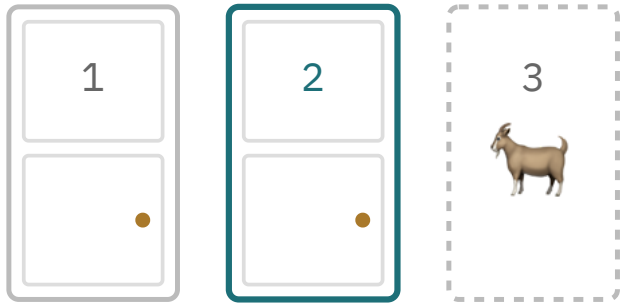
概率成为扩展的逻辑

主持人打开一扇门。你的直觉说：换不换无所谓。但直觉这东西，三分之二的概率是错的。



● 你选了 1 号门 · 主持人打开 3 号门 · 该换到 2 号门吗？

一档七十年代的电视游戏节目，竟把整个贝叶斯推理的精要藏在了门后。

你选了 1 号门。三扇门后，一辆跑车静待赢家，另外两扇各拴一只山羊。主持人心中有数，踱到 3 号门前，轻轻一推，露出一只羊，然后笑意盈盈地问：「要不要换到 2 号门？」剩下两扇门，一辆车。这不就是五五开吗？换与不换，能有什么差别。

差别大了。坚持到底，赢车概率只有三分之一；改换门庭，胜率跃升至三分之二——你几乎什么都不用做，只消改一个主意，胜算便翻了一倍。这就是蒙提霍尔问题。1990 年，它登上杂志专栏，随即引爆了数学史上最大规模的集体误判。今天我们会看到，正确答案为何不仅成立，而且「不可避免」；而解开这道谜题的那套机制，也正是人类在不确定世界中推理所能依赖的最深刻理论。

在第 1 日，我们认识了置信度——那只从 0 到 1 的信念旋钮——以及荷兰赌论证：旋钮若不自洽，便会被人组出一套稳赚不赔的赌局。今天我们学习证据降临时，这只旋钮必须怎样转动：贝叶斯定理。在第 2 日，我们看到科学如何在信号与噪音之间艰难划线，复现危机正是那条战线上的真实炮火；而今天的前沿——一场以「下注」取代 p 值的静悄悄革命——正是为了根治这一痼疾。今日点亮的线索有：信息（证据如比特，更新信念）、计算（心智与实验室都是推理引擎），以及「贝叶斯大脑」重现时一闪而过的能量。

集体翻车

全国最聪明的一批人，同时栽了跟头

1990 年 9 月，玛丽莲·沃斯·莎凡特——《吉尼斯世界纪录》认证的最高 IQ 纪录保持者，在《Parade》杂志[1]主持「问玛丽莲」专栏——回答了一位读者关于游戏节目的提问。换门，她写道，三分之二的概率赢。答案没错。随后，她的信箱炸了。

蒙提·霍尔的真实现场说明，这个谜题远非文字游戏：主持人不是随机开门的旁观者，而是掌握内情的行动者——他的每一步都在泄露信息。

据她自己统计，来信将近一万封，绝大多数在告诉她错了——其中约一千封来自博士。数学家写信训诫她。一位教授留下了那句「经典」评语：

> *「你搞砸了，而且搞砸得一塌糊涂！……这个国家的数学文盲已经够多了，不需要全世界 IQ 最高的人再来添乱。丢人！」*
>
> ——斯科特·史密斯博士，佛罗里达大学，1990 年致《Parade》杂志信

真正搞砸的恰恰是他。严格按概率来算，他的大多数同行也没好到哪去。莎凡特在接下来的三篇专栏里寸步不让，最后干脆请全美教师带着学生用纸杯和硬币做实验。最终她获得了数据的支持，正如她所言：换门胜率是不换的两倍。教授们这才慢慢、且大多不那么体面地认了输。

