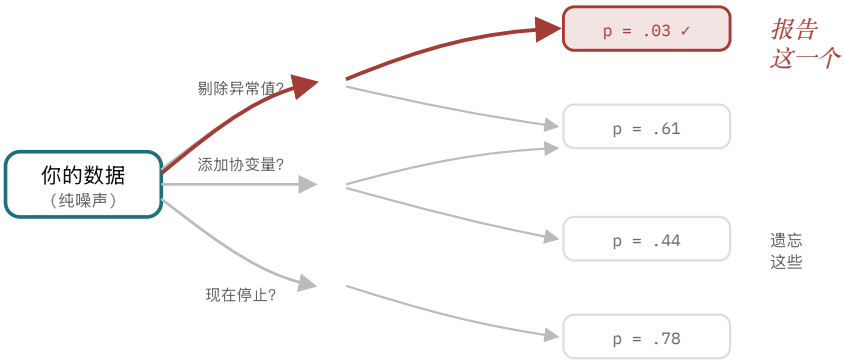
统计学与「不自我欺骗」的艺术

在你的数据中，可能存在一条通往 $p < .05$ 的路径。只要有足够多可辩护的选择，许多数据集里都会出现这样的路径。而统计学这门学科的全部意义，就是为了不让你在无意间踏上它。



分岔路径的花园。数据纯粹是噪声——但只要分支足够多，总能保证其中一个通向「发现」。

2011 年，三位心理学家决定证明一件不可能的事。他们找来 20 名本科生，给其中一半播放披头士的《当你六十四岁时》（When I'm Sixty-Four），另一半则播放一首对照曲目，然后询问了一系列问题——包括每个人的出生日期。在进行了一次完全常规的统计分析后，他们宣布了发现：听《当你六十四岁时》能让人年轻一岁半。不是感觉年轻，而是真的变年轻了——出生日期证明了这一点，且 $p = .04$。

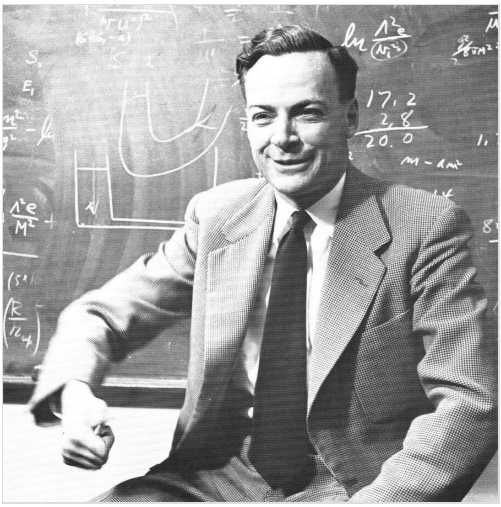
这当然是荒谬的。一首歌不可能跨越时空修改你的出生年份。而这恰恰是重点所在。约瑟夫·西蒙斯（Joseph Simmons）、莱夫·纳尔逊（Leif Nelson）和尤里·西蒙逊（Uri Simonsohn）使用的全是标准且受尊重的统计工具，做出的也全是任何在职科学家每周都会做的选择——却变戏法般地得出了一个统计学上显著的「不可能」。他们的论文是一场刻意荒谬的演示，用来暴露统计流程中的脆弱点。今天，我们将了解它揭露了什么，以及科学界如何学会停止自我欺骗。

在 第 2 日，我们通过一组触目惊心的数字了解了复现危机（原始心理学研究有 97% 「成功」，但仅有 36% 能被重复），并提到了 **p 值操纵**（p-hacking）[1]，但尚未深入探讨。今天，我们将打开这个黑箱。这是危机背后的统计学动力室，也是 第 1 日 提到的盖梯尔担忧（通过运气而非必然联系获得的真实信念）在整个科学文献规模上的体现。我们将深入理解 第 4 日 的教训：p 值并不是「零假设为真的概率」（这种混淆本质上是伪装后的基础率谬误）；并重温 第 5 日 的警告：选择「控制哪些变量」本身就是路径的分岔口。今日线索：信息（信号与噪声）和计算（作为易错推理机的实验室），以及对涌现的初步预览——科学作为一个比任何单个分析师都更宏大的错误修正系统。

