信水平* $1 - \alpha$ *下，双侧置信区间为：*

$$\bar{\hat{\theta}} \pm z_{\alpha/2} \cdot \text{SE}(\hat{\theta}) \tag{5.8}$$

其中 $z_{\alpha/2}$ *为标准正态分布的上* $\alpha/2$ *分位数。最常见的置信度为 95%（*$z_{0.025} \approx 1.96$*）和 99%（*$z_{0.005} \approx 2.576$*） [46]。*

需要注意的是，置信度描述的是方法的长期表现，而非单次计算得到的置信区间包含总体参数真值的概率。即：95% 置信度意味着若重复 100 次求索，每次获得求索结果的样本，并构建相应的置信区间，大约有 95 个区间会包含总体参数真值。

[5]在不同上下文里，使用准确度或准确性。

5.10.3　求索质量的指标

求索基本量定义 11（有效程度（effectiveness））. 有效程度指求索达成预定目标的程度 *[136]*。设求索目标所要求的质量阈值为 Q^*（例如准确度、覆盖度等），实际达成的质量为 Q，则有效程度可定义为：

$$\eta_{\text{eff}} = \frac{Q}{Q^*} \tag{5.9}$$

$\eta_{\text{eff}} \geq 1$ 表示达到或超过目标。

求索基本量定义 12（效率（efficiency））. 效率指达成单位质量所需的资源消耗，后者称为开销（*overhead*）*[136]*。设求索过程消耗的总资源为 C（时间、计算量、人力等），则效率可定义为：

$$\eta_{\text{effi}} = \frac{Q}{C} \tag{5.10}$$

效率越高表示单位质量消耗的资源越少。

在不同的求索里，尽管有效程度与开销的内涵不同，但它们构成了一对核心权衡：这一权衡贯穿所有求索类型：

- 测量中，更高的准确度（有效程度）要求更精密的仪器和更严格的校准链（开销）[28, 2]；
- 测试中，更高的覆盖度（有效程度）要求更多的测试用例，带来更大的时间和计算成本（开销）[327, 135]；
- 推理中，更严格的证明（有效程度）要求更多的推导步骤和前提验证，带来更高的认知和计算负荷（开销）[282]。

求索基本量定义 13（有效性（validity））. 有效性度量刻画求索结论的准确度和适用范围 *[236]*，分为两个层面：

- 内部有效性（*internal validity*）：指能够更高程度地排除当前求索条件下混杂因素的干扰，使得求索结论具有更高的准确度。
- 外部有效性（*external validity*）或者泛化能力（*generalizability*）：指求索结论能推广到更通用的求索条件的能力。一个外部有效性高的求索结论不依赖于特定的求索条件配置。

有效性虽然是定性维度，但可通过威胁因素的数量和严重程度加以间接量化。常见的内部有效性威胁包括 *[236]*：混杂变量未控制（*uncontrolled confounding variables*）、选择偏差（*selection bias*）、成熟效应（*maturation effect*）、测量工具变化（*instrumentation change*）等。

设待评估的求索过程面临 m 类内部有效性威胁因素，其中已被有效控制的有 m_c 类，则内部有效性的控制比例为：

$$V_{\text{int}} = \frac{m_c}{m} \tag{5.11}$$

V_{int} 越接近 *1*，表明对内部有效性威胁的控制越充分。

常见的外部有效性威胁包括：样本不具代表性（*unrepresentative sample*）、求索条件过于特殊（*overly specific interrogation conditions*）、时间效应（*time effects*）等 *[236]*。

类似地，外部有效性可通过求索条件的覆盖范围与目标推广范围的比值来刻画。

求索基本量定义 14 (简单性 (simplicity))**.** 简单性量化对象建模假设的约束条件的多少——越少越简化——或者对象模型相对于对象的简化程度。其核心是通过尽量减少对象建模假设的约束条件以及对象的结构、属性 和机制的数目来降低模型的复杂性。简单性可通过模型保留的建模假设的约束条件，结构、属性和机制相对原始的比例来量化。设对象包含的组成总数为 N，简化的模型保留的组成总数为 M，则简单性可定义为：

$$S = 1 - \frac{M}{N} \tag{5.12}$$

S 越接近 *1*，表明复杂性越低、简单性越高；当 $S = 0$ 时，表示未做任何简化。

这里的组成总数可以是单一的结构、属性和机制，或者归一化的结构、属性和机制总和。

5.11　测量的指标

测量基本量定义 1 (测量不确定度 (Uncertainty))**.** 测量不确定度是对测量可信性或者可信程度的度量。由于真值和误差通常不能直接获得，不确定度成为实际计量中度量测量结果质量的量 *[28, 2]*。

根据测量不确定度表示指南（GUM），不确定度分量的评定方法分为两类 [2]：

- **A 类评定**：通过对多次求索结果的统计分析获得。进行多次独立重复观察并以观察结果的算术均值作为测量的估计值，在这种情况下，以该均值的标准误差作为采用 A 类方法评定得到的标准不确定度。
- **B 类评定**：通过非 A 类评定的统计方法获得，例如查阅校准报告、数据手册或基于经验估计。

求索结果的合成标准不确定度由 A 类和 B 类方法评定得到的标准不确定度通过平方和根法（Root-Sum-Square）合成。当各标准不确定度分量已经换算为其对输出量 y 的贡献 $u_i(y)$，且这些分量彼此不相关时，合成标准不确定度为：

$$u_c(y) = \sqrt{\sum_i u_i^2(y)} \tag{5.13}$$

其中 $u_i(y)$ 为求索结果 y 的第 i 个不确定度分量。

GUM 不确定度评定的核心思想是：首先建立被测量与各输入量之间的 测量模型；然后分别用 A 类或 B 类方法评定各输入量的不确定度；最后根据测量模型，计算这些输入不确定度对最终测量结果的影响，从而得到输出量的合成标准不确定度 [2]。GUM 评定核心步骤 [2]：

1. 明确被测量与建立测量模型　定义被测量 Y，写出其与输入量 $X_1, X_2, \ldots, X_n$ 的函数关系：

$$Y = f(X_1, X_2, \ldots, X_n) \tag{5.14}$$

函数 f 应包含所有影响测量不确定度的输入量。

2. 识别所有不确定度来源　识别不确定度来源，常见来源包括测量重复性（随机效应），仪器校准、分辨率或漂移，标准物质或参考值不确定度，环境因素（如温度、湿度、气压），以及方法近似、人员读数、样品前处理等。

3. 评定各分量的标准不确定度 $u(x_i)$　评定分为 A 类评定与 B 类评定，均以标准不确定度为表征。

A 类评定通过对评定分量的一系列观察值进行统计分析获得。在最常见的简单情形中，设对输入量 X_i 进行 n 次相互独立、测量条件相同的重复观察，得到观察值 $x_{i,1}, \ldots, x_{i,n}$，其算术平均值为：

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{i,k}. \tag{5.15}$$

取输入估计值 $x_i = \bar{x}_i$。这些观察值的实验标准偏差为：

$$s_i = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_{i,k} - \bar{x}_i)^2}. \tag{5.16}$$

则输入估计值 x_i 的 A 类标准不确定度，即均值的实验标准偏差，为 [2]：

$$u(x_i) = \frac{s_i}{\sqrt{n}}. \tag{5.17}$$

B 类评定根据校准证书、制造商说明书、参考数据、以往测量数据或经验等现有信息，为输入量建立适当的概率分布，并以该分布的标准偏差作为标准不确定度。当输入量 X_i 被认为位于以其估计值 x_i 为中心、半宽为 a 的对称区间内时，可根据所假设的概率分布将区间半宽换算为标准不确定度 [2]：

$$u(x_i) = \frac{a}{d} \tag{5.18}$$

其中 d 为由假设分布决定的换算除数。例如，对于均匀分布（矩形分布），$d = \sqrt{3}$；对于正态分布（置信度为 95%），d 近似取为 2。

若校准证书给出扩展不确定度 U 及其包含因子 k，则相应的标准不确定度为 [2]：

$$u(x_i) = \frac{U}{k} \tag{5.19}$$

4. 计算合成标准不确定度 $u_c(y)$　基于不确定度传播规律（泰勒展开，一阶近似），当输入量独立时，合成标准不确定度的计算公式为：

$$u_c(y) = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i)} \tag{5.20}$$

其中 $c_i = \frac{\partial f}{\partial x_i}$ 为灵敏系数，反映输入量对输出量的影响权重。若令 $u_i(y) = c_i u(x_i)$ 表示第 i 个输入量对输出量 y 的不确定度贡献分量，则上述公式可简写为 [2]：

$$u_c(y) = \sqrt{\sum_{i=1}^{n} u_i^2(y)}. \tag{5.21}$$

5. 计算扩展不确定度 U　计算扩展不确定度的目的是给出可能包含被测量真值的数值范围。计算公式为

$$U = k \cdot u_c(y) \tag{5.22}$$

其中 k 为包含因子。在常规测量中，若有效自由度足够大且合成不确定度接近正态分布，通常取 $k = 2$；若在小样本或需要特定置信水平的严格场合下，应通过计算有效自由度查 t 分布表确定 k 值 [2]。

测量结果可按下列形式报告：

$$Y = (y \pm U) \tag{5.23}$$

其中，y 为被测量 Y 的估计值，U 为扩展不确定度。报告时应注明所采用的包含因子 k 以及说明 k 的确定方法 [2]。

5.12　总结

本章阐述了我们对计量学的新的理解，也就是计量学是量的定义、实现与赋值的科学。同时，在阐述基本量的定义和实现的同时，我们也定义了求索的基本量和测量的基本量。

第六章

测试学：检验与测试的科学

本章阐述了我们对测试的理解，包括测试的基本概念、测试学的定义、问题定义、基础假说、基本作用、方法论以及典型的测试案例。

我贡献了第6.1、6.2、6.3、6.4和6.5节；我和王磊贡献了第6.7和第6.8节；王磊贡献了第6.6、6.9、6.10和6.11节。

6.1 什么是测试学？

测试学概念定义 1 (测试对象)**.** 测试对象是被测试的对象或者衍生对象，它可以是对象及其命题或模型。

测试学概念定义 2 (被测量 (testrand))**.** 测试对象待测试的行为量称为被测量 (*testrand*)。

测试学概念定义 3 (检验基准 (test oracle))**.** 检验基准是关于测试对象的事实 *(fact)*、基准事实 (*ground truth*) 或已经检验通过的命题或者模型。如果测试对象是一个人造物或者自然界中的对象，检验基准也可以是测试对象的预期行为。

检验基准本质上是用于参考的知识，我们在第 5.2节中定义了参考，从而使测试人员能够区分可接受和不可接受的结果。

根据知识来源猜想，检验基准有两个来源，一个是智能生命的感官经验，例如预期的行为；另一个来源是智能生命的心智能力，例如检验通过的理论。

检验基准所定义的预期结果可以是确定的 (确定值) 或统计性的 (统计值)。此处 “确定的” 含义和第 3.5.2节定义的 “确定的” 含义相同，但统计性不同于真随机性的含义，而是一种表面随机性。

测试学概念定义 4 (测试输入 (test input))**.** 测试输入 指的是提供给测试对象的数据 *[135]* 或激励 (*stimuli*) *[200]*，用于驱动测试。这些可能包括参数、用户输入或配置设置。

例如，在一个计算机实现的登录系统中，输入可能是用户名和密码。

测试学概念定义 5 (测试前提条件（test precondition）)**.** 测试前提条件是为了确保测试对象在执行测试前处于有效状态而设定的条件 *[135]*。前提条件可能涉及系统配置或需要满足的外部条件。

例如，对于同一个计算机实现的登录系统，一个前提条件可能是用户账户必须已经存在于数据库中，才能测试登录功能。

测试学概念定义 6 (测试执行过程（test execution procedures）)**.** 测试执行过程定义了执行测试所采取的步骤或操作 *[135]*。

例如，对于登录系统的登录功能测试，物理的执行过程可能包括打开登录页面、输入用户名和密码，并点击“登录”按钮。

测试学概念定义 7 (测试用例)**.** 测试用例是依据检验基准为测试预定义的求索条件 *(interrogation condition)*，包括测试输入、测试前提条件和测试执行过程；一个检验基准至少对应一个测试用例。测试用例包括设计（模型）和实现，分别能虚拟执行和物理执行测试对象，以验证实际结果是否与检验基准定义的预期结果一致。