非要眼见才肯信的那一位

就连保罗·埃尔德什[2]——史上最高产的数学家之一，他证明的定理多数人连题目都读不懂——也拒绝接受这个答案。朋友安德鲁·瓦兹森尼把逻辑讲给他听，他不为所动。直到瓦兹森尼跑了一次计算机模拟，重复几百轮，眼睁睁看着换门在约三分之二的情况下获胜，埃尔德什才勉强点头。即便如此，他仍闷闷不乐：模拟只告诉他「确实如此」，却没告诉他「为何如此」。（见保罗·霍夫曼传记《只爱数字的人》，1998。）如果连埃尔德什都在这件事上栽了跟头，你的困惑便没什么好丢人的。

这场风波暴露了一件事。蒙提霍尔问题不是花招，不是文字游戏——它的答案可证明、可模拟，板上钉钉。它揭示的是：人类对不确定性的直觉存在系统性偏差，我们迫切需要一件形式化的工具来矫正它。这件工具就是今天的主题。但在此之前，先让我们重新看看我们的直觉——再重新塑造它。

蒙提霍尔结果表

初次选择	主持人动作	坚持	换门
跑车，概率 1/3	打开任意一扇山羊门	赢	输
山羊，概率 2/3	被迫打开另一扇山羊门	输	赢

因此，坚持保持最初的 1/3；换门则收割了「第一次选错」那 2/3 的概率。

为何如此

主持人在帮你，也在泄密

最干净的感受方式是记住一点：你第一次选中跑车的概率，只有三分之一。这个数从不会改变。当你指向 1 号门时，车在那扇门后的概率是 1/3，而在「另外两扇门之一」后的概率是 2/3。接着主持人打开了一扇山羊门——关键就在这里：他不是随机挑的。他知道车在哪，而且必须露出山羊。于是，原本平摊在两扇门上的那 2/3 概率，被一股脑压到他唯一没有打开的那扇门上。

主持人的动作不是噪音，而是信息——五条贯穿全书的线索之一，首次以严格的数量形式登场。第 1 日那座停走的钟告诉我们：靠运气说对，不等于知识；而在这里，知情主持人在规则约束下采取的动作就是证据，会推动置信度旋钮。换门，你押的是那沉甸甸的 2/3；坚持，你守的只是最初那孤零零的 1/3。

如果直觉还在抵抗，就把问题放大。想象面前有一千扇门。你随手选了一扇——中奖概率千分之一。知情的主持人随后打开 998 扇，每一扇后面都是山羊，最后只剩下你的门和另一扇。你还会觉得这是一半一半吗？车几乎可以肯定就在主持人刻意避开的那扇门后。三扇门是同一逻辑，只是规模太小，直觉来不及反应。

比电视节目更古老

这个谜题并非始于蒙提霍尔。统计学家史蒂夫·塞尔文早在 1975 年致《The American Statistician》的一封信中就提出了它——而他的后续回应，也是「蒙提霍尔问题」这一名称首次见诸印刷。它的骨架还可追溯得更远：与伯特兰箱子悖论[3]（约瑟夫·伯特兰，1889）和马丁·加德纳的三囚犯问题[4]（1959）在结构上如出一辙。数学家称之为真实悖论（veridical paradox）——答案看似荒谬，却可严格证明。这又是一次趋同再发现，正如第 1 日的盖梯尔案例：当一个世纪里无数聪明头脑反复被同一块石头绊倒，那块石头一定是真的。

模型

贝叶斯定理：信念的更新法则

我们刚才对门做的那套手工操作，其实有个名字，也有个公式。公式看起来冷冰冰，想法却很直接：贝叶斯定理只是在证据到来之后，重新给仍然可能的世界分配权重。

贝叶斯像一只筛子

证据 E =「主持人打开了 3 号门」。每个假设一开始都有同样的先验；随后，谁越能预言这个具体揭示，谁留下的权重就越大。

$$P(H \mid E) = P(H) \times P(E \mid H) \,/\, P(E)$$

H 本行的藏车位置假设	P(H) 本行假设为真的先验概率	P(E | H) 若本行假设为真，主持人打开 3 号门的概率	P(H) × P(E | H) 本行留下的权重	P(主 号 概
H：车在你选的 1 号门后	1/3	1/2	1/6 仍可能，但证据只支持一半	1/
H：车在 2 号门后	1/3	1	1/3 最强幸存者：这个揭示是被迫的	1/
H：车在 3 号门后	1/3	0	0 被排除：主持人不能打开藏车的门	1/