引子

最容易被愚弄的人

> *「首要原则是你不能欺骗自己——而你恰恰是最容易被自己愚弄的人。」—— 理查德·费曼，加州理工学院毕业典礼，1974 年*



1959 年，理查德·费曼站在黑板前。上方引文出自他 1974 年在加州理工学院关于「货物崇拜科学」的毕业典礼演讲。图片：*The Big T* 1959，经 Wikimedia Commons，美国公有领域。

费曼谈论的是一种任何公式都无法提供的正直。现代统计学最残酷的转折在于，如果按照人类天生的倾向去使用那些公式，它们反而会成为共犯。它们为你想看到的一切结论，都披上了数学那层温暖而权威的外衣。

这个陷阱有个学名：研究者自由度（researcher degrees of freedom）[2]。每一项真实分析背后都隐藏着几十个微小但可辩护的决定。剔除哪些异常值？是否控制年龄？性别？两者都控制？是现在停止收集数据，还是再找 20 个参与者？你测量了三个指标，该报告哪一个？每一个选择单独看都是清白的。但如果是在瞥了一眼数据走向之后才做决定，它们加在一起就成了一台伪造「发现」的机器。

西蒙斯、纳尔逊和西蒙逊在用披头士做实验之前，先用算术证明了这一点。通过模拟，他们展示了仅仅捆绑四个普通的「自由度」——偷看数据并在需要时增加样本；在两个相关的结果指标中二选一；添加或删去性别等协变量；剔除三个实验条件中的一个——就能将假阳性的概率从宣称的 5% 飙升至 61%。只要拨动足够多清白的开关，像抛硬币一样的随机噪声就能变成近乎确定的「成功」。

下方参考表以静态形式展示同一台机器：每增加一种研究自由度，纯噪声就多了一条看起来像「发现」的路径。

假阳性工厂

分析自由度	具体变动	为何会抬高假阳性率
两个结果指标	测量两个相关结果，报告有效的那个。	噪声获得了两次跨过 $p < .05$ 的机会。
选择性停止	在 n=20 时观察，需要时再增加参与者。	停止规则本身成了数据中的另一条路径。
灵活的协变量	尝试整体结果及子组拆分。	如果在看到数据后才选择，合法的控制变量就变成了多次比较。
剔除实验条件	运行三个小组，报告最好的一对。	报告的对比结果是被特意挑选出来的。

西蒙斯、纳尔逊和西蒙逊展示，即使数据中没有真实效应，结合这四种自由度也能将名义上 5% 的假阳性率提高到约 61%。

核心模型

p 值到底在说什么（以及关于它的六种误读）

为了不被愚弄，你必须确切知道测量仪器报告的是什么。许多读者，包括受过训练的科学家和相当数量的统计学讲师，都会在这里出错。所以，请仔细阅读：

定义 · P 值

p 值（p-value）[3]是在零模型[4]、零假设[5]及所有建模假设均成立的前提下，获得与实际观察到的结果至少同样极端的结果的概率。

换成一个小例子：假设零模型说两组均值相同。如果在这个模型之下，出现当前这么大或更大的差异的概率是 3%，那么 p = .03。这绝不是「零假设为真的概率只有 3%」。

请读两遍。它是关于「给定假设下的数据」的陈述——写作 P(数据 | 零假设)。它绝对不是关于「给定数据下的假设」的陈述，即 P(零假设 | 数据)。混淆这两者正是 第 4 日 在患病测试问题中警告过的「逆概率」错误：P(阳性 | 患病) 并不等于 P(患病 | 阳性)，而在没有先验概率的情况下，无论你盯着结果看多久，都无法将前者转化为后者。以下是需要从你的本能中清除的六种误读——摘自格陵兰（Greenland）及其同事列出的 25 种常见误读清单：