测试学概念定义 8 (执行检验，检验)**.** 执行检验 是一个原语（*primitive*）的比较过程，它运行测试用例，以模型或者物理的方式执行测试对象，并且将运行结果与检验基准定义的预期结果进行比较。检验包括定义检验基准、设计或者实现测试用例以及以虚拟执行或物理执行检验。

不同领域的测试有不同的含义，例如人造物或者自然界对象行为的测试（testing），二分类系统（binary classification system）的分类，如诊断测试，假设检验 (hypothesis testing)、理论验证（Verification of theory）。但都是基于检验针对不同测试对象的测试。

基于此，我创造了一个新的术语“测试学（Testology）”来涵盖这一领域，并提出了跨不同领域的通用测试原理和方法。

测试学概念定义 9. 测试学 是检验与测试的科学：检验是唯一能验证命题和模型有效性的方法；测试基于检验，定义、实现和赋值测试对象的行为量（被测量）。实质上它是求索科学体系下的计量学在测试领域的应用。

6.2　概念定义

由于测试学是计量学在测试领域的应用，第 五章定义的测量学概念稍作修改可以用于测试学概念的定义。

测试学概念定义 10 (测试标准（test standard）)**.** 测试标准是测试对象的行为量的实现，它是一个参考，不同的实现存在准确度和开销的差异。反过来对象行为量的定义可以视为测试标准的模型，也就是测试标准的简化表示。

测试学概念定义 11 (测试仪器（test instrument）)**.** 测试仪器是参考测试标准实现的装置，用于测试对象行为量的赋值。

测试学概念定义 12 (可测试性（testable or testability）)**.** 是指一个测试对象可以被执行，其结果可以与检验基准定义的预期结果进行比较。

测试学概念定义 13 (执行测试)**.** 执行测试是基于检验对对象的行为量赋值。

测试学概念定义 14 (测试系统)**.** 测试系统是参与测试过程的对象与衍生对象组成的系统，包括对象的行为量的定义，测试标准，测试仪器，测试主体，测试方法学，测试过程等。

测试学概念定义 15 (对象失效)**.** 对象失效是指对象行为偏离预期。故障是失效的原因。

6.3 问题定义

测试问题可以表述如下：

问题定义 2 (测试问题定义)**.** 给定一个测试对象，如何设计检验基准和测试用例，在此基础上定义对象的行为量；实现测试用例，在此基础上实现对象的行为量。最后通过执行测试为对象的行为量赋值。输出需要满足利益相关方对于测试准确度和开销的要求。

6.4 基础假说

作为计量学的应用，测试学的基础假说和前者相同。
第一条量的真值假说意味对象行为量具有真值。
第二条可利用属性假说指的是可利用属性是检验基准。
第三条可获得性假说指测试标准可获得。

6.5 测试的基本作用

测试可以为对象的行为量赋值。因此，它和测量一起成为最基础的求索方法，只有它们能够为对象的量赋值，是其他求索的基础。在测量和测试的基础上，可以形成命题或者模型，也可以推断对象的影响。

测试也是唯一能够验证命题或模型有效性的求索手段。它能用于证实和证伪对象或者对象关系的命题或者模型。证伪只需要一次检验，即反例。而证实则需要完备性。

从这个角度来看，测试充当了推理的支柱，我们将在第七章讨论推理，因为推理会产生不同的命题或模型，而这些只能通过测试来证伪或者证实。

6.6 测试的指标

测试基本量定义 1 (测试正确性（correctness）)**.** 测试正确性指测试用例（虚拟或者物理的）执行结果符合检验基准定义的期望结果的程度。

一个非 1 即 0 的正确性测试例子如下：设输入为 V_i，检验基准定义的期望结果为 $R_{\text{expected}}(V_i)$，测试对象的实际输出为 $R_{\text{actual}}(V_i)$，则单个测试用例判为：

$$\text{Pass}(V_i) = \begin{cases} 1, & \text{若} R_{\text{actual}}(V_i) = R_{\text{expected}}(V_i) \\ 0, & \text{否则} \end{cases} \tag{6.1}$$

测试基本量定义 2 (测试覆盖度（coverage）).

设测试对象的状态空间为 S_{total}，$|S_{total}|$ 表示对应状态空间中元素的数量；已被测试覆盖的子空间为 $S_{covered}$，$|S_{covered}|$ 表示对应状态空间中元素的数量。则覆盖度为：

$$Coverage = \frac{|S_{covered}|}{|S_{total}|} \tag{6.2}$$

根据量的真值假说，在理想情况下可以实现 Coverage = 1。在实践中，覆盖度受到资源约束，需要在覆盖度与开销之间权衡。

测试基本量定义 3 (通过率（pass rate）).

通过率是虚拟或者物理执行测试用例通过的比例，是测试结果最直接的量化指标。设共执行 N 个测试用例，其中通过 N_{pass} 个，则：

$$Pass\ Rate = \frac{N_{pass}}{N} \tag{6.3}$$

测试基本量定义 4 (显著性水平 α).

显著性水平 α 是假设检验中拒绝真实零假设 H_0（发生第一类错误）的容忍概率，并作为判断检验结果是否具有统计显著性的决策阈值 *[208]*。即：

$$\alpha = P(\text{拒绝} H_0 \mid H_0 \text{ 为真}) \tag{6.4}$$

在实际检验过程中，我们使用两个检验统计量 z 标准分数（*z-score*）或者 t 标准分数（*t-score*）来度量一个数据点 (data point) 或者样本均值" 偏离总体平均数有多少个标准误差"。这里 "偏离总体平均数有多少个标准误差" 本质上是一个统一的量（single universal scale）。

当总体标准差已知时，可以采用 z 标准分数检验，即在零假设成立时，检验统计量服从标准正态分布。当总体标准差未知并使用样本标准差进行估计时，可以采用 t 标准分数检验，即在零假设成立时，检验统计量服从相应自由度的 t 分布。

通常在实际系统中，测试的总体标准差是未知的，因此普遍使用 t 标准分数检验。计算公式如下：

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \tag{6.5}$$

其中，$\bar{X}$ 为样本均值，μ_0 为零假设下指定的总体均值（hypothesized population mean），例如某一固定参考值或基线配置（baseline configuration）的已知平均性能，s 为样本标准差，n 为测量样本量。

将实际测试样本数据代入公式计算所得的检验统计量具体数值记为 t_{obs}。它反映了样本均值偏离假设的总体均值有多少个标准误差，在此基础上进一步计算在零假设 H_0 成立的前提下，假如仅由随机误差导致变化，观察到当前或更极端的检验统计量的概率，也就是p 值（p-value）。

在统计学假设检验中，p 值（p-value）是指在原假设 H_0 完全成立的前提下，获得与实际观测到的结果至少同样极端的测试结果的概率 [85]。

其数学表达式为：

$$p\text{-value} = P(|T| \geq |t_{\text{obs}}| \mid H_0) \tag{6.6}$$

T 是在零假设 H_0 成立时服从自由度 $\nu = n - 1$ 的 t 分布的随机变量，t_{obs} 为根据样本数据计算出的检验统计量。

若实验设计为单侧检验，则公式可替换如下：

- 右侧检验

$$p\text{-value} = P(T \geq t_{\text{obs}} \mid H_0) \tag{6.7}$$

- 左侧检验

$$p\text{-value} = P(T \leq t_{\text{obs}} \mid H_0) \tag{6.8}$$

决策规则为：若 p-value $\leq \alpha$，则拒绝 H_0。

测试基本量定义 5 (灵敏度与特异度（sensitivity and specificity）)**.**
在二分类判别系统中（如诊断测试或分类器），当检验基准能够区分正例与负例时，可以根据测试结果与真值的对照构造判别矩阵：真阳性 TP（真值为正且判别为正）、假阴性 FN（真值为正却判别为负）、真阴性 TN（真值为负且判别为负）、假阳性 FP（真值为负却判别为正）。

在此基础上，灵敏度，又称真阳性率，衡量测试将真值正例判别为正的能力，特异度，又称真阴性率，衡量测试将真值负例判别为负的能力 *[7]*：

$$\mathit{Sensitivity} = \frac{TP}{TP + FN}, \qquad \mathit{Specificity} = \frac{TN}{TN + FP} \tag{6.9}$$

灵敏度反映测试“不漏判”的能力（降低假阴性），特异度反映测试“不误判”的能力（降低假阳性）。二者通常存在权衡：调整判别阈值会使一方上升而另一方下降，因此实践中需要结合测试目的在两者之间权衡。

测试基本量定义 6 (测试可靠性（reliability）)**.**
测试可靠性 $R(t)$ 指系统在规定时间 t 内无故障的概率 *[61]*。当系统故障服从指数分布时，可靠性可进一步表示为：

$$R(t) = P(T > t) = \exp\left(-\frac{t}{\text{MTBF}}\right), \tag{6.10}$$

其中，t 为设定的时间，T 为系统从开始运行到发生故障的时间，MTBF 为平均无故障工作时间。$R(t)$ 的取值范围为 $[0, 1]$，其值越大，表示系统在规定时间内无故障运行的可能性越高。

测试基本量定义 7 (可用性（availability）)**.** 系统可用性度量系统无故障状态的概率 *[61]*，可表示为：

$$A = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}, \tag{6.11}$$

其中，MTBF 为平均无故障时间，MTTR 为平均修复时间。A 的取值范围为 $[0, 1]$，其值越大，表示系统在任意时刻处于无故障状态的可能性越高。

在人工智能特别是大语言模型推理测试中，主流指标通常不是直接测试“推理过程本身”，而是测试任务输出结果。例如，多选题、问答题和数学题常用准确度（accuracy）或精确匹配率（exact match）。

测试基本量定义 8 (任务结果指标（task outcome metrics）)**.** 任务结果指标指在人工智能和自动化推理测试中，通过标准任务、测试集或基准来度量推理系统输出是否正确的指标 *[116, 51]*。

准确度（*accuracy*）或精确匹配率（*exact match*） *[234]*:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\hat{a}_i = a_i\}. \tag{6.12}$$

其中，$\hat{a}_i$ 是系统输出的答案，a_i 是标准答案。

通过率或 *pass@k* *[51]*:

$$pass@k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\text{第}i\text{ 个任务的}k\text{ 次候选输出中至少一次通过}\}. \tag{6.13}$$

通过率或 $pass@k$ 这类指标广泛用于 MMLU、数学推理、代码生成等基准测试 [116]。它们能衡量推理系统在任务上的结果表现，但不一定能充分揭示推理过程是否真实、稳健或可解释。

6.7　基本原理

如图 6.1所示 [1]，测试的核心是对象行为量（被测量）的定义、实现和赋值。行为量的定义需要基于检验基准和测试用例的设计（模型）。以验证本书的正确性为例，我需要定义一组检验基准以及相应的测试用例设计，只有基于它们才能定义书的正确率。

同样，行为量的实现需要基于测试用例的实现。行为量的实现就是测试标准，它是可以物理执行的。同样，以本书的正确性验证为例，同一测试用例的设计可以有不同的实现，测试标准的实现需要基于测试用例的实现。测试仪器是参考测试标准实现的测试装置。

测试是基于执行检验的聚合过程 (aggregate process)，它基于执行检验，为行为量赋值。也就是说，执行检验是执行单个测试用例的过程，测试是多个检验过程的聚合。以测试通过率指标为例，如果我们检验 100 个测试用例的测试正确性，通过了 90 个，则测试通过率为 90%。

与计量学的概念一致，可以通过评价学实现测试结果相对于测试单位（行为量的测量单位）的可溯源性。

测试学概念定义 16 (测试的可溯源性（test traceability）)**.** 测试的可溯源性是测试利益相关方为了保证测试的准确度对测试过程的一种要求，它要求测试结果与测试单位的定义以及测试标准通过有据可查、不间断的校准链建立关系，具体体现为测试仪器或者测试标准中每一级都应使用更高一级（具有更高准确度）进行校准，以帮助识别和减小系统偏差。