P(E) 是证据筛过之后留下的总权重：**1/6 + 1/3 + 0 = 1/2**。把每个幸存权重都除以这个 1/2，未打开的 2 号门就拿走了剩余概率中的 **2/3**。

$$P(H \mid E) = P(H) \times P(E \mid H) \,/\, P(E)$$

后验（看到证据后的置信度）= 先验（之前的置信度）× 似然（H 预测 E 的能力），再用 证据总量 归一化

用一句话说：看到证据 E 后，你对假设 H 的后验置信度，等于你的先验置信度乘以似然——即 H 预测你会看到 E 的力度——再除以 E 本身出现的总体预期。强有力的证据，是你的假设能预见、而对手却预见不到的东西。整台引擎就这么多。信念永远流向最能解释已发生之事的那一方。

回到蒙提霍尔。令 H =「车在 2 号门后」，E =「主持人打开 3 号门」。如果车确实在 2 号门后，主持人只能开 3 号门（不能开你的门，也不能开藏车的门），所以似然为 1。但如果车在你选的 1 号门后，他本可以开 2 号或 3 号，所以开 3 号的似然只有 1/2。正是这个似然上的不对称，把后验推到了支持换门的 2/3。公式替我们做的，不过是直觉算不好的那笔账。

连医生都会掉进去的陷阱

贝叶斯定理拯救的不只是游戏节目选手。它还能抓住一个著名研究中大多数医生都犯的错误。

下方表格按默认医学检测案例展开这个陷阱；它值得记住，因为它存在于每一次体检、每一个垃圾邮件过滤器、每一道机场安检。

基础概率陷阱

组别	每 1000 人	阳性数
患病者	1	约 1 个真阳性
健康者	999	约 50 个假阳性
全部阳性	约 51	其中只有约 1 人真的患病

因此后验概率约为 0.99 / 50.94，即 1.9%——这就是 Casscells 结果，只是把灵敏度也明说出来。

深层思想

为什么叫「扩展的逻辑」，而不只是一个公式

这就引出了今天的主题。普通演绎逻辑——第 3 日的三段论——是确定性的逻辑：凡人皆有死，苏格拉底是人，所以苏格拉底会死。到此为止，没有余地。但现实生活中几乎没有什么是确定的。我们需要一种逻辑，覆盖「肯定为真」（概率 1）与「肯定为假」（概率 0）之间那片辽阔的灰色地带。令人惊讶的结论是：这样的逻辑本质上只有一种，就是概率演算。

物理学家 R. T. 考克斯在 1946 年把这一结论变成了定理。他问：假设你想给「在已知前提下，这件事有多可信？」指定一个数值，并且只坚持几条常识——可信度必须能用实数表示；同一件事用两种正确方法算出来必须得到同一个可信度（「一致性」）；「非 A」的可信度只取决于「A」的可信度。仅凭这几条朴素要求[5]，考克斯证明，你就「不得不」——不是被建议，而是被强迫——接受标准的概率规则。经过一次不改变实质的重新刻度后，否定必须像 1 - P(A) 那样运作，合取必须服从乘法规则，证据 E 到来时就必须对 E 做条件化。任何自洽的分级信念系统，换件马甲就是概率论。

物理学家 E. T. 杰恩斯的遗著《概率论：科学的逻辑》（2003）正建立在这块基石上。他的口号是：演绎逻辑不过是概率论的特例——所有概率恰好取 0 或 1 的那个特例。概率，就是把逻辑扩展到不确定性领域——也就是扩展到现实世界。请注意，这已经是通往同一终点的第三条独立路径：荷兰赌论证（第 1 日）从「别让人钻空子」出发；而决策论稍后将从「别做被支配的选择」也抵达此处。自洽、无确定损失、一致推理——三条路指向同一套演算。