人们最常混淆的三种概率

α 水平[6]说的是：如果零模型为真，一个检验程序的长期第一类错误率是多少；α = .05 意味着真实零假设的检验中大约 5% 会误过门槛。**错误发现率**[7]问的是另一个问题：在你宣布为「显著」的发现中，有多大比例是假的？这取决于基础率、功效和选择机制，而不只取决于 α。**后验概率**[8]又是第三件事，例如 P(零假设 | 数据)，它需要先验概率。5% 的 α 既不是 5% 的错误发现率，也不是零假设为真的概率为 5%。

x 「p = .03 意味着零假设为真的概率只有 3%。」

错误。那是 P(零假设 | 数据)——这是一个需要先验概率的贝叶斯量。而 p 值本身就已经「假设」了零假设为真。

x 「p = .05 意味着结果纯属偶然的概率是 5%。」

错误。「偶然」（零假设）正是计算所依据的前提；它不能既是前提又是怀疑的对象。

x 「p > .05 意味着没有效应。」

错误。证据的缺失并不等于缺失的证据——功效不足的研究经常无法检测出真实的效应。

x 「1 − p 是备择假设为真的概率。」

错误。p 值及其补数都不是关于任何假设的概率。

x 「p 值告诉你结果是否可以重复。」

错误。来自单次噪声研究的低 p 值对下一次研究的参考价值极小。

x 「显著的结果意味着效应很大或很重要。」

错误。只要样本量足够大，微不足道的差异也能变得「显著」。显著性 ≠ 大小 ≠ 重要性。

最后一点是无声的杀手，因此值得单独展开一节。

显著性不等于大小：认识效应量

p 值会把效应量、噪声、样本量和模型假设纠缠在一起。只要投入足够的参与者，即使是微小到毫无意义的差异也能跨过显著性门槛。因此，成熟的统计学坚持单独报告 效应量（effect size）[9]：例如 Cohen's d[10]（以标准差为单位的均值差异）或相关系数 r。雅各布·科恩（Jacob Cohen）提供了粗略的分级——d ≈ 0.2 为小，0.5 为中，0.8 为大——同时也坦然承认这些分级是主观的：「虽然是主观的，但理性的人会发现这些提议的惯例是合理的。」重点不在于标签，而在于：如果一个数字没有附带效应量，它几乎没有告诉你任何值得了解的信息。

显著性与大小

情境	P 值	均值差的 95% CI	COHEN'S D	实际解读
均值差 0.30，n = 20，噪声 = 1.00	.34	[-0.32, 0.92]	0.30	小效应，估计太不精确。
均值差 0.30，n = 500，噪声 = 1.00	< .001	[0.18, 0.42]	0.30	统计上清楚，但大小仍然很小。
均值差 0.80，n = 50，噪声 = 1.00	< .001	[0.41, 1.19]	0.80	按科恩惯例属于大效应。
均值差 0.30，n = 50，噪声 = 2.00	.45	[-0.48, 1.08]	0.15	相对于噪声而言非常小。

要点不是「p 值越小越好」。p 值会变小，可能因为效应更大，也可能因为样本巨大、噪声更低，或三者同时改变。请同时报告效应量和区间。

你也一直在误读的「区间」

置信区间（confidence intervals）[11]通常被推销为 p 值的温和替代方案，它们确实更好——但也隐藏了自己的陷阱。

> 定义 · 95% 置信区间
>
> 95% 置信区间是由一个程序产生的，该程序在许多次假设的重复实验中，有 95% 的次数能涵盖真实值。这里的「95%」描述的是方法的长期可靠性，而不是你眼前这一个特定区间的胜算。