检验基准可以分为两个主要类别。第一类验证测试对象的实际结果是否符合检验基准所规定的预期结果。例如，排序算法应该对任何有效的输入序列返回一个非递减顺序

[1] 在我绘制的草图基础上，王磊绘制了该图。

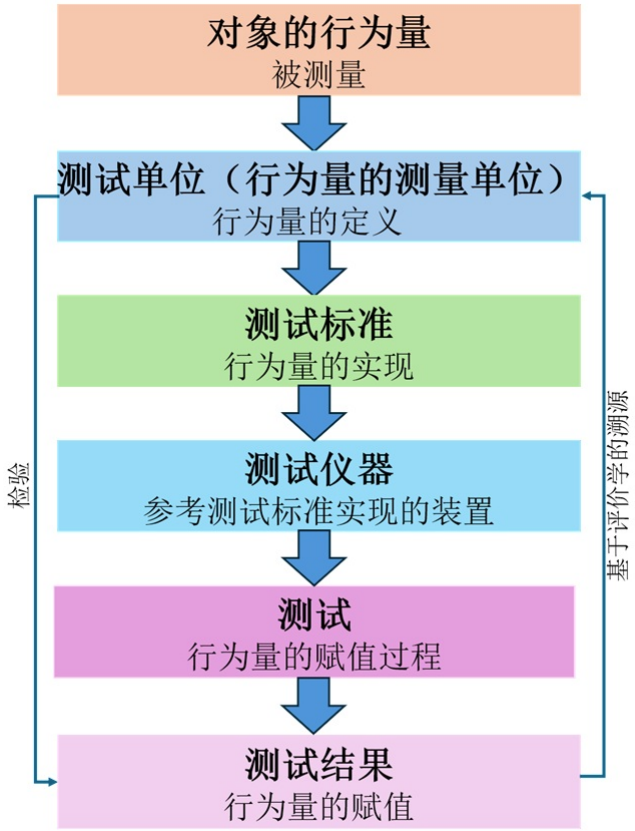

图 6.1: 简化而系统的测试学概念框架。因为空间限制，没有添加推理学，它可以为测试学提供更严谨的推理体系。

的输出结果。第二类则侧重于测试对象运行的不同环境，目标是验证测试对象在不同的环境中是否能表现出由检验基准规定的预期结果。这类测试的目的是确保人造物在其设计的目标环境下正常工作，能够与外部对象正确、适宜地交互，遵守环境约束，并在现实操作条件下保持健壮性。

我们进一步阐述两种典型测试方法的原理：基于概率的测试（*Probability-based Testing*）和确定性测试（*Deterministic Testing*），它们在检验基准上有所不同。

6.7.1　基于概率的测试方法原理

基于概率的测试依赖于一个广义的“基准事实”：在某个假设（hypothesis）下，如果仅仅是由随机误差导致的变化，当前求索结果样本的平均值偏离总体平均值的程度出现的概率极低，我们宁愿拒绝这个假设。从不同的角度看，求索到的结果可能是观察到的结果、实验结果，或者测量、测试结果。从这个意义上来说，我使用了“求索结果”这个更广义的词，在前文中有准确的定义。

在第3.6节中，我们已经正式定义了假设的含义。基于概率的测试的一个常见例子是假设检验，它用于检验求索结果是否不太可能发生。这里的检验基准是：如果 p 值（即在零假设为真的前提下出现的概率）小于预先设定的显著性水平 α（例如 0.05），则拒绝零假设 H_0。

例如，在评估一种新药的临床试验中，零假设 H_0 可能断言该药物对血压没有影响，这意味着干预组（treatment group）的平均变化为零。将实验数据针对零假设进行假设检验。如果计算出的 p 值小于预先确定的统计显著性水平 α（例如 0.05），则表明在仅由随机误差导致变化的假定下，观察到样本的平均值偏离总体平均值的当前程度出现的概率极低，因此观察到的结果具有统计学显著性，因此应拒绝原假设。表明该药物可能对血压存在影响。

6.7.2　确定性测试方法原理

确定性测试用于检验系统在预定义条件下是否完全符合预期行为，得出明确的测试通过或失败的结论 [327]。

在确定性测试中，检验基准通常是预先定义的一组规范（specification）或要求，比较系统的行为是否符合预期，在此基础上可以形成命题。正确性通过比较系统的行为是否符合预期来确定，从而检验命题的真假。

例如，在计算机领域，两种主要的测试形式是“软件测试”和“硬件测试”。软件和硬件本质上都是人造物，我们在第3.4节中对人造物进行了定义。这些测试方法可以扩展到其他人造物或者自然界的对象。

- 软件测试（*Software testing*）使用检验基准（*test oracles*），如规范、参考实现或先前的系统版本，来定义预期的行为 [57]。

  软件测试检验不通过时，会出现对象失效（failure）。对象失效可以分为功能性失效：实际结果偏离预期结果；或者当违反了诸如内存、性能或兼容性等约束时 [20]，会出现环境性失效。这些失效可能是级联的 (cascading)：例如，内存不足（环境性失效）可能导致应用程序崩溃，从而引发功能性失效。

  一个具体的例子是测试一个 Web 应用程序的登录：检验基准规定有效的凭证应返回成功的登录信息。如果应用程序崩溃，或具备有效的凭证却返回错误，这就是功

能性失效。如果服务器内存不足，导致应用程序在重负载下变慢，或无法响应，这就是环境性失效，可能间接触发功能性失效。

- 测试向量是硬件测试中广泛使用的检验基准，用于 CPU 和数字电路的测试 [200]。一个测试向量（*test vector*）由输入序列、预期结果和故障模型组成，它模拟真实世界的操作，以检测设计或制造缺陷。

  例如，考虑 64 位 CPU 的 64 位无符号加法器。一个测试向量输入 `A=0x0000000000000001` 和 B=0x0000000000000002。预期结果是 S=0x0000000000000003，进位 =0。如果 CPU 产生错误和/或进位，表明发生了功能失效。

6.8　基本方法学

测试的基本方法学遵循行为量的定义、实现和赋值的过程。测试过程可以概括为一个统一的高层次框架，包含以下顺序步骤：

1. 行为量的定义：定义检验基准，规定测试对象的预期行为结果，依据检验基准，设计测试用例，在此基础上定义被测对象的行为量。
2. 行为量的实现：依据检验基准，实现测试用例，在测试用例实现的基础上实现被测对象的行为量，也就是测试标准。参考测试标准可以实现测试仪器。
3. 行为量的赋值：基于测试标准或测试仪器的测试结果，为测试对象的行为量赋值，实际上就是测试的过程。

在接下来的章节中，我们将介绍测试学在三个领域的应用：物理学理论的验证、统计学中的假设检验以及人造物的测试（包括计算机中的软件和硬件测试）。物理学理论的验证、统计学中的假设检验这两个案例属于证伪，只需要一个反例即可，因此，我们着重阐述检验过程。而后面的案例（软件和硬件测试），测试结果的解释取决于测试覆盖度这个行为量的定义和实现，我们详细阐述量的定义、实现和赋值的三个过程。

测试学还有其他应用，我们考虑未来就这一主题撰写一本书。

6.9　假设检验

假设检验（hypothesis testing）是一种统计过程，它利用样本数据来检验零假设（H_0）与备择假设（H_1） [207]。它采用检验统计量来衡量基于样本数据的观察结果与 H_0 预期结果之间的差异，并通过决策规则–比较 p 值与统计显著性水平 α（significance level）–来决定是拒绝还是不拒绝零假设 H_0。

在假设检验中，检验基准的定义由零假设（H_0）、备择假设（H_1）和显著性水平（α）组成，这三者共同设定了检验对象行为是否显著偏离预期的标准。

测试用例是以样本数据作为测试用例输入，使用样本数据根据检验基准计算检验统计量。最后检验是基于计算得到的检验统计量和决策准则（decision criteria），做出最终决策，拒绝 H_0 或不拒绝 H_0，这是一个通过/失败的判断。

在测试学框架下，假设检验可以描述如下。

1. 定义检验基准: 在假设检验中，检验基准由零假设（H_0）、备择假设（H_1）和统计显著性水平（α）组成，这三者共同设定了检验对象行为是否显著偏离预期的标准。

   - 零假设 *(H_0)*: 关于总体参数或分布的命题，即总体与理论分布模型一致，表示无差异。例如，服用新药的患者的平均血压等于总体平均值，即 $\mu = \mu_0$。
   - 备择假设 *(H_1)*: 与零假设相对立的命题，即在总体与零假设下预期存在统计学显著偏差。例如，服用新药的患者的平均血压不同于总体平均值，即 $\mu \neq \mu_0$。

   决策规则依赖于显著性水平（α）和 p 值:

   - 显著性水平 *(α)*: 事先规定的当 H_0 为真时拒绝 H_0 的概率。通常使用 $\alpha = 0.05$，计算公式为公式 6.4。
   - p 值: 假设 H_0 为真时，获得当前观察到的检验统计量或更极端检验统计量的概率。p-value 的计算依赖于检验类型，即双侧检验或单侧检验，具体公式为公式 6.6、公式 6.7和公式 6.8。
   - 决策规则：如果 p-value $\leq \alpha$，则拒绝 H_0。

2. 测试用例设计与实现: 测试用例基于样本数据。检验统计量对样本估计值与零假设下的参数值之间的差异进行了量化。通常表示为:

$$X = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})}, \tag{6.14}$$

   其中，$\hat{\theta}$ 是从样本数据中计算得到的估计量，用于估计总体参数 θ；θ_0 是零假设 (H_0) 下给定的参数值；$\mathrm{SE}(\hat{\theta})$ 表示 $\hat{\theta}$ 的标准误差。

   测试用例设计指定了如何根据检验基准计算检验统计量，包括:

   - 测试输入（样本数据）：用于计算总体参数 θ 的估计值 $\hat{\theta}$;
   - 前提条件: 主要包括关于底层分布的假设，方差是否已知，样本的独立性等。
   - 执行程序:
     (a) 计算检验统计量*:* 根据前提条件，常见的检验统计量包括:
       – t 标准分数检验（当总体方差未知，使用样本方差进行估计时）: 采用公式 6.5。
       – z 标准分数检验（当总体方差已知时）:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{6.15}$$

       其中 $\bar{X}$ 是样本均值，μ_0 是零假设下的总体均值，σ 是总体标准差，n 是样本大小。

     (b) 推导 p 值*:* 在获得检验统计量的观察值 t_{obs}（t 标准分数的观察值）或 z_{obs}（z 标准分数的观察值）后，根据所选检验统计量在 H_0 下的分布，基于公式 6.6、公式 6.7和公式 6.8计算。

测试用例的实现将测试用例转化为可执行的程序，例如，通过查找统计表中的值，或者编写计算机程序实现自动化计算。

3. 执行检验：执行测试用例获得结果：

   (a) 计算观测到的检验统计量 x_{obs}；

   (b) 计算相应的 p 值；

   (c) 获取检验结果：

   - 如果 p-value $\leq \alpha$，拒绝 H_0；
   - 如果 p-value $> \alpha$，不拒绝 H_0。

为了解释这个框架，考虑一个用于评价新药对血压影响的临床试验。零假设和备择假设分别为：$H_0 : \mu = 0$，药物对血压没有影响；$H_1 : \mu \neq 0$，药物影响血压。μ 是治疗组中血压变化的总体均值。

样本数据来自 20 名患者，样本均值和标准差为：$\bar{X} = 0.774$，$s = 2$。由于总体方差未知，我们使用 t 标准分数检验。t 标准分数观察值计算如下：$t_{obs} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{0.774-0}{2/\sqrt{20}} = 1.73$。

双侧检验的 p 值为：p-value $= P(|T| \geq |1.73| \mid H_0) = 0.1$。由于 p-value $= 0.1 > 0.05$，我们不能拒绝零假设。因此，该实验数据没有足够的统计证据表明该药物显著影响血压。

总体而言，假设检验提供了一种基于概率的决策框架，用于检验观察到的样本结果在零假设 (H_0) 成立时是否仍然合理。如果在 (H_0) 成立的前提下，观察到当前样本结果的概率足够小，即 p 值小于显著性水平 α，则拒绝 (H_0)，并认为数据提供了支持备择假设 (H_1) 的统计依据。

6.10　理论验证

我们以宇称不守恒的验证为例来说明如何验证一个理论。

宇称（P）对称性（Parity symmetry）是经典物理学和早期量子理论中的一条基本对称性，描述了物理定律在空间反演下的不变性 [104]。

形式上，宇称变换将系统的空间坐标 (x, y, z) 替换为 $(-x, -y, -z)$，有效地将右手坐标系转换为左手坐标系。一个物理过程如果在宇称变换下保持不变，就被称为 宇称守恒 ，这意味着它的镜像反射版本——用相反手性（opposite-handed）的坐标系表示——在物理上与原过程无法区分。如果一个物理过程在这种变换下保持不变，则称其为宇称守恒，这意味着其镜像反射版本——以相反的坐标系表示——在物理上与原始版本无法区分。