一个诚实的脚注

考克斯的原始证明略有疏漏。1999 年，计算机科学家约瑟夫·哈尔彭指出，要让证明完全严密还需补一条技术假设（在某些有限域上可能失效），后来的作者做了妥善修补。因此准确的说法不是「概率是不确定性唯一可想象的逻辑」——那过于绝对——而是「在合理条件下，自洽的分级信念必然落入概率公理的框架」。定理依然成立，只是它的桂冠比杰恩斯笔下某些豪言所暗示的要小一号。［已确立，附前提］

争论

两大阵营，同一个方程

概率这样优美而统一的理论，为何能在统计学内部引发一场百年内战？因为方程本身无人质疑，争议在于「那些数字意味着什么」。两派使用的是同一套演算——安德烈·柯尔莫戈洛夫 1933 年写下的公理。这些公理刻意不回答概率「是什么」，只规定它「如何表现」。在这副中立骨架上，两派披上了不同的外衣。

频率派

概率 = 长期频率

- 概率是事件在无限次重复中出现的频率。「硬币公平」意味着抛掷无穷多次，正面比例趋近一半。

贝叶斯派

概率 = 置信度

- 概率是一种置信度——你在已知条件下理性地有多确信（直接来自第 1 日的那只旋钮）。

- 参数是固定但未知的常数；数据才是随机的。你关心的是：你的方法有多大可能误导你。
- 工具：p 值、置信区间、第一/第二类错误（费希尔；奈曼与皮尔逊，1920–30 年代）。
- 说不出「火星上曾有生命的概率是 70%」——火星要么有过生命，要么没有，不存在可重复的样本可供计数。

- 参数自身也获得概率分布；你随数据不断用贝叶斯定理更新它们。
- 工具：先验、后验、贝叶斯因子。谱系：拉普拉斯 → 杰弗里斯 → 拉姆齐 → 德·菲内蒂 → 萨维奇。
- 可以理直气壮地说「火星上曾有生命的概率是 70%」——一次性事件无法重复，但置信度恰好为此而生。

频率派在 20 世纪独领风骚，一半靠道理，一半靠运气。道理在于：它的创立者追求客观性，不信任贝叶斯派的先验，认为那是暗中塞入的主观意见。（费希尔把「逆概率」斥为「必须彻底摒弃」。）运气在于：贝叶斯方法需要大量计算，而廉价计算机姗姗来迟。贝叶斯派至今最敏感的痛点仍是先验——你那个「事前」信念从哪来？凭什么让别人信你的？客观贝叶斯派（杰弗里斯、杰恩斯）寻找规则化的「无信息先验」；主观贝叶斯派则耸耸肩：所有推理总得有个起点。

「概率不存在」

意大利人布鲁诺·德·菲内蒂在专著开篇劈头盖脸就是四个大写英文单词：PROBABILITY DOES NOT EXIST（概率不存在）。他的观点蓄意挑衅：世界上并不存在像质量或电荷那样独立「在那儿」的概率——存在的只是一个理性主体自洽的下注行为。他用一条真正的定理为这句口号背书（1937 年的表示定理）：如果你把一系列观测视为「可交换的」——先后顺序对你无关紧要——那么数学上你就必须表现得仿佛存在一个固定但未知的频率，而你对它持有一个先验。主观信念与看似客观的参数，原来是同一枚硬币的两面。一份用数学写就的停战协议。

从这盘棋里还能落下一条实践智慧：克伦威尔法则（丹尼斯·林德利以克伦威尔 1650 年的恳求命名：「我求你，在心底想一想，你也有可能是错的」）。永远

不要把先验精确地设为 0 或 1，因为贝叶斯定理此后再也改变不了它——被绝对确信的东西，按定义就是不会再因世界而变动的。林德利写道：哪怕给「月亮是绿奶酪」留一丝怀疑余地也好，否则哪怕宇航员真从月球带回了奶酪样本，也休想撼动你分毫。我们再次讨论了校准这一贯穿整个模块的暗线。