因此，人们自然会说「真实值在这个区间内的概率是 95%」——但在频率学派的世界里，这句话是错误的。你的区间要么包含真相，要么不包含；在你完成研究的那一刻，骰子就已经掷下了。（第 4 日 提到的贝叶斯「可信区间」确实允许这种说法——但那是在你承诺了先验概率之后。工具不同，承诺也不同。）置信区间也不意味着你 95% 的数据都落在其中，更不意味着区间外的数值就不可能。它带给我们的清爽转变是：它给出了与数据相容的效应量范围。这种从「是/否」到「有多少、有多准」的框架转换，正是改革者所称的「新统计学」的核心。

为什么小样本研究会夸大它们侦测到的东西

犯错有两种方式：第一类错误（Type I error）[12]（假警报——宣称存在实际并不存在的效应）和 第二类错误（Type II error）[13]（漏报——未能发现确实存在的效应）。统计功效（statistical power）[14]是你捕捉到真实效应的机会；更多的数据意味着更高的功效。你可能会认为功效不足的研究只会得出「没有结果」的耸肩。但更糟的是：它会主动毒害科学文献。当功效较低时，只有那些由于运气好而被偶然性极度放大的估计值才能跨过显著性门槛——这就是 赢家诅咒（winner's curse）[15]。再加上偏好奖励「显著」发现的发表实践，你就制造出了一份既夸大其词又脆弱不堪的研究记录。2013 年的一项清醒审计（Button 等人的《功效失败》）指出，神经科学研究的功效中位数仅为约 21%；Nord 等人在 2017 年的再分析则有益地限定了这个头条数字：并非每个神经科学子领域都同样功效不足。这个限定不削弱今天的教训：在研究功效不足的地方，真实效应会被漏掉，而浮出水面的效应往往会被放大。

更深层的陷阱

即使不作弊，也会被愚弄

这是最令正直的人夜不能寐的部分。你可能读完了上述内容，誓不操纵 p 值，只运行一个预先计划好的分析并忠实报告——但你的假阳性率依然可能超标。这就是安德鲁·格尔曼（Andrew Gelman）和埃里克·洛肯（Eric Loken）所说的 分岔路径的花园（garden of forking paths）[16]。

他们的洞察细微而又令人不寒而栗。损害并不需要你运行多次分析，而只需要你运行的那唯一一次分析是取决于你恰好看到的数据。假设你的数据稍微有点不同——这里多点噪声，那里有个聚类，你就会做出另一个同样合理的选择：控制年龄而不是收入，对比女性子组而不是全样本，使用中位数而不是平均值。其他分析路径是隐形的，因为你从未运行过它们。但在邻近的世界里，正是你本会采取的所有路径，推高了你的错误率。正如格尔曼和洛肯所言，即使假设是预先设定的，你不需要任何「钓鱼摸鱼」或 p 值操纵，也依然能得到假阳性结果。

注意这与 第 5 日 的呼应。在那里，潜伏的决定是「选择哪些变量进行条件化」——而对 对撞变量（collider）[17] 进行条件化会无中生有地「制造」出相关性。分岔路径则是这种危险的泛化：每一个可辩护的分析选择都是一个交叉口，而数据会静静地把你推向那条更有利于你假设的分支。正直的分析师和 p 值操纵者可能得出完全相同的错误结果，唯一的区别在于，前者并不知道自己是怎么做到的。

针对病症的药方

如果问题在于决策是在看到数据之后才做出的，那么最彻底的疗法就是在看到数据之前就做决策：研究预登记（preregistration）[18]（提前公开记录分析计划）和 注册式报告（Registered Reports）[19]（期刊在结果出来前就评审并接收你的「计划」）。我们在 第 2 日 见过它们；现在我们明白了它们为何有效。它们冻结的是验证性路径；偏离计划的分析仍可报告，但应标为探索性，而不是隐藏在确认性结论里。

争论

是降低门槛，还是彻底废除？

如果 $p < .05$ 如此容易被滥用，该用什么来取代它？在这个问题上，统计学界分成了几个派系——这不是民科与专家的对立，而是严肃统计学家之间的真实分歧。我们可以将这些观点排在一条直线上，从「打补丁」到「彻底推倒」。