长期以来，许多相互作用，包括电磁相互作用和强相互作用，被认为遵循宇称对称性。然而，1956 年，李政道和杨振宁对实验数据和理论假设进行了深入分析，提出了宇称在弱相互作用中可能不守恒的假设 [173]。他们还提出了具体的实验构想，以检测粒子发射方向中的潜在不对称性。

宇称不守恒意味着一个物理过程的镜像反射版本——即其在相反手性坐标系的实现——在自然界中可能不会发生，从而导致粒子行为中可观察到不对称性。换句话说，在弱相互作用下，自然本身（Nature）能够区分左手坐标系和右手坐标系。

1957 年，吴健雄及其团队通过对钴-60（Co^{60}）核的实验首次直接验证了弱相互作用中的宇称不守恒 [331]。在实验中，Co^{60} 核通过 β 衰变发射电子。团队使用*绝热去磁*在极低温下极化 Co^{60} 核自旋，并采用*单一蒽晶体闪烁探测器*。通过反转极化磁场并观察同一探测器计数率的变化，他们测量了电子角分布的不对称性。这种方法有效消除了由于多个探测器效率差异引起的系统误差。实验结果清晰地显示电子优先沿与核自旋相反的方向发射，这一现象后来通过控制实验得到证实，从而排除了可能的磁性伪影（magnetic artifacts）。这个发现彻底推翻了之前广泛认为的“所有相互作用中宇称是守恒的”的观点，对粒子物理理论和实验方法产生了深远的影响。

为了系统地理解吴实验（Wu's experiment）的科学方法，可以采用*测试学框架*。

定义检验基准阶段，吴健雄的验证实验定义了一个*检验基准*：在 β 衰变过程中，如果宇称是守恒的，那么沿核自旋轴方向的电子发射空间分布应当是对称的。任何在电子发射方向上观测到的偏离 (bias [2]) 都将表明宇称不守恒。当然还存在其他的检验基准。吴实验本质上是一个证伪的过程，因而不需要强调检验基准集合的完备性。

测试用例设计和实现阶段，测试用例设计是一种预先定义的实验设置，包括测试输入、前提条件和执行步骤，具体如下：

- *输入*: 高纯度的 Co^{60} 放射性样品（已知衰变半衰期和强度）。
- *前提条件*:
  1. *核极化装置*: 实验使用了 Rose-Gorter 方法，通过对铈镁硝石晶体的绝热去磁来极化 Co^{60} 核。
  2. *集成 β 粒子探测器*: 一块薄的蒽晶体闪烁探测器被放置在真空腔体内，位于 Co^{60} 源上方，用于检测发射的 β 粒子。
  3. *极化监控系统*: 安装了两个 NaI 伽马射线闪烁计数器——一个位于赤道面，另一个位于近极位置，用于监测样品的极化状态。
  4. *样品准备*: Co^{60} 样品作为薄晶体层生长在优质铈镁硝石单晶的上表面。
  5. *控制实验准备*: 准备了两个对照样本，以消除实验过程中由磁效应引起的潜在伪影。
- *执行步骤*:
  1. *样品安装*: 将 Co^{60} 样品晶体安装在去磁装置的中心。
  2. *极化过程*: 进行绝热去磁，随后启动垂直螺线管以对齐自旋（20 秒过程）。
  3. *数据采集*: 立即开始 β 粒子和伽马射线计数：
     - 使用 10 通道脉冲高度分析仪对 β 脉冲进行分析（每 1 分钟间隔）。
     - 连续监测伽马各向异性，作为极化状态的指示。
  4. *场反转*: 在随后的实验中反转极化场方向，以验证任何观察到的偏离的真实性。
  5. *预期结果*: 实验数据清楚地显示出不对称效应，电子发射明显偏向于反自旋方向。

[2]这里将 bias 翻译为偏离，与统计中的*偏差*相区分。

6. 控制验证：控制实验确认在非极化条件下没有观察到不对称性。

实验的具体操作流程为：在低温条件下对原子核进行磁化以其自旋取向；沿核自旋轴方向对探测器进行校准并排列；并对仪器灵敏度和背景噪声进行严格控制，以确保数据的可靠性。

执行检验阶段，通过测试用例的执行，沿核自旋方向及其相反方向的电子计数率被记录下来。如果宇称是守恒的，这两个方向上的分布应当是对称的。然而，吴健雄的实验观测到电子发射在反自旋方向有明显偏离，从而为弱相互作用中的宇称不守恒提供了直接证据。

测试学中的统一方法学可用于验证其他理论。

6.11　人造物测试

本章阐述人造物的测试，以计算机软件和计算机硬件测试为例。这一类测试通常可以将正确性测试转换为检验命题的真假。

6.11.1　软件测试

软件测试（software testing）旨在验证软件系统行为是否符合预期。测试期间检测到的故障可分为功能故障和环境故障。当系统产生错误结果时，发生功能故障，而环境故障（如内存不足或执行缓慢）可能会间接导致功能故障 [20]。

本节以测试覆盖率这个行为量为例，阐述行为量的定义、实现和赋值。在测试学框架下，软件测试可以描述如下。

1. 行为量的定义：首先，需要定义测试覆盖度这个测试对象的行为量，它不是抽象的，而是定义在检验基准和测试用例设计的基础之上。只有在清晰和完备地定义检验基准和测试用例之后，测试覆盖率这个行为量才有实际意义。在这个案例中，我们以检验基准覆盖的比例作为测试覆盖率的行为量的定义。可以有其他的定义，例如可以将测试覆盖率这个行为量定义为检验基准覆盖的比率、测试用例设计覆盖的比率、测试用例实现覆盖的比率等。

   在本案例中，假如测试对象的测试状态空间可以建模为 100 个检验基准，如果测试用例覆盖了 80 个检验基准，则测试用例预期的测试覆盖度为 0.8。

   检验基准被定义为指定系统或软件组件的预期输出。其目的是给出系统实际行为与预期结果是否一致的检验依据。例如，一个登录子系统测试的检验基准设计为：如果用户输入了合法的用户名和正确的密码，预期结果应为登录成功并将用户重定向至主页。检验基准所对应的测试用例设计是包含特定输入、前置条件和执行步骤的方案，用于将系统的行为与检验基准所描述的预期结果进行比较。一个检验基准可以对应多个测试用例。例如，对应登录子系统的检验基准，测试用例 1 可以是：输入用户名“admin”和正确的密码；先决条件是用户账户“admin”已存在于系统数据库中；执行步骤则包括打开登录页面、输入用户名和密码，并点击“登录”按钮。测试用例 2 可以是：输入用户名“root”和正确的密码；先决条件是用户账户“root”已存在于系统数据库中；执行步骤则包括打开登录页面、输入用户名和密码，并点击“登录”按钮。

2. 行为量的实现：尽管可以基于检验基准和测试用例的设计，明确定义覆盖率。但是，同样的测试用例设计可以有不同的实现，因此，测试覆盖率这个行为量的实现还依赖于测试用例的实现。一个参考或者规范的行为量的实现实际上就是测试标准。

   根据测试用例设计，测试人员通过创建测试工件（test artifacts），如测试脚本、输入文件或执行指令，搭建所需的测试仪器，并确保测试输入、前置条件和执行步骤被正确实例化，从而实现测试标准。例如，在实现登录测试用例 1 时，测试人员编写一个测试脚本，打开登录页面，输入用户名"admin" 和正确的密码，点击" 登录" 按钮，并验证系统是否重定向到主页。该过程确保测试用例具有可执行性，并能够在预期测试环境中验证系统行为是否符合检验基准。一个测试用例的设计可以对应多个测试标准的实现，例如登录系统的测试用例 1 的实现可以使用 Java 程序编写测试脚本也可以使用 Python 程序编写测试脚本。通过执行实现的测试标准可以实现被测对象的行为量。

3. 行为量的赋值：基于测试标准的测试结果，可以对测试覆盖率这个行为量赋值。例如，执行基于 Java 实现的登录测试用例 1 的测试标准时，运行测试脚本可以执行以下步骤：打开登录页面，输入用户名 "admin" 和正确的密码，点击 "登录" 按钮，并观察系统响应。将实际输出结果与检验基准所定义的预期结果（重定向至主页）进行比较，若用户页面成功被重定向，则判定检验通过；若未出现预期行为，则判定检验不通过。

   最后，基于所有测试标准的检验结果获得测试结果，为被测对象的行为量赋值。例如，测试对象的状态空间建模为 100 个检验基准，其中有 80 个检验基准被测试用例设计所覆盖；这 80 个检验基准共对应 100 项测试用例设计，进一步实例化为 200 个可执行的测试标准。若 200 个测试标准全部检验通过，则最终测试覆盖率赋值为 0.8。若 200 个测试标准中有 150 个检验通过，同时仅有 60 个检验基准所对应的全部测试标准均通过检验，则最终测试覆盖率赋值为 0.6。这里，我们采用的测试覆盖率的判定规则如下：若一个检验基准对应多条测试标准，只要存在任意一条对应的测试标准未通过，该检验基准就不能判定为被覆盖。

软件测试方法可以根据不同的维度进行分类。根据对系统内部的了解程度，我们有：

- 黑盒测试。黑盒测试侧重于在不了解系统内部运作的情况下验证其功能。这种测试完全基于系统的输入和输出。例如，在电子商务网站上，可以使用 "等价类划分（equivalence partitioning）" 来测试登录表单的各种输入，比如确保不同格式的电子邮件地址（有效或无效）能触发系统做出恰当的响应。"边界分析（boundary analysis）" 可用于测试密码长度，确保系统正确实施了最小和最大字符限制。

- 白盒测试。白盒测试确保系统的内部逻辑正常运行，并旨在最大限度地提高测试覆盖率。例如，在电子商务网站上，可以使用 "代码覆盖率（code coverage）" 来验证代码库的每个部分（如运费计算）都得到了充分的测试。"变异测试" 或者 "突变测试" 涉及对代码库进行小的、受控的更改（变异），然后运行测试以确保测试套件能够检测到这些错误，最终确认测试的全面性。

这些方法可以通过不同的途径进一步实施：

- 手动测试。手动测试的目标是从用户的角度确保系统按预期运行。例如，以一个电子商务网站为例；测试人员可能会手动模拟结账体验，以验证所有表单字段是否正确填写，以及用户是否能够成功完成购买。

- 自动化测试。自动化测试利用脚本和工具来执行测试用例，无需人工干预，从而实现高效且频繁的回归验证。这种方法确保在系统更新、变更或新功能部署后，先前已验证的功能仍能正常运行。例如，在一个电子商务网站中，可以编写自动化测试程序来模拟各种用户操作，如用户注册、产品搜索、将商品添加到购物车以及完成结账流程，并自动验证每一步的实际结果是否与预期结果相符。自动化测试的关键优势在于能够快速、准确地重复这些关键验证步骤，与手动执行相比，显著提高了测试覆盖率和效率。

6.11.2　硬件测试

测试向量（test vector）方法在 CPU 和数字电路的硬件测试（hardware testing）中广泛使用，它将预定义的输入序列应用于待测电路（或硬件系统），根据检验基准验证其正确性 [200]。

在测试学框架下，硬件测试向量测试可描述如下。

1. 行为量的定义：首先，与软件测试一致，需要在检验基准和测试用例设计的基础之上定义测试覆盖度这个测试对象的行为量。

   硬件测试向量测试中，一个检验基准是一个参考，用于确定给定输入向量 V_i 的正确预期结果。它定义了待测系统的预期行为，为输入向量指定预期结果 $R_{\text{expected}}(V_i)$。一个测试用例由输入、前置条件和执行步骤组成。输入向量 V_i 即为测试输入。测试用例的前置条件包括必要的测试状态，例如给设备通电、初始化寄存器，以及确保电路处于已知的状态。执行步骤则定义了将 V_i 施加到电路并测量输出的具体操作。一个检验基准可以对应多个测试用例。例如对应于向量 V_i 的检验基准，测试用例 1 可以是系统上电后，电路处于空闲状态，施加向量 V_i；测试用例 2 可以是电路正在执行数据运算，施加向量 V_i。