前沿 · 2026

针对 p 值的静默兵变

一个世纪以来，频率派的 p 值一直是科学的守门人：跌破 0.05，结果便可称为「显著」。在第 2 日我们看到「显著」再也不是不可动摇的结论了——复现危机中，成山的「显著」发现一经复测便烟消云散。一个主要元凶是结构性的：p 值太脆弱。实验做到一半查看一次数据，发现 p 已跌破 0.05 就立刻收手——你的假阳性率就这样被悄悄抬高。这个过失太过常见，甚至有了专门名称：「选择性停止」（optional stopping）。如今统计学界正在流传一套新框架，从地基开始重建假设检验，正是为了解决这个问题。它的核心对象不是概率，而是一场「赌局」。

前沿 01 [已确立]

e 值：用下注来检验假设

e 值是你对原假设[6]下注后获得的回报。你押上 1 美元赌原假设为假，这份赌约被设计成「在原假设为真时公平」——也就是说，如果原假设确实成立，你便会亏得占不到任何长期便宜（用符号说：e 值在原假设下的期望值至多为 1）。所以，如果最终你的赌注翻了二十倍，那原假设一定哪里出了问题：要么它为假，要么你中了天文数字级别的头彩。一个很大的 e 值，字面意思就是你从原假设身上赢到的真金白银，而你累积的财富就是你的证据。它的倒数 1/e 行为上像个保守的 p 值，但下注的场景才是精髓。

在硬币例子里，原假设很具体：「硬币公平，P(正面) = 0.5」。e 值是两张似然比[7]赌票合起来的财富。一张赌票押「正面偏多」的硬币，即 P(正面) = 0.60：每出现一次正面，这张票乘以 0.60 / 0.50 = 1.2；每出现一次反面，则乘以 0.40

/ 0.50 = 0.8。另一张镜像赌票押「反面偏多」，即 P(正面) = 0.40，倍率正好反过来。把起始的 1 美元平均分到两张票上，无论硬币朝哪边持续偏，都可能让财富增长。如果硬币真的公平，每张票每轮的期望倍率都是 1；这场赌局在原假设下就是公平的。在这个玩具赌局里，「赢」就是财富大到足以拒绝「硬币公平」；「输」就是财富停滞或缩水，说明你还没有赢到反对公平的证据。

这不是松散的比喻，而是一套严格的纲领——「博弈论统计学」，由格伦·谢弗与弗拉基米尔·沃夫克用二十年时间建立，现由阿迪亚·拉姆达斯、彼得·格伦瓦尔德、王若度等人继续推进。谢弗的宣言《以赌注检验》于 2020 年在英国皇家统计学会宣读，2021 年发表于该会《期刊》A 辑。他抱怨 p 值的一个理由正是它太难向人解释；而「我赌这个假设不成立，赢了 20 块」——这话任何人都能听懂。

前沿 02 ［已确立］ ［争议］

为什么下注好过 p 值：实时的局部结论

赌局会复利累积。如果你对原假设下了一场公平的赌，再下一场，再下一场，手头财富就构成了数学家所说的鞅[8]（martingale），而一条经典定理（维勒不等式[9]）保证：若原假设为真，它几乎不可能膨胀出天文数字。这赋予了 e 值一项 p 值望尘莫及的近乎魔法的性质：任意时刻有效性。你可以盯着实验进展，随时喊停，觉得有希望就继续加数据——中途查看多少次都可以——你的错误保证依然成立。格伦瓦尔德、德·海德与库伦称之为「安全检验」（发表于 RSS《期刊》B 辑，2024）；更完整的框架——包括每时每刻都有效的置信区间——叫做「安全任意时刻有效推断」（拉姆达斯、格伦瓦尔德、沃夫克与谢弗，《统计科学》，2023）。e 值合并起来也极为方便：独立的 e 值直接相乘，相依的 e 值取平均，结果仍是有效 e 值——这让跨研究汇总变得干净利落，而 p 值则会一头扎进多重比较的雷区。