图示 · 改革光谱

与 p 值共存的四种方式

这条线上的所有人都同意现状已支离破碎。他们的分歧在于疗法需要多彻底。

重定义 —— 本杰明（Benjamin）等 (2018) 保留了门槛的想法，但将其大幅下调：将 $p < .005$ 称为「发现」，而 .005–.05 仅称为「有启发性」。简单直接，但批评者认为这只是治标不治本。

论证你的 α —— 拉肯斯（Lakens）等 (2018) 拒绝任何通用数字。他们认为应该根据具体情况，权衡「假警报」与「漏报」的真实成本，审慎选择你的门槛，并展示你的推理过程。

弃用显著性 —— 阿姆瑞恩（Amrhein）、格陵兰和麦克沙恩（McShane）(2019) 在《自然》杂志上联合 800 多名签名者，呼吁废除「显著 / 不显著」这种二元对立的习惯——即不再把 $p = .04$ 和 $p = .06$ 看作两个不同的世界。这并不是禁止 p 值，而是禁止那条刚性的分界线。

统计学界的学会 —— 美国统计协会（ASA）曾两度发声：先是在 2016 年发表了包含六项原则的审慎声明，随后在 2019 年发表了一篇更为激进的社论，敦促大家彻底停止使用「统计学显著」一词。（2019 年那篇是编辑观点，并非 ASA 的正式政策——这一区别本身也引发了一场争论。）

从谨慎派到激进派，整个改革光谱的共同点在于一种迁移：不再寄希望于通过一个魔术数字给出「是/否」的判定，而是转向报告效应的大小、不确定程度，以及结论对分析选择的敏感程度。最后这个「敏感程度」（或称结论的脆弱性），正是 21 世纪 20 年代最前沿的阵地所在。

前沿 · 2024–2026

让脆弱性可见——以及前沿校准器

如果单次分析会产生误导，现代的做法极其简单粗暴：运行所有分析，并展示其分布。目前有两种主要方法在做这件事，还有第三种思路通过动员全球科学家对同一份数据进行分析来测试这种理念。一如既往，每项主张都有其对应的标签。

边缘 01 [已确立] [争议/炒作]

多元宇宙分析与设定曲线分析

不再是只为一种分析辩护，而是列举出每一种合理的选择组合——保留还是剔除异常值、是否进行对数转换、控制这个还是那个协变量——并计算出每种组合下的结果。多元宇宙分析（multiverse analysis）[20]（Steegen 等，2016）展示了所有可能结果构成的云图。设定曲线分析（specification-curve analysis）[21]（Simonsohn 等，2020）则将数百种设定排列成一条曲线，并追问：在所有切分数据的合理方式中，效应是否依然稳固？还是只要稍微拨动一下选择就会烟消云散？这种方法诚实且直观，越来越多地受到审稿人的期待。下方的互动环节将让你切身体会这一点。

这里还必须守住一条纪律：有些「合理设定」会改变估计目标[22]。控制 W 只有在 W 是有因果理由的控制变量时，才是在估计同一个问题；秩变换、剔除异常值或删除子组，可能把目标从均值差异改成有序关联、稳健效应，或某个特定人群中的效应。设定曲线能暴露这个分岔，但不能替你决定你原本想问的究竟是哪一个问题。

不过，前沿校准器在这里也很重要。这些是绝佳的「透明度」工具，但作为「推断」工具则较为薄弱。设定曲线分析的作者们自己也承认了一个陷阱：决定哪些设定是「合理」的本身就是一个主观判断，算法无法代劳，而且「消除主观性的目标是无法实现的」。一个意志坚定的辩论者仍然可以精心挑选他的多元宇宙。目前也还没有一种公认的方法能从曲线中计算出单一的有效结论——