2. 行为量的实现：与软件测试一致，测试覆盖率这个行为量的实现依赖于基于测试用例的实现，也就是测试标准。硬件测试向量测试的测试标准包括准备输入向量和搭建硬件测试仪器。具体操作包括将输入向量加载到测试设备中、初始化被测硬件，以及验证所有前置条件已满足，以确保激励可靠地施加。一个测试用例的设计可以对应多个测试用例的实现，例如测试用例 1 对应的测试标准在搭建硬件测试仪器时，可以使用不同的逻辑分析仪和信号发生器，从而实现多个测试标准。通过执行实现的测试标准可以实现被测对象的行为量。

3. 行为量的赋值：基于测试标准的测试结果，可以对测试覆盖率这个行为量赋值。执行测试标准将输入向量 V_i 施加到测试硬件上，并观察实际输出 $R_{\text{actual}}(V_i)$。然后将实际输出与检验基准定义的预期输出进行比较。如果 $R_{\text{actual}}(V_i) = R_{\text{expected}}(V_i)$，通过该输入向量的检验；否则，检验不通过。最后，与软件测试一致，基于所有测试标准的检验结果获得测试结果，为被测对象的行为量赋值。

用于 CPU 验证的硬件测试方法可以从不同维度进行分类。按照对系统内部微体系结构了解程度的不同，可分为以下几类：

- 黑盒测试（black-box testing）：在 CPU 测试中，黑盒测试侧重于在不了解处理器内部实现细节的情况下，验证其指令集体系结构（instruction set architecture，ISA）层

面的行为是否正确。该方法主要测试处理器是否能够按预期执行指令。其中，“等价类划分（equivalence partitioning）”常用于验证特定指令的行为，如“ADD”操作，确保不同输入范围均能被正确处理。例如，可以通过对溢出边界进行检查，验证两个大数相加时不会因溢出而产生错误结果。黑盒测试方法中，另一种常见技术是“非法操作码的模糊测试（fuzzing illegal opcodes）”，即向 CPU 输入随机或畸形（malformed）的指令，用以发现潜在的安全漏洞或未定义行为。

- 白盒测试（white-box testing）：白盒测试关注对 CPU 内部逻辑的验证，确保处理器的内部结构和流水线能够正确处理所有可能的执行路径。一种常用技术是“路径覆盖（path coverage）”，即保证处理器流水线中的所有可能路径（例如各种异常处理路径）都被测试到。另一种方法是“变异测试（mutation testing）”，通过对处理器逻辑进行小而可控的修改（例如翻转某个比较器的判断条件），来检验现有测试套件是否能够检测这些错误。

这些方法还可以通过不同的技术手段加以实现：

- *基于仿真的测试*（*simulation-based testing*）：这是最常见的方法，通过在仿真环境中运行测试程序，对 CPU 的功能进行动态验证。例如，工程师会编写或自动生成大量测试向量程序，并在仿真器中执行，再将执行结果与黄金参考模型（golden reference model）进行对比。该方法的优势在于能够灵活地模拟复杂场景和较长的执行序列。

- *形式化验证*（*formal verification*）：形式化验证是一种静态验证方法，它不运行测试系统，而是利用数学方法彻底地证明设计正确性的某些方面。例如，可以对微体系结构中的关键属性进行形式化验证，如流水线中的数据一致性，证明其在所有可能的指令序列下都成立。该方法在发现随机测试难以覆盖的隐蔽且深层次的设计错误方面尤为有效。

- *混合方法*（*hybrid methods*）：复杂的 CPU 验证通常采用混合技术。例如，*并发混合验证*（*concurrent hybrid verification*）将仿真的灵活性与形式化验证的彻底性相结合。在运行仿真的同时，形式化工具会分析从当前状态可达的状态空间，并可能自动生成新的测试向量以覆盖尚未探索的路径，从而更加系统地覆盖和触发各种边界情况。

6.12　总结

本章提出测试学（Testology）并将其定义为检验与测试的科学。测试学作为计量学在测试领域的应用，基于检验对测试对象行为的量进行定义、实现和赋值。测试包括不同领域中的人造物或自然对象的行为测试、二分类系统中的分类测试、假设检验以及理论验证，尽管测试对象、检验基准和具体方法各异，但其检验的本质一致：在规定的条件下执行测试用例，并将实际结果与检验基准所规定的预期进行比较。而测试是通过检验对测试对象行为的量赋值的过程。本章界定了测试的基本概念，并提出了测试方法学。一方面，测试分为证实与证伪：一个反例即可证伪命题，而证实通常需要较为完整的测试覆盖。另一方面，根据检验基准（test oracle）的形式测试可分为基于概率的测试和确定性测试。前者依据检验统计量、p 值和显著性水平作出统计判断，后者则依据规范和参考实现来判断测试是否通过。此外，在实际测试中，由于穷尽式测试往往不可行。因此测试未发现问题仅表明在所执行的测试用例和规定条件下尚未发现问题，并不能证明命题在所有条件下成立，也不能证明所测试的系统完全不存在缺陷。

第七章

推理学：生成新命题和新模型的科学

本章系统性地介绍推理学（*Inferology*）的基本概念、定义、问题定义、基本假说、基本作用、基本原理与边界、统一框架、推理类型的分类与各类推理的详细分析，以及总结。本章由范帆达与我共同撰写。

7.1 推理基本概念定义

推理学概念定义 1（推理或推断（reasoning or inference））. 推理或推断是生成新命题和新模型的一种心智潜力（*capacity*）、实际能力（*capability*）以及过程 *(process)*。推理或推断基于前提 *(premise)*、事实或真理 *(fact 或 truth)*、已有命题和模型以及求索（*interrogation*）结果，按照一定规则生成有关对象或者对象关系的新命题或新模型。

推理学概念定义 2（证据（evidence））. 证据是有关对象或者对象关系的求索结果，可用于支持、补充或反驳某一命题或模型 *[311]*。

推理学概念定义 3（结论（conclusion））. 结论是通过推理得到的命题或模型 *[12, 99]*。

推理学概念定义 4（推理规则（inference rule））. 推理规则是规定如何从已有前提、命题或模型生成出新命题或新模型的规则。不同推理类型可以依赖不同形式的推理规则。

推理学概念定义 5（形式系统（formal system））. 形式系统是由形式语言、公式形成规则、公理和推理规则等要素（组件）构成的结构化系统，用以规定哪些表达式是合法公式，以及哪些公式可以在系统内部由给定前提或公理推出 *[123, 262]*。

推理学概念定义 6（推理合法性（legality of inference））. 在给定形式系统中，推理合法性指一个推导步骤或推理模式符合该形式系统规定的公理和推理规则。从前提集合 Γ 到结论 C 的推导，如果每一步要么是给定的前提，要么是公理，要么可由前面已经得到的公式通过允许的推理规则推出，则该推导在该形式系统中是合法的 *[123, 99]*。

7.2　什么是推理学？

7.2.1　已有学科定义的推理概念

需要特别说明的是，推理并不是一个全新的概念。不同学科有各自的推理术语和问题定义。例如，逻辑学的演绎推理（deductive reasoning）、归纳推理（inductive reasoning）和溯因推理（abductive reasoning）[12, 226, 111]；统计学的统计推断（statistical inference）和估计（estimation）[46, 174]；人工智能的模型推理（model inference）、知识表示与推理（knowledge representation and reasoning）和自动定理证明（automated theorem proving）[29, 95, 17, 263]；因果研究的因果推断（causal inference）、干预和反事实推理（counterfactual reasoning）[220, 224]。

在逻辑学（logic）中，推理通常围绕前提与结论之间的关系展开。演绎推理关注结论是否由前提必然推出，归纳推理关注如何从有限观察形成一般性命题，溯因推理关注如何为已观察到的现象提出可能的解释。逻辑学因此为推理学提供了基本的推理规则、推理合法性保证和解释结构等基础 [12, 92, 99, 226, 111]。在本书中，逻辑学是推理学最重要的基础部分。

在统计学（statistics）中，推理（inference）的核心问题可以表示为“如何从观察到的有限样本（sample）中得出关于总体（population）的结论”[46]。具体而言，统计推断（statistical inference）包括参数估计（parameter estimation）、假设检验（hypothesis testing）和预测（prediction）等基本任务 [46, 84, 207]。与演绎推理不同，统计推断不追求必然为真的结论，而是在数据有限、存在不确定性（uncertainty）和误差（error）的条件下，给出总体参数、分布（distribution）或模型 [174, 24]。

在人工智能领域里，推理（reasoning）或推断（inference）通常指机器系统在给定上下文（contexts）、规则（rules）、模型（models）或数据（data）的条件下，产生命题或预测结果的过程 [29, 95]。

早期人工智能更强调基于逻辑和知识表示的符号推理（symbolic reasoning）[212, 95]，例如自动定理证明（automated theorem proving）[17]。现代人工智能中的推理进一步扩展到不确定性条件下的概率推断（probabilistic inference）：系统可以用概率模型、贝叶斯网络或其他图模型表示变量之间的不确定关系，并在获得新证据后更新信念、计算后验概率或推断隐藏变量 [65, 29, 263]。同时，机器学习模型也承担了大量预测推理（predictive reasoning）任务：系统从训练数据中学习输入与输出之间的映射、分布或表示，并将学习到的模型应用到新样本，或者未发生或不能观察到的求索条件下，生成预测结果 [112, 29, 187]。

在因果研究中，因果推断（causal inference）、干预（intervention）和反事实推理（counterfactual reasoning）关注的是如何从因果假设、观察、实验、模型出发，推断反事实的求索条件下的原因、影响和干预结果 [220, 224, 133, 229]。

7.2.2　本书对推理学的重新界定

我们的目标是在求索科学的体系下，把不同学科中关于推理的概念、问题和方法统一到一个更基础的框架上——把推理视为智能生命的基础求索方法之一。这个框架以“生成新命题和新模型”为核心目标，关注主体如何基于已有命题、模型、事实或真理以及其他求索结果，在一定规则、假说或模型的支持下，生成关于对象或对象关系的新命题或

新模型。在这个意义上，逻辑学（logic）是推理学最重要的基础组成部分；统计推断、人工智能推理和因果推断是推理学在不同领域需要解决的问题。本章的目标是把不同学科里的不同推理形态重新组织为求索科学中的一类基本求索方法：推理。

推理学概念定义 7 (推理学（Inferology))**.** 推理学是研究如何基于已有的前提、事实或真理、命题、模型以及求索结果生成新的命题和新的模型的科学。

7.3　问题定义

在求索过程中，对象的某些属性、结构、机制或者行为可能无法直接测量或测试，隐藏对象可能无法直接观察，对象在未发生或不能观察到的求索条件下的变化也无法通过测量或者测试直接获得。在这些情况下，主体需要基于已有命题、模型、事实或真理以及其他求索结果，生成对象新的命题或模型。因此，推理问题可以表述如下：

问题定义 3 (推理问题定义)**.** 主体如何在一定求索条件下，基于已有命题、模型、事实或真理以及其他求索结果，按照明确的假说、推理规则或模型，生成对象或对象关系的新命题或新模型。

新命题或者新模型可以是特定、未发生或者不能观察到的求索条件下的对象或对象关系的陈述或者简化表示。

推理问题至少包含三个组成部分：第一，推理所依赖的输入，包括前提、事实或真理、已有命题、已有模型以及其他求索结果；第二，推理所采用的依据，包括假说、推理规则或模型；第三，推理所产生的结果，即新命题或新模型。

7.4　基本假说

推理学假说 1 (可检验的推理假说)**.** 只有通过检验的输入才能成为有效的推理输入。只有通过检验的推理规则才能成为有效推理规则。推理生成的新命题或新模型需要被检验。

7.5　推理的基本作用

推理是一种心智潜力（capacity）和实际能力（capability），只有主体才具备这种能力。主体可以通过心智活动来取代物理世界中的真实活动。

推理有以下基本作用:

第一，推理增强了主体理解对象或者对象关系的能力。主体能够通过推理生成关于对象或者对象关系的新的命题或模型；作为一种心智活动，推理可以在一定程度上替代现实世界中的物理活动，并通过推理来减轻部分实际任务。

第二，推理从有限的命题、模型、事实或真理以及求索结果中形成更一般的命题或模型。主体可以基于前提、事实或真理、已有命题和模型以及求索结果，通过归纳推理生成对象的属性、结构、机制、行为或对象关系的命题或者模型。

统计推断、统计学习和机器学习都属于这一作用的具体表现 [46, 112, 29]：它们从有限样本或观察结果中生成关于对象或对象关系的新模型，或将已有命题或模型应用到尚未观察或者不能观察的求索条件下。主体可以基于已有命题、模型、历史求索结果，通过预测推理产生对象未来状态或者不同求索条件下的命题和模型。