下方表格概括同一个对比：脆弱、怕中途查看的 p 值，与诚实的 e 值。

这个玩具任务故意很窄：它要拒绝的只是「这枚硬币是公平的」这个命题；它不是在估计硬币的精确偏差，也不是在绝对证明硬币不公平。

e 值账本

量	含义	用途
E = 1	对原假设没有净赢面	起点
硬币演示赌票	似然比回报：押中的一面出现时乘以 1.2，另一面出现时乘以 0.8	若硬币的真实 P(正面) = 0.5，则期望上公平
E = 20	原假设下公平赌约的二十倍回报	0.05 水平的拒绝阈值，因为 1 / 20 = 0.05
滚动财富	检验鞅或 e 过程	可持续监控，同时控制第一类错误

代价是偏保守：当所有建模假设完全正确时，任意时刻有效的账本可能需要比固定样本量[10]检验更强或更持久的证据。

换到科学场景会是什么样？在一项持续更新的临床元分析[11]里，原假设可能是「BCG 疫苗对医护人员感染 COVID-19 没有临床相关效果」。新的随机试验会在不同时间报告结果，研究者希望一有新数据就更新综合分析，同时又不希望每看一次结果，假阳性风险就悄悄升高。ALL-IN 元分析框架正是为这类场景设计的：它允许后续试验的证据陆续加入，同时保留第一类错误率与区间覆盖率保证。在一个 BCG/COVID 应用中，对证据过程来说，「赢」本来意味着累积到足以支持临床相关获益的强证据；但这项任意时刻有效分析没有发现 BCG 能临床相关地降低感染，而住院结局因事件太少，仍不足以下定论。这和硬币玩具是同一结构，只是把正反面换成了医学终点和陆续到来的试验数据。

前沿 03 [已确立] [炒作]

这场兵变究竟蔓延了多远?

到了区分实诚与吹嘘的时候了。e 值的数学已经确立且优美——经过本领域最顶尖期刊的同行评议（《统计学年鉴》、RSS 两辑《期刊》、《统计科学》），并在 2024 年预印本之后由拉姆达斯与王若度汇集成一本 390 页的《Foundations and Trends》专著。这一部分[已确立]，无可争议。

真实世界中的采纳则是更窄、也更诚实的故事。最清晰的落地在科技公司 A/B 测试 [12]——因为持续查看数据本身就是它们的日常：Optimizely 围绕「始终有效推断」重建了整个平台（Johari、Koomen、Pekelis 与 Walsh），Netflix 与 Adobe 则公开使用任意时刻有效的置信序列 [13]，让产品团队能持续监控实验而不在统计上作弊。这是真正的生产环境应用——但距离全球的生物统计、心理学和物理学共同体还很远，那里 p 值依旧根深蒂固。

新工具也不是毫无代价。在固定样本量比较中，e 值可能需要比 p 值更极端的数据才能跨过同样的拒绝门槛；谢弗的回应是，这是让证据这把尺子变诚实的代价，而非简单缺陷。你的赌局效率取决于下注策略的好坏——说白了，这跟贝叶斯派选择先验时面对的建模判断如出一辙，只是换了套行头。塞缪尔·帕维尔与莱昂哈德·赫尔德等批评者警告：把检验标榜为「安全」或「始终有效」可能有误导之嫌，因为这些保证同样依赖假设（模型设定正确、无发表偏倚），而那些假设可能跟别的假设一样失效。诚实的裁决是：它是 p 值的一个[前景可期]、严格、真正有用的补充——但绝不是科学范围内的全面替代品，至少现在不是。

什么能真正推动局面？如果 FDA 或 EMA 这类药物监管机构批准 e 值用于确证性临床试验，或者某家顶级综合科学期刊把它写进投稿指南，「取代」的口号才有可能从炒作变成现实。让我们拭目以待。

开放问题

真正尚未解决的

- 概率到底是什么？是世界中的频率、心智中的置信度，还是一个公平的赔率？三个世纪过去了，诠释之争有过停火（德·菲内蒂），但从没有投降。
- 先验从何而来？是否存在一种有原则、客观的方式来设定你的「事前」信念，还是一切推理终究立足于一个数学无法替你辩护的选择？