这就是为什么 2024 年出现了新的工具（如 PIMA，见下文）试图填补这一空白。 ［已确立］［争议/炒作］

边缘 02 ［已确立］

多人同数：房间里最有力的证据

这项实验应该会改变你阅读每一条头条新闻的方式。拿出一份数据集，设定一个清晰的问题，然后将相同的副本分发给几十个专家团队。他们会得出同一个判定吗？很多时候，不会。

在 Silberzahn 等 (2018) 的研究中，29 个团队（61 位分析师）被问了一个简单的问题：足球裁判是否更倾向于给深色皮肤的球员出示红牌？他们估计的优势比（OR）从 0.89 到 2.93 不等；20 个团队发现了显著的正向效应，9 个团队则没有。耐人寻味的是，分析师的先验信念甚至他们的统计专业知识都无法解释这种差异。在 Botvinik-Nezer 等 (2020) 的研究中，70 个团队（180 位研究者）针对同一份脑成像数据集分析了九个预设假设——结果没有两个团队使用了完全相同的流程；即使他们的基础统计图谱高度相关，他们的「是/否」结论也发生了严重分歧。Breznau 等 (2022) 将相同的数据和假设（移民是否会削弱对社会政策的支持？）交给 73 个团队，观察到估计值从明显的负相关一直散布到明显的正相关。甚至金融领域也有自己的版本：Menkveld 等人的《非标准误差》(2024) 让 164 个团队在相同的市场数据上测试相同的假设，发现团队间的变异性与普通的统计误差旗鼓相当——令人欣慰的是，当增加额外的同行评审环节时，这种变异性有所收缩。

这就是 第 1 日 提到的盖梯尔问题在单份数据集内部的具象化。每个团队的信念都被其专业的分析所「证成」；但它是否为「真」——即它是否连接到现实，而非仅仅连接到他们在花园中走过的那条特定路径——则因团队而异。一次分析只是一根树枝。请以此心态看待它们。

边缘 03 ［前景可期］ ［已确立］

填补推断鸿沟——以及第 4 日的低调替代方案

上述所有内容中尚未解决的问题是如何在多元宇宙中进行诚实的「推断」：你如何从一千个纠缠在一起的分析中得出一个有效的结论？2024 年的一项研究提出了 **PIMA**（多元宇宙分析中的选择后推断；Girardi 等，2024），它提供了一种基于符号翻转的测试方法，旨在为多元宇宙提供真正的联合误差保证，跨越了 2020 年设定曲线的线性模型限制。虽然这是一个经过同行评审且极具意义的工具，但它还很新，尚未成为标准做法——这属于有前景的线索，而非定论。

与此同时，那些无聊但耐用的修补方案正在不断普及：「注册式报告」格式目前已有数百家期刊提供，研究预登记也正在从一种美德变为默认操作。还记得第 4 日 提到的 **e 值**（e-values）[23]和「通过投注进行测试」吗？这种 p 值的替代方案即使在你「偷看」数据或随时停止实验时依然有效——这正对着你在上方的工厂中拨动的「选择性停止」自由度。目前这仍处于研究前沿而非主流实践——前景可期，值得持续关注。

设定曲线摘要

选择	对分析的影响	解读风险
剔除异常值	在估计关联之前排除极端值。	可能使结果更稳定，也可能选择性地移除不利的数据点。
控制变量 W	调整同时驱动 X 和 Y 的背景混杂变量。	通常会缩减虚假关联，但这种控制必须有因果论证支持。
秩变换	在计算关联之前用排名代替原始值。	可以降低对分布形状的敏感性，但也会改变估计的目标。
删除子组	仅分析数据的子集。	可以测试真实的边界条件，也可以制造出一个讨巧的子集。

诚实的报告是展示所有合理设定下的分布情况；而不诚实的报告则是只展示曲线上最漂亮的那一点。

开放问题

尚未定论的领域

- 「统计显著性」是否应该彻底消失？ 改革者在「约束门槛」和「废除门槛」之间摇摆不定。但决策总要在某个地方做出——一种药物要么获批要么不获批——废除派需要解释该如何进行这种决策。
- 多元宇宙是否真能产生一个诚实的判定？ 还是说，对「合理设定」的选择本身就是一种无法消除的主观行为，只是将分岔路径向上传递了一个层级？ PIMA 等工具是早期的尝试，目前尚无定论。
- 结论的变异等同于现实的变异吗？ 2023 年的一项细致重分析指出，在一些「多人同数」研究中，引人注目的分歧其实是关于「显著性」判定的，而底