7.6　推理的指标

推理基本量定义 1 (演绎有效性（deductive validity）)**.** 演绎有效性指在给定推理规则或形式系统下，结论是否可以由前提必然推出 *[99, 92]*。

设 $\Gamma = \{P_1, P_2, \ldots, P_n\}$ 表示前提集合，C 表示结论。如果在某一演绎系统中有

$$\Gamma \vdash C, \tag{7.1}$$

则表示 C 可以由 Γ 形式推导出来。若从语义角度看，演绎有效性通常写作

$$\Gamma \models C, \tag{7.2}$$

表示在所有使前提（$\in \Gamma$ ）为真的解释或模型中，C 也为真。演绎有效性是逻辑学中最核心的推理质量概念之一。

推理基本量定义 2 (演绎健全性（soundness）)**.** 对于一个具体演绎论证，演绎健全性指该论证形式有效且前提为真 *[99]*。

对于具体论证，可以表示为：

$$\mathrm{Sound}(\Gamma, C) \iff \big(\Gamma \vdash C\big) \wedge \big(\forall P_i \in \Gamma,\ P_i \text{ 为真}\big). \tag{7.3}$$

演绎健全性比演绎有效性有更强的约束条件；一个论证可以形式有效，但如果其前提不真，则它不是健全的论证。

对于一个形式系统，演绎健全性指系统中可证明的命题在相应语义下都有效。

对于形式系统，健全性通常表示为：

$$\Gamma \vdash C \ \Rightarrow\ \Gamma \models C. \tag{7.4}$$

推理基本量定义 3 (推理一致性（consistency），推理完备性（completeness）)**.**

一致性要求系统不能同时推出某命题 P 及其否定：

$$\neg\big(\Gamma \vdash P \wedge \Gamma \vdash \neg P\big). \tag{7.5}$$

推理完备性则要求语义上有效的结论能够在系统中被证明 *[99, 123]*。

$$\Gamma \models C \ \Rightarrow\ \Gamma \vdash C. \tag{7.6}$$

推理基本量定义 4. 统计推断误差（*statistical inference error*）：
统计推断误差指从有限样本推断总体参数、分布或模型时，推断结果相对于目标真值或参考值的偏离 *[46, 174]*。

在统计推断和归纳推理中，常用指标包括偏差（*bias*）、均方误差（*mean squared error, MSE*）、一致性（*consistency*）和置信度（*confidence level*） *[174, 24]*。若 $\hat{\theta}$ 是总体参数 θ 的估计量，均方误差为

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2. \tag{7.7}$$

推理基本量定义 5（预测误差（prediction error））. 预测误差刻画预测结果与后续观察到的求索结果之间的偏离 *[131]*。

在预测推理中，常用点预测误差包括均方误差（*mean squared error, MSE*）、平均绝对误差（*mean absolute error, MAE*）和均方根误差（*root mean squared error, RMSE*）：

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \\ \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}. \end{aligned} \tag{7.8}$$

其中，y_i 是观察到的求索结果，$\hat{y}_i$ 是预测结果。

7.7　推理的基本类型

本章在传统哲学三分法的基础上 [12, 226, 111]，根据生成新命题或新模型的方式，将推理分为四种类型：演绎推理（*deductive reasoning*）、归纳推理（*inductive reasoning*）、溯因推理（*abductive reasoning*）和预测推理（*predictive reasoning*）。我们沿用了基本的传统哲学三分法，不过对一些基本概念进行了泛化。同时，我们加入了预测推理这一新的推理类型，预测对象在尚未观察或者不能观察到的求索条件下的状态。

演绎推理从已接受的前提、规则或公理系统出发，推出在给定系统中逻辑上必然成立的新命题 [12, 123, 99]。归纳推理从已有命题、已有模型或其他求索结果出发，生成超出已有输入的新命题或新模型；在传统的逻辑学（logic）里它可以表现为对一般规律的概括 [1]（generalization），在这里也可推广为对总体、分布或模型参数的推断 [226, 46, 29]。溯因推理从已观察到的结果、现象或证据出发，提出能够解释该结果的原因、机制、假设或模型；其结论是候选解释，需要在后续求索中被检验和修正 [226, 111, 179]。预测推理则基于已有命题、模型、历史求索结果或已归纳出的规律，推断尚未观察到的对象状态或尚未观察到的求索条件下的结果 [131, 29]。

7.8　演绎推理

7.8.1　定义与功能

推理学概念定义 8（演绎推理（deductive reasoning））. 演绎推理是指在推理过程中，若前提为真，则结论必然为真的推理形式 *[12, 123, 99]*。

演绎推理的核心特征在于：推理不增加新的信息内容，而是将已隐含于前提中的逻辑关系显式地展开。传统逻辑学中也常把这类推理称为保真推理（truth-preserving reasoning）[99]。

[1] 逻辑学中 generalization 通常翻译为概括，而在机器学习社区中通常被翻译为泛化。

演绎推理为生成新命题提供了最严格的推理形式。它不是从经验观察中概括一般规律，而是在已接受的前提、定义、公理或推理规则之内，推出其逻辑上必然成立的结论。只要推理步骤符合相应规则，结论就随之成立。从这一意义上说，演绎推理的确定性来自前提与结论之间的逻辑关系：它保持的是“若前提为真，则结论必然为真”的关系，而不是依赖新的经验观察来增加结论的可靠性。

7.8.2　形式化基础与边界

形式逻辑（formal logic）是研究、表示和检验演绎有效性的理论工具。现代形式逻辑通常通过*形式系统（formal systems）* [123, 262] 刻画演绎推理。形式系统通常至少包括符号语言、公式形成规则、公理以及推理规则，用以规定哪些表达式是合法公式，以及哪些结论可以从给定前提中有效推出 [92, 123, 262]。在诸多形式系统中，*一阶逻辑（first-order logic, FOL）* [92] 提供了最具代表性的基础框架。它的核心要素包括：

- *变量（Variables）* $x, y, z, \ldots$
- *谓词（Predicates）* $P, Q, \ldots$
- *逻辑联结词（Logical connectives）* $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$
- *量词（Quantifiers）* $\forall, \exists$

一个典型的表达式形式为 $\forall x\,(P(x) \rightarrow Q(x))$，它表示一个带有全称量化的蕴含命题。

在形式系统中，公理无需在系统内部证明而被接受，推理规则规定从已有命题推出新命题的合法步骤。演绎系统建立在一组公理与推理规则之上，能够生成大量有效命题。因此，演绎推理可以被理解为对这些逻辑变换的有序、受约束的运用，从而推导出那些已然隐含于前提之中的结论。

经典逻辑通常由形式语言、语义解释、公理和推理规则共同刻画；不矛盾律（law of non-contradiction）是其核心原则之一，但不一定在每个形式系统中都作为公理列出。

亚里士多德在其《形而上学》一书中 [13] 表述了这一基本法则：

$$\neg(P \wedge \neg P), \tag{7.9}$$

该法则指出，任何命题 P 不可能在同一时间既为真又为假。这一原则构成了一切一致性推理的基础：相互矛盾的主张不可能同时支持一个有效的结论。

形式系统为演绎推理提供了严谨框架，但形式系统本身也存在边界。库尔特 · 哥德尔提出的不完备性定理 [98] 揭示了这一点。

针对任何一致的、可有效公理化的且具有足够表达能力以刻画基本算术（如皮亚诺算术）的形式系统，库尔特 · 哥德尔提出的不完备性定理揭示了以下两点：

第一不完备性定理：该形式系统是不完备的。这意味着，在该系统中总会存在一些在标准算术解释下为真的命题，但无法仅依赖该系统自身的公理与推理规则加以证明。换言之，该形式系统在可证明性上必然存在“空隙”。

第二不完备性定理：该形式系统无法仅依赖其自身的公理与推理规则来证明自身的一致性。若该系统是一致的，则断言“该系统是一致的”这一命题在系统内部是不可证明的。

需要强调的是，哥德尔不完备性定理并不否定演绎推理的有效性。它揭示的是：对于足够强的形式系统，演绎推理能够保持前提到结论的逻辑必然性，但不能保证该系统能够证明所有可表达的真命题，也不能仅凭系统自身完成对自身一致性的证明。

在上述形式化基础上，演绎推理可以表示为前提集合到结论的有效推导。设 $P_1, P_2, \ldots, P_n$ 表示一组前提（*premises*）（命题或模型），并设 C 表示一个结论（*conclusion*）。记为

$$P_1, P_2, \ldots, P_n \vdash_{\text{ded}} C, \tag{7.10}$$

表示在演绎推理规则之下，结论 C 可以由前提集合 $\{P_i\}$ 必然推出。其中，$\vdash_{\text{ded}}$ 表示演绎推导关系（*deductive derivability relation*），用于描述前提集合与结论之间的推导关系，强调该推理是否有效取决于前提、结论和推理规则之间的形式关系，而不取决于新的经验观察。

$\vdash_{\text{ded}}$ 不同于 $P \to Q$，后者是命题内部的逻辑联结词，表示“若 P，则 Q”。二者层次不同，不能混用。

一个典型的例子是肯定前件规则（*modus ponens*）。给定一个条件命题 $P \to Q$[2] 以及前提 P，在演绎推导关系下可以推出：

$$(P \to Q),\ P \vdash_{\text{ded}} Q. \tag{7.11}$$

这一推理模式与 P 和 Q 的具体语义内容无关——其合法性完全源于逻辑结构本身。

演绎推理的一个典型示例

前提 *1*：所有偶数都能被 2 整除。
前提 *2*：14 是一个偶数。
演绎结论：因此，14 能被 2 整除。

7.8.3　方法学范式

上述内容说明了演绎推理的定义和形式化基础。本小节说明演绎推理在不同领域和工程实践中的几种典型实现方式。

三段论演绎（Syllogistic Deduction） 最早的演绎框架可以追溯至亚里士多德的《工具论（*Organon*）》 [12]。三段论具有如下的典型形式：

$$\text{所有具有属性}A\text{ 的对象都具有属性}B;\quad \text{对象}c\text{ 具有属性}A;\quad \therefore \text{对象}c\text{ 具有属性}B. \tag{7.12}$$

其中，A 和 B 表示对象的属性，c 表示一个具体对象。第一项前提断言：只要某个对象具有属性 A，它也具有属性 B；第二项前提断言：对象 c 具有属性 A。在这两个前提成立的条件下，可以以逻辑必然性的方式推出结论：对象 c 具有属性 B。

这里的 $\therefore$ 符号表示”因此”。

[2] $P \to Q$：读作“若 P，则 Q”

公理化证明（Axiomatic Proof） 在现代数学和形式逻辑中，演绎推理通常表现为公理化证明：主体从公理、定义、已证命题和推理规则出发，经过一系列有效推导，得到新的命题或定理 [92, 123, 262]。例如：

$$\forall x\ (A(x) \to B(x)),\ A(a) \ \vdash\ B(a). \tag{7.13}$$

这个式子表示：如果所有满足 A 的对象也满足 B，并且对象 a 满足 A，那么可以推出对象 a 满足 B。数学证明是公理化证明的典型实例。它要求每一步都符合相应的推理规则，从而保证只要前提成立，结论就随之成立 [123]。直接证明（direct proof）、反证法（proof by contradiction）和数学归纳法（mathematical induction）都可以作为数学证明的方法。这里的数学归纳法虽然名称中含有“归纳”，但它依赖自然数的递归结构和归纳公理进行演绎推导，而不是从有限经验观察中进行概率性概括。

自动化演绎（Automated Deduction） 当演绎推理由计算系统按照给定规则自动执行时，它就表现为自动化演绎。典型工具包括逻辑程序设计 [161]（如 Prolog）、基于归结的定理证明器 [16, 17]、SAT/SMT 求解器 [14, 100] 以及描述逻辑推理器 [312]。这些工具的共同特点是：把前提、规则或约束转化为可计算的形式，并通过搜索、匹配或约束求解来判断某个结论是否可以推出。在人工智能领域，自动化演绎可用于知识库维护 [319]、规则一致性检查以及安全约束的强制实施 [172]。