- 基于下注的统计学真能接管吗？还是只会沦为序贯实验的专用工具，而 p 值继续统治其余领域——而且，「选你的赌注」真的比「选你的先验」更不主观吗？
- 大脑真的在运行贝叶斯吗？第 1 日的预测加工线索说，感知就是神经组织中的贝叶斯推断。今天为这个主张提供了规范性骨架——但「大脑近似贝叶斯」和「大脑就是贝叶斯」是两笔截然不同的赌注，我们将在第 119 日重返这个话题。
- 考克斯定理真的对任何理性主体都有效吗——包括人工主体——还是只对那些已经接受了它的一致性公理的主体有效？（这个问题对 AI 板块格外要紧，第 138–145 日。）

◆ 今日三句话

大观念

概率不只是骰子和硬币的工具——它是不确定性领域中逻辑的唯一延伸（考克斯、杰恩斯），而贝叶斯定理是它的运动定律：信念流向最能解释你实际所见之事的假设。

最佳类比

蒙提霍尔打开一扇山羊门——知情者的选择把 2/3 的概率倾注到仅剩的那扇门上；以及赌徒的账本——反对某个假设的证据，字面意思就是你赌它不成立而赢来的钱。

当下争议

频率派与贝叶斯派围绕概率「是什么」的分裂，如今又添了一场 2020 年代的兵变：用脆弱、怕中途查看的 p 值换取 e 值——数学已确立，科技已采纳，但远未成为最狂热支持者许诺的那种科学级革命。

今日线索 › 信息（主持人的揭示和 e 值都是更新信念的证据）· 计算（心智与实验室作为推理引擎）· 能量（对「贝叶斯大脑」的一次轻回调）——而校准这条暗线，从第 1 日和第 2 日一路延续至此。

明日 → 第 05 日

因果

今天我们学会了如何根据证据更新信念——但这些证据只告诉了我们「相关」。冰淇淋销量和溺水人数同步上升，但谁也不是谁的因。明天我们要面对推理中最艰难的一次升级：区分什么只是跟着某物一起起伏，什么才真正「驱动」了它。混杂因素、反事实，以及朱迪亚·珀尔的 do-演算[14]——这套工具问的不是「我预期什么？」而是「如果我插手干预，会怎样？」带上今天的贝叶斯直觉——你需要学会它的边界。

说明

1. 《Parade》是美国发行范围很广的周日报纸副刊杂志；「问玛丽莲」是莎凡特的问答专栏。
2. 埃尔德什发表约 1500 篇论文，深刻影响组合数学、图论、数论和概率方法；数学界还用「埃尔德什数」衡量合作距离。
3. 其中有三只盒子：金金、银银、金银；抽到一枚金币后，另一枚也是金币的概率是 2/3，而非 1/2。
4. 其中三名死囚有一人将被秘密赦免；得知另一名囚犯会被处决后，人们容易作出同样错误的 1/2 更新。
5. 这里的要求指一套理论必须满足的基本条件。
6. 原假设是统计检验试图拒绝的默认说法，通常是「没有效应」或「没有差异」。
7. 似然比比较同一批数据在两个假设之下分别有多可能出现。
8. 鞅是公平赌局的数学模型：在已知过去的条件下，下一步的期望值等于当前值。
9. 维勒不等式限制了公平赌局的财富过程仅凭运气膨胀到很大数值的概率。
10. 固定样本量检验会预先决定样本数，并只在计划好的终点分析一次。
11. 元分析会用统计方法合并多项研究的结果；持续更新的元分析会随着新研究出现而更新。
12. A/B 测试会把用户随机分到产品版本 A 或 B，比较哪个版本表现更好。
13. 置信序列是一组随着数据累积而更新、且在每个时点都保持有效的置信区间。
14. do-演算是珀尔用于推理干预的形式系统：如果我们把某个变量设成某个值，会发生什么？

来源

来源与延伸阅读

1. Selvin, S. (1975). "A Problem in Probability" (Letter to the Editor). The American Statistician 29(1): 67. ——以及后续回应 "On the Monty Hall Problem," 29(3): 134，为该名称首次见诸印刷。