层的效应量其实相当一致且微小。对结论的分歧可能会超过对数字本身的分歧。不要过快得出「一切都毫无希望」的教训。

- 未来的天下属于 p 值还是 e 值？ 基于投注的工具优雅地解决了「选择性停止」问题，但可能需要更多数据，并引入了新的建模负担（该怎么下注？）。两者共存似乎比一方征服另一方更有可能。
- 以及 AI 单元中等待着的问题： 当一个在数百万篇论文上训练出来的模型报告一个「稳健」的结果时，它是在推理证据，还是在模仿那些让我们陷入现状的「分岔路径」习惯？ （见 第 138-145 日）。

◆ 今日总结：三句话

核心理念

统计工具并不能保护你免于自我欺骗——如果带着人类天生的灵活性去使用，它们反而会助纣为虐。因为任何分析中都存在几十个合情合理的选择（即「分岔路径的花园」），足以让噪声伪装成发现。因此，疗法在于透明度、研究预登记、附带不确定性的效应量，以及在多种合理分析下的稳健性，而非盯死一个魔术般的门槛。

最佳类比

假阳性工厂：给自己足够的「自由」，就能让抛硬币般的虚无变成 61% 的「成功」机会；以及分岔路径的花园，即使你只运行了一个分析，你「本会」在平行世界里运行的其他路径也会推高你的错误率。

前沿争议

是下调显著性门槛（$p < .005$）、按具体情况论证门槛，还是彻底弃用「统计显著性」——多元宇宙和「多人同数」研究揭示了单份数据集的结论是多么惊人地取决于分析它的人。

今日线索 › 信息（将信号从噪声中分离；p 值作为一种常被误读的证据度量）· 计算（作为易错推理机的实验室；多元宇宙作为穷举法的应用）· 涌现（科学作为一个大于任何分析师的纠错系统——研究预登记、多人同数众测、元分析）。

明日预告 → 第 07 日

信息论

今天的主题是不把噪声误认为信号。明天我们将学习如何精确地「测量」信号。克劳德·香农的「比特」、作为惊喜的「熵」、二十个问题的游戏、信道容量——以及通往物理学的惊人桥梁：兰道尔原理，即抹除一个比特信息所需的不可削减的能量成本。从第 1 日起我们就一直在悄悄拉动的那根「信息」线，终于将迎来属于它自己的数学体系。

说明

1. p 值操纵意味着尝试足够多的分析方式、剔除标准、结果指标或停止规则，直到随机噪声看起来具有统计学显著性。
2. 研究者自由度是指研究者在数据收集、清理、建模和报告过程中，可以做出的各种看起来合情合理的分析选择。
3. p 值是特定零模型下数据的尾部概率，而非零假设为真的概率。
4. 零模型是 p 值计算所假定的世界，通常包括无效应模型以及所有抽样和建模假设。
5. 零假设是正在被测试的基准断言，通常指没有效应或没有差异。
6. α 水平是在查看数据之前承诺的假警报率：在零模型为真且检验程序已指定时，长期第一类错误率是多少。
7. 错误发现率是在所有被宣布为发现的结果中，实际上为假的比例。
8. 后验概率是在结合数据与先验信息之后，某个假设为真的概率。
9. 效应量报告的是差异或关联的大小，而不只是它是否跨过了显著性门槛。
10. Cohen's d 以标准差为单位表示两个平均值之间的差异。
11. 置信区间是由某种具有特定长期覆盖率（如 95%）的方法产生的范围。
12. 第一类错误是假阳性：错误地拒绝了一个真实的零假设。
13. 第二类错误是假阴性：漏掉了真实的效应。
14. 统计功效是指如果某种效应真实存在，研究能检测出该效应的概率。
15. 赢家诅咒是指在充满噪声的研究中，被选中的显著估计值往往会夸大真实的效应。
16. 分岔路径的花园是指研究者在看到同一份数据的不同模式后，本可以合理选择的所有分析路径的集合。
17. 对撞变量是受两个其他变量影响的变量；对其进行条件化可能会在它的原因之间制造出误导性的关联。
18. 研究预登记意味着在查看结果之前，公开记录假设、数据处理和分析计划。
19. 注册式报告是在结果揭晓前，仅凭研究问题和方法通过同行评审并被期刊原则上接收的文章。
20. 多元宇宙分析运行多种可辩护的分析选择，并报告结果的分布，而非单一的推导路径。
21. 设定曲线分析将许多可辩护的模型设定按估计效应排序，以揭示结果的稳健性或脆弱性。