形式验证（Formal Verification） 形式验证是演绎推理在计算系统正确性证明中的应用。其目标是证明软件或硬件系统满足预先给定的性质或要求。主要方法包括模型检测（model checking）[60, 91]，即系统地检查系统可能状态是否满足给定性质；以及定理证明（theorem proving），即通过交互式或自动化证明工具构造证明。形式验证的作用在于把系统正确性问题转化为演绎推导问题：如果系统描述、性质要求和推理步骤都被接受，那么证明得到的正确性结论也随之成立。

7.9　归纳推理

7.9.1　定义与功能

推理学概念定义 9 (归纳推理（inductive reasoning）)**.** *归纳推理是指主体基于适用于有限求索条件下的命题、模型或求索结果，概括更一般求索条件下的新命题或新模型的推理形式 [226, 46, 29]。*

传统逻辑学中也常把这类推理称为扩展推理（ampliative reasoning）[226]，因为其结论超出了单个前提直接包含的信息。

归纳推理使主体能够从有限输入中形成更一般的命题或模型。例如，主体可以从若干次观察中概括对象的某种属性，也可以从有限样本中估计总体参数，还可以从训练数据中学习一个用于刻画对象关系的模型。因此，本章定义的归纳推理概念比传统逻辑学中的归纳概念范畴更广：传统的归纳只是其中一种基本形态，归纳推理还包括统计推断、统计学习、机器学习和模型拟合等，它们是归纳推理在统计学和人工智能中的现代实现。

归纳推理的一个简单示例

观测 *1*：昨天太阳从东方升起。
观测 *2*：今天太阳从东方升起。
归纳结论：太阳明天将从东方升起。

该结论并不能从任何一次单独观察中以逻辑必然性的方式推出；它是主体基于多个观察结果之间的相似性形成的候选命题。新的观察结果可能支持、修正或推翻这一候选命题。

7.9.2　归纳支持与可靠性约束

演绎推理的关键问题是推理步骤是否保持逻辑必然性；归纳推理的关键问题则是有限输入能够在多大程度上支持新的命题或模型。由于本章采用的是更广义的归纳推理概念，归纳结论的可靠程度通常可以从输入、假说和输出不确定性三个方面来约束。

1. 输入代表性（*input representativeness*）。归纳推理依赖有限输入，生成超出输入范围的结论，因此有限输入是否能够代表目标对象、对象属性或对象关系，是归纳结论能否泛化的关键。若样本（sample）、观察（observation）或求索结果（interrogation outcomes）存在偏差（bias），归纳结论也会随之偏离 [46, 112]。

2. 假说合理性（*assumption plausibility*）。归纳推理并不是从数据（data）中自动产生真理，而是在一定假说（assumptions）、模型（models）或学习规则（learning rules）的支持下生成新命题或新模型。例如，统计推断依赖分布假说（distributional assumptions）或抽样假说（sampling assumptions）；机器学习依赖模型结构（model structure）、损失函数（loss function）和训练规则（training rules） [174, 29]。

3. 输出不确定性（*output uncertainty*）。归纳结论通常需要附带不确定性刻画（uncertainty characterization），例如估计误差（estimation error）、置信区间（confidence interval）、后验分布（posterior distribution）、泛化误差（generalization error）或预测区间（prediction interval）。这些刻画不确定性的量并不能把归纳结论变成演绎结论，而是用于量化归纳结论在给定假说和数据条件下的支持程度 [174, 29]。

7.9.3　不确定性建模

归纳推理从有限的求索结果概括更一般的命题或模型，因此需要同时建模抽样波动、参数未知、模型选择以及从已观察条件推广到目标条件时的不确定性。不同方法学范式采用的表示方式并不相同：

1. 经验泛化。对缺少概率模型的经验归纳，不确定性主要通过限制命题的适用条件、保留反例以及比较独立求索结果的一致程度来表示 [226]。此时较稳妥的输出是“在已观察范围内得到支持的可修正命题”，而不是无条件的全称命题。观察范围越窄、反例越多或对象差异越大，概括到未观察条件时的不确定性越大。

2. 频率学派统计推断。该范式把样本 D 看作在重复抽样机制下产生的随机结果，通过估计量的抽样分布、标准误差和置信区间刻画有限样本所导致的波动 [46, 174]。

置信水平为 $1-\alpha$ 的区间程序满足

$$\Pr_{\theta}\{\theta \in CI_{1-\alpha}(D)\} \geq 1-\alpha, \tag{7.14}$$

其中概率针对重复抽样得到的随机区间，而不是在数据已经固定后直接给参数赋予概率 [46, 174]。

3. *贝叶斯归纳*。贝叶斯方法通过先验分布、似然函数和后验分布联合表示参数与模型不确定性：

$$p(\theta, M \mid D) \propto p(D \mid \theta, M)\, p(\theta \mid M)\, p(M), \tag{7.15}$$

M 表示候选模型，θ 表示模型参数，$\propto$ 表示“正比于”（proportional to），即左侧等于右侧乘以一个归一化常数。从该后验分布可进一步导出可信区间、特定假设的后验概率以及模型平均预测等推理结果 [139, 29]。

4. *统计学习与机器学习*。这类方法通过训练集与验证集划分、交叉验证、重抽样 [3]或自助法估计模型对有限样本的敏感程度，并通过正则化、泛化误差界、集成模型或参数分布刻画模型选择和参数估计的不确定性 [317, 112, 29]。训练误差只能反映已观察样本上的拟合程度，不能替代对未观察对象或求索条件下泛化不确定性的估计。

上述量化结果都以抽样机制、分布形式、模型集合或数据生成条件等假说为条件。未被纳入模型的选择偏差、测量误差和分布变化不会自动反映在置信区间或后验分布中，因此还需要通过稳健性分析、敏感性分析或更换假说来考察结论是否稳定。

7.9.4　方法学范式

本小节总结归纳推理在传统逻辑、统计学和人工智能中的几种典型方法学范式。

经验泛化（Empirical Generalization）　经验泛化是传统归纳推理的基本形态。主体从若干观察、测量、测试或其他求索结果中识别相似性，并据此形成关于对象属性、结构、机制、行为或对象关系的一般命题 [226]。经验泛化能够扩展知识，但其结论始终是可修正的候选命题，需要在新的求索结果中继续接受检验。

统计推断与估计（Statistical Inference and Estimation）　统计推断与估计是归纳推理的量化实现。它从有限样本（sample）出发，对主体关心的总体（population）作出推断，例如估计总体参数（population parameters）、刻画分布（distributions）、建立统计模型（statistical models）或给出不确定性范围（uncertainty ranges）。

点估计(point estimation)给出未知参数的单一近似值；区间估计(interval estimation)通过置信区间（confidence interval）或可信区间（credible interval）刻画估计的不确定性；贝叶斯推断（Bayesian inference）[22, 139, 201, 128] 则将先验知识与观察数据结合，通过后验分布（posterior distribution）更新关于参数或模型的支持程度。因此，统计推断并不是演绎地证明有关总体的命题，而是在抽样假说、统计模型和概率度量的支持下，对总体作出可量化、可修正的归纳判断 [46, 83, 174, 24]。

[3]重抽样（resampling）指从已有样本中重新抽取样本来估计统计量分布的方法，如 Bootstrap 等。

统计推断作为归纳推理的一个简单示例

数据：从某批产品中随机抽取 100 件，发现 8 件不合格。
点估计：该批产品的不合格率约为 $\hat{p} = 0.08$。
区间估计：该批产品真实不合格率可能落在一个围绕 $\hat{p}$ 的区间内。
归纳结论：在给定抽样假说下，这批产品的不合格率大约为 8%，但该结论带有由有限样本导致的不确定性。

这一过程体现了统计推断的归纳性质：主体并没有检查整批产品，而是从有限样本出发，对总体的未知属性作出量化推断，同时刻画推断的不确定性。

统计学习与机器学习（Statistical Learning and Machine Learning） 在统计学习和机器学习中，归纳推理表现为从训练数据（training data）中学习模型。统计学习 [317] 通常把学习过程形式化为在候选函数集合中寻找一个函数 $\hat{f}$，使其在训练数据上的损失较小，同时控制模型复杂度（complexity）以降低对未观察数据预测的偏离风险：

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda R(f). \tag{7.16}$$

其中，x_i 表示第 i 个样本（sample）的输入，y_i 表示训练数据中观察到的求索结果，即真实输出或目标值，$f(x_i)$ 表示模型输出或预测值，L 度量模型输出与观察到的求索结果之间的差异，R 用于控制模型复杂度，λ 用于平衡拟合程度与复杂度。需要区分的是：从训练数据中学习模型属于归纳推理；当训练后的模型被用于新对象、未来状态或特定求索条件下的结果推断时，它属于预测推理。

归纳推理从传统经验泛化发展到统计推断、贝叶斯更新和机器学习模型拟合，形成了一组跨逻辑学、统计学和人工智能的方法学范式。它们的共同点都是从有限输入生成超出已有输入的新命题或新模型；它们的差异在于：各自采用不同的假说和模型，并以不同方式建模输出不确定性——从经验泛化的可修正候选命题，到统计推断的置信区间与后验分布，再到机器学习对泛化误差的控制。

7.10　溯因推理

7.10.1　定义与功能

推理学概念定义 10（溯因推理（abductive reasoning））**.** *溯因推理是指主体基于已观察到的现象、证据或其他求索结果，提出能够解释这些结果的新命题或新模型的推理形式 [226, 111, 179]。*

溯因推理由皮尔士（Peirce）首先提出，他将其刻画为“发现的逻辑（the logic of discovery）” [227]。

溯因推理的核心目标不是证明某一解释必然为真，而是生成候选解释，包括可能的原因、机制假设或解释模型。传统上也将其称为解释推理（explanatory reasoning），但本章采用“溯因推理”作为这一基本推理类型的名称。

演绎推理从前提、规则或公理系统出发导出在给定系统中必然成立的新命题；归纳推理从有限观察、样本或其他求索结果中概括出更一般的新命题或新模型。与二者不同，溯因推理从已经出现的现象或证据或其他求索结果出发，反向提出能够解释这些现象或证据的候选假设。它提出的问题不是“什么必然为真”，也不是“什么可能为真”，而是“什么可能解释这些现象或证据”。

溯因推理从观察到的求索结果出发，生成需要进一步检验的解释性命题或模型。由于这些解释只是候选解释，溯因推理的结论具有可撤销性（defeasibility），需要在后续求索中通过测量、测试、评价进一步修正和检验。

形式化表示　设 E 表示已观察到的证据、现象或求索结果，$\mathcal{H} = \{H_1, H_2, \ldots, H_k\}$ 表示一组候选解释假设。每个 H_i 都可能在一定程度上解释 E，但并不必然为真。溯因推理的任务是在候选假设集合中选择当前条件下解释能力最强或最合理或最值得进一步检验的假设：

$$H^* = \arg\max_{H_i \in \mathcal{H}} S(H_i, E), \tag{7.17}$$

其中，$S(H_i, E)$ 表示假设 H_i 对证据或求索结果 E 的解释得分。该得分可以来自解释性、简单性、一致性、机制合理性、后验概率或与已有模型的相容性。

这种表示方式避免把溯因推理解释为一种确定性的证明。即使 H_i 能够解释 E，也不能推出 H_i 必然为真；能解释一个现象，并不等于这个解释一定成立。溯因推理在本质上是非单调的（non-monotonic）：一个当前合理的解释可能会被新的证据或替代性假设（alternative hypotheses）推翻 [231, 77]。

溯因推理的一个简单示例

观察：今天早上地面是湿的。
可能的解释 H_1：昨晚下过雨。
可能的解释 H_2：夜间洒水系统运行过。
溯因结论：最为合理的解释是昨晚下过雨。

如果出现新的证据，例如洒水系统的运行记录，该结论可能会被修正。因此，溯因推理体现了一种*以解释为导向但可能犯错*（*fallible*）的推理方式。

7.10.2　不确定性建模

溯因推理的不确定性来自证据可能不完整或有误、多个假设都能解释同一证据，以及候选假设集合本身可能遗漏真实解释。因而，溯因输出不应只给出一个未经限定的“最佳解释”，还应保留候选解释之间的相对支持程度及其对新增证据和建模假说的敏感性 [179, 77]。

1. 最佳解释推断。对定性的最佳解释推断，可以分别记录简单性、一致性、解释范围、机制合理性等解释性优点，形成候选假设的偏序或排序。最高得分与次高得分之间

的差距、评价准则权重变化后排序是否改变，以及是否存在同样能够解释证据的替代假设，共同表示解释选择的不确定性。解释得分本身通常不是经过校准的概率，不能直接解释为假设为真的概率 [179, 180]。