2. vos Savant, M. "Ask Marilyn." Parade (Sept 9, 1990, and follow-ups 1990–91). marilynvossavant.com/game-show-problem ——专栏、读者来信，以及约一万封信 / 约一千位博士的估计（莎凡特本人统计）。

3. Tierney, J. (July 21, 1991). "Behind Monty Hall's Doors: Puzzle, Debate and Answer?" The New York Times. nytimes.com ——包括蒙提·霍尔与 Persi Diaconis 关于主持人协议附注的讨论。

4. Hoffman, P. (1998). The Man Who Loved Only Numbers. Hyperion. ——埃尔德什 / 瓦兹森尼模拟轶事。

5. Bertrand, J. (1889). Calcul des probabilités. Gauthier-Villars. ——伯特兰箱子悖论，结构上的祖先。 另见 Gardner, M. (1959), "Mathematical Games," Scientific American (Three Prisoners)。

6. Casscells, W., Schoenberger, A. & Graboys, T. B. (1978). "Interpretation by Physicians of Clinical Laboratory Results." New England Journal of Medicine 299(18): 999–1001. doi:10.1056/NEJM197811022991808. ——60 名临床医生中只有 11 人给出约 2% 的答案。

7. Cox, R. T. (1946). "Probability, Frequency and Reasonable Expectation." American Journal of Physics 14(1): 1–13. ——迫使概率规则成立的那些条件。

8. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press (ed. G. L. Bretthorst). ——概率成为扩展的逻辑。

9. Halpern, J. Y. (1999). "A Counterexample to Theorems of Cox and Fine." Journal of Artificial Intelligence Research 10: 67–85. ——关于考克斯定理严谨性的附注。

10. Kolmogorov, A. N. (1933). Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of the Theory of Probability). Springer. ——诠释中立的公理。

11. de Finetti, B. (1937 / 1974). "La prévision..."; Theory of Probability (Eng. trans.). ——"PROBABILITY DOES NOT EXIST"；表示定理。

12. Lindley, D. V. (1991). Making Decisions, 2nd ed. Wiley. ——克伦威尔法则（第 104 页）。

13. Shafer, G. (2021). "Testing by Betting: A Strategy for Statistical and Scientific Communication." Journal of the Royal Statistical Society Series A 184(2): 407–431. doi:10.1111/rssa.12647. rss.onlinelibrary.wiley.com ——含发表讨论（包括沃夫克的评论，JRSS-A 184(2): 445–446）。

14. Vovk, V. & Wang, R. (2021). "E-values: Calibration, combination, and applications." The Annals of Statistics 49(3): 1736–1754. doi:10.1214/20-AOS2020. pdf

15. Grünwald, P., de Heide, R. & Koolen, W. (2024). "Safe Testing." Journal of the Royal Statistical Society Series B 86(5): 1091–1128. doi:10.1093/jrsssb/qkae011 (read paper, with discussion incl. Shafer, Pawel & Held). academic.oup.com

16. Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2023). "Game-Theoretic Statistics and Safe Anytime-Valid Inference." Statistical Science 38(4): 576–601. doi:10.1214/23-STS894. arXiv:2210.01948

17. Ramdas, A. & Wang, R. (2025; first posted 2024). "Hypothesis Testing with E-values." Foundations and Trends in Statistics 1(1–2): 1–390. arXiv:2410.23614 ——综合专著。

18. ter Schure, J., Ly, A., Belin, L. et al. (2022). "Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers." Prospective ALL-IN meta-analysis preprint. Amsterdam UMC ——在持续更新的临床元分析中使用 exact e-value logrank tests 与任意时刻有效置信区间。

19. Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2022). "Always Valid Inference: Continuous Monitoring of A/B Tests." Operations Research 70(3): 1806–1821. doi:10.1287/opre.2021.2135 ——Optimizely 的部署；参见 Netflix Research 关于任意时刻有效推断的研究，以及 Adobe Experience Platform 置信序列。

20. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values." The American Statistician 70(2): 129–133. —— 以及 Amrhein, Greenland & McShane (2019), "Retire statistical significance," Nature 567: 305–307。

第 04 日完 · 余下 176 次深入