2. 概率溯因。对候选集合 $\mathcal{H} = \{H_1, \dots, H_k\}$，贝叶斯方法用后验分布保留各解释的相对可能性：
$$\Pr(H_i \mid E, \mathcal{H}) = \frac{\Pr(E \mid H_i)\Pr(H_i)}{\sum_{j=1}^{k} \Pr(E \mid H_j)\Pr(H_j)}. \tag{7.18}$$
后验概率、后验优势比或后验熵可以刻画当前证据能在多大程度上区分候选解释；获得新证据后，后验分布随之更新 [139, 263]。

3. 基于模型的溯因。在因果图、知识库或生成模型中，可以把未观察原因、模型参数和结构选择表示为隐藏变量、参数分布或多个候选模型，并计算与证据相容的原因集合或其后验分布 [220, 231, 29]。通过在这些原因、参数和模型上进行边缘化，可将输入与模型不确定性传播到解释结论 [65, 187]。

上述后验概率始终以候选集合 $\mathcal{H}$、先验和似然模型为条件。如果真实解释不在 $\mathcal{H}$ 中，即使某一候选假设的归一化后验概率很高，也不能据此断言它是真实解释。因此，应保留“其他解释”的可能性，并通过扩大候选集合、改变先验、扰动证据或比较替代模型进行敏感性分析。

7.10.3　方法学范式

溯因推理方法学范式主要包括：

最佳解释推断（inference to the best explanation）

推理学概念定义 11（最佳解释推断（inference to the best explanation））**.** *最佳解释推断 [180] 是溯因推理的一种形式：在若干候选假设中，基于特定的解释性优点（explanatory virtues），选择最能解释已观察现象、证据或求索结果的假设。*

解释性优点通常包括简单性（simplicity）、一致性（coherence）、解释范围（breadth）和解释能力等。

简单性要求假设不过度复杂；一致性要求假设与已有知识或模型相容；解释范围要求假设能够解释更广泛的现象；解释能力要求假设能够以清晰、充分的方式说明现象或证据为何出现。

当存在多个候选假设 $\{H_1, H_2, \dots\}$ 且它们均能够解释证据 E 的情况下，溯因推理的任务在于选取解释性得分（explanatory score）最高的假设：
$$H^* = \arg\max_{H_i} \ \mathrm{Explains}(H_i, E), \tag{7.19}$$
其中，$\mathrm{Explains}(H_i, E)$ 用于表示假设 H_i 对证据 E 的解释能力。该框架通常具有定性成分，依赖于主体基于比较对候选假设的解释性优点进行赋值。

概率溯因（probabilistic abduction）

推理学概念定义 12 (概率溯因)**.** 概率溯因是在给定观察到的现象或证据 E 的条件下，选择后验概率最大的假设 H：

$$H^* = \arg\max_H P(H \mid E), \tag{7.20}$$

其中，$P(H \mid E)$ 表示在观察到现象或证据 E 之后，假设 H 的后验概率 *[232]*。

相比最佳解释推断，概率溯因采取了一种定量化的方法，从贝叶斯推断的视角来理解溯因推理。在这一框架下，该方法通过贝叶斯规则，将假设的先验概率 $P(H)$ 与证据在该假设下出现的似然 $P(E \mid H)$ 相乘，并通过归一化得到后验概率 $P(H \mid E)$。与最佳解释推断相比，概率溯因更强调量化比较；但后验概率最大的假设仍然只是在当前模型和证据条件下最受支持的候选解释，并不能自动成为事实或真理。

基于模型的溯因（model-based abduction）

推理学概念定义 13 (模型溯因)**.** 模型溯因是借助对象及其结构、机制、行为、属性或者对象关系的模型进行推理。

基于模型的溯因被广泛应用于诊断分析与科学建模中，可用于推理的的模型包括因果图 [230]、知识库 [194] 或生成机制 [30]。在这一范式下，如果将某一假设 H 纳入模型后，能够更好地生成、预测或解释已观察到的现象或证据 E，则 H 可以被视为一个合理的候选解释。

7.11　预测推理

预测在现代科学、工程、统计学和人工智能中具有核心地位 [131]。

7.11.1　定义与功能

推理学概念定义 14 (预测推理（predictive reasoning）)**.** 预测推理是指主体基于已有命题、模型、事实或真理以及历史求索结果，推断特定 (尚未发生或不能观察) 求索条件下对象求索结果的推理形式 *[46, 131]*。

预测与估计（estimation）的核心区别在于：估计通常针对当前或过去的未知量、总体参数或模型参数，例如总体均值、回归系数或隐变量状态；预测则针对尚未发生或者不能观察到的求索结果，例如未来结果、新对象的结果、未测试条件下的结果或替代求索条件下的结果。

按照预测对象和输出类型，预测推理主要生成三类新命题或新模型：

1. 关于未来求索结果的新命题或新模型。主体基于已有命题、模型或历史求索结果，推断对象在未来时间点或未来时间段中的求索结果。

2. 关于新对象或新求索条件下的新命题或新模型。主体将已有模型应用到尚未观察的新对象或新求索条件下，推断这些对象或求索条件下可能出现的结果。

3. 关于替代条件下结果的新命题或新模型。主体在给定条件、干预条件或反事实条件下，推断对象状态、结果分布、状态轨迹或替代影响路径。

预测推理的一个简单示例

已有模型：根据过去 10 年的气象数据，建立了某地区月均温度的时间序列模型。
历史证据：今年 1–6 月的月均温度分别为 $3, 5, 12, 18, 24, 29\,^{\circ}\mathrm{C}$。
预测命题：模型预测 7 月月均温度为 $31 \pm 2\,^{\circ}\mathrm{C}$（95% 预测区间）。

这一过程体现了预测推理的核心特征：基于已建立的模型与历史数据，对尚未观察到的未来状态进行推断，并给出不确定性的量化度量。

7.11.2　不确定性建模

预测推理面向尚未观察或不能观察的求索结果，单一的点预测不能完整表达其不确定性。较完整的预测输出应当是预测分布、预测区间或预测集合 [97, 131]。设 D 表示历史求索结果，x_* 表示新的求索条件，θ 表示模型参数，则贝叶斯后验预测分布为

$$p(y_* \mid x_*, D) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid D)\, d\theta. \tag{7.21}$$

该分布同时传播结果本身的随机波动和参数未知造成的不确定性。二者可以通过全方差公式区分：

$$\begin{aligned} \mathrm{Var}(Y_* \mid x_*, D) = {} & \mathbb{E}_{\theta \mid D}[\mathrm{Var}(Y_* \mid x_*, \theta)] \\ & + \mathrm{Var}_{\theta \mid D}[\mathbb{E}(Y_* \mid x_*, \theta)]\,. \end{aligned} \tag{7.22}$$

第一项表示在模型和参数给定后仍存在的结果波动，第二项表示有限数据导致的参数不确定性 [29, 187]。

不同预测范式据此采用不同的具体方法：

1. 统计模型预测通过后验预测分布、预测区间、分位数或重抽样分布表示未来结果的不确定性 [131]。参数的置信区间与未来观测的预测区间含义不同，后者还必须包含结果本身的波动。

2. *AI* 模型预测可以使用模型集成、贝叶斯近似、分位数回归或共形预测生成概率、区间或预测集合 [29]。在交换性等条件下，共形预测可为有限样本提供边际覆盖率保证 [320]；分类概率和区间还应通过校准程度与经验覆盖率进行检验。

3. 机制模型与推演预测为初始状态、边界条件、模型参数和外部情景指定概率分布或取值范围，并通过蒙特卡洛模拟、区间传播和敏感性分析得到可能的状态轨迹集合 [265]。若机制结构本身不确定，应分别比较多个结构模型，而不能只传播单一模型的参数不确定性。

4. 条件、干预与反事实预测分别对条件分布、干预分布或个体反事实结果进行推断。除样本与参数不确定性外，还需要明确因果结构、可识别性和反事实假说所引入的不确定性；对无法由观察数据识别的量，应给出界、敏感性分析或在明确先验下的后验分布，而不应把模型输出当作直接观察到的事实 [220, 133]。

无论采用哪种表示方式，预测不确定性都需要在新的求索结果出现后检验。概率预测应考察校准性，预测区间或预测集合应考察经验覆盖率及宽度；当新的求索条件偏离训练或建模条件时，还应单独报告分布变化或模型外推造成的不确定性。

7.11.3　方法学范式

预测推理的方法学范式可以按照预测所依赖的输入来分类，但不同范式之间并不互斥。

统计模型预测（statistical model prediction）

推理学概念定义 15 (统计模型预测（probabilistic model）)**.** *统计模型预测以概率模型（probabilistic model）、回归模型（regression model）、时间序列模型（time series model）或状态空间模型（state-space model）为依据，从已有样本或历史序列推断尚未观察到的结果。*

统计模型预测的典型例子包括基于回归模型预测新对象的响应值，基于自回归移动平均模型（ARIMA） [34] 或状态空间模型 [110] 预测未来时间点的状态，以及基于统计学习模型推断新输入 x_{new} 的输出 $\hat{y} = \hat{f}(x_{\text{new}})$ [112, 29]。这类方法的核心是从有限数据中估计可外推泛化的统计关系，并将该关系应用到未观察样本、未来时间或新的求索条件中。

AI 模型预测（AI model prediction）

推理学概念定义 16 (AI 模型预测)**.** *AI 模型预测是指将训练好的人工智能模型应用到新的输入、对象或求索条件上，生成标签、数值、序列、文本、图像或其他输出结果。*

例如图像分类模型（image classification model）根据输入图像预测类别；语音识别模型根据声学信号预测文本；大语言模型（large language model）根据上下文预测下一个 token；基于深度学习的时间序列模型（deep learning based time series model）根据历史序列预测未来状态 [318, 101, 29, 38, 304, 325]。此类预测的模型是从训练数据（training data）中学习到的参数、表示或输入输出映射；其可靠性取决于训练数据、模型结构、训练过程以及新求索条件与训练条件之间的一致性 [217]。

机制模型与推演预测（mechanistic model and simulation-based prediction） 当主体能对对象的机制进行建模时，可以基于机制模型进行预测。机制模型可以是微分方程（differential equation） [15]、规则系统（rule-based system） [95]、领域方程（domain equation） [150]、专家模型（expert model） [185] 或对象因果模型（object causal model）。

预测过程通常表现为前向模拟（forward simulation）：给定初始状态、边界条件和模型规则，推导系统在未来时间或特定求索条件下的状态。例如，物理系统可以用动力学

方程推算未来状态，流行病学中的 SIR 类模型可以预测传染病传播趋势 [153]。当系统具有复杂反馈、高维状态或非线性动态时，也可以采用系统动力学、基于主体的仿真或世界模型进行推演 [88, 108]。

条件预测、干预预测与反事实预测（conditional prediction, interventional prediction, counterfactual prediction） 条件预测、干预预测与反事实预测是预测推理在不同求索条件下的特殊形式。一般预测关注在当前命题、模型和求索条件继续成立时，尚未观察到的结果如何出现；条件预测关注在给定某些求索条件后，结果会如何出现；干预预测关注对某个对象、变量或条件进行控制后，系统会如何变化；反事实预测则关注在已经观察到某个事实之后，在替代求索条件下的潜在结果 [224, 253, 175]。

条件预测、干预预测与反事实预测通常基于已有模型、已有证据或已观察事实，以及某个控制的求索条件，生成关于尚未观察、不能观察或者替代求索条件下的新命题或新模型。这里的已有模型可以是结构因果模型、对象因果模型、状态空间模型、机制模型或世界模型；被设定的求索条件可以是干预条件、反事实条件或替代轨迹；推理结果可以是关于结果分布的命题，也可以是在特定条件下生成的状态轨迹、替代影响路径或更新后的模型。

7.12 总结

推理并不是一个全新的研究对象，逻辑学、统计学、认知科学、人工智能、因果研究和论证理论已经形成了关于推理、推断、证明、预测、解释和论证等复杂的概念、理论和方法。本书的工作不是彻底放弃这些学科的既有定义，而是基于精简和一致的基础概念体系，在求索科学的框架下将它们统一在更一般性的“推理学”问题定义上：即主体如何基于已有命题、模型、事实或真理以及历史求索结果，生成新的命题或新的模型。本章围绕“推理学”建立了一个统一的推理框架，包括演绎推理、归纳推理、溯因推理和预测推理。