22. 估计目标是研究试图估计的精确量，例如原始关联、因果效应，或某个特定人群中的效应。
23. e 值通过针对零假设的投注策略的回报来衡量证据强度。

参考文献

来源与延伸阅读

1. Feynman, R. P. (1974). "Cargo Cult Science." *Engineering and Science* 37(7): 10–13. 加州理工学院 1974 年毕业典礼演讲；开篇引文来源。 calteches.library.caltech.edu

2. Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science." *Science* 349(6251): aac4716. doi:10.1126/science.aac4716；97% 原研究显著 / 36% 复现显著统计的来源。 doi.org/10.1126/science.aac4716

3. Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11): 1359–1366. doi:10.1177/0956797611417632。 doi.org/10.1177/0956797611417632 – 《当你六十四岁时》演示；5%→61% 的模拟。

4. Gelman, A. & Loken, E. (2014). "The Statistical Crisis in Science." *American Scientist* 102(6): 460–465. – 「分岔路径的花园」；更早为 2013 年哥伦比亚大学工作论文。 americanscientist.org

5. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. & Altman, D. G. (2016). "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." *European Journal of Epidemiology* 31(4): 337–350. doi:10.1007/s10654-016-0149-3。 doi.org/10.1007/s10654-016-0149-3 – 25 种常见误读的权威清单。

6. Haller, H. & Krauss, S. (2002). "Misinterpretations of Significance: A Problem Students Share with Their Teachers?" *Methods of Psychological Research Online* 7(1): 1–20. – p 值误解即使在统计学教师中也持续存在的调查证据。 epub.uni-regensburg.de/34338

7. Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. & Longobardi, C. (2016). "Misconceptions of the p-value among Chilean and Italian Academic Psychologists." *Frontiers in Psychology* 7: 1247. doi:10.3389/fpsyg.2016.01247 。 doi.org/10.3389/fpsyg.2016.01247

8. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2): 129–133.

doi:10.1080/00031305.2016.1154108 。 doi.org/10.1080/00031305.2016.1154108 — 六项基本原则。

9. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019). "Moving to a World Beyond 'p < 0.05'." *The American Statistician* 73(sup1): 1–19. doi:10.1080/00031305.2019.1583913。 doi.org/10.1080/00031305.2019.1583913 — 43 篇文章特刊的社论；「停止说统计学显著」（编辑观点，非 ASA 正式政策）。

10. Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). "Redefine statistical significance." *Nature Human Behaviour* 2(1): 6–10. doi:10.1038/s41562-017-0189-z 。 doi.org/10.1038/s41562-017-0189-z

11. Lakens, D., Adolfi, F. G., Albers, C. J., et al. (2018). "Justify your alpha." *Nature Human Behaviour* 2(3): 168–171. doi:10.1038/s41562-018-0311-x 。 doi.org/10.1038/s41562-018-0311-x

12. Amrhein, V., Greenland, S. & McShane, B. (2019). "Scientists rise up against statistical significance." *Nature* 567(7748): 305–307. doi:10.1038/d41586-019-00857-9；800 余名共同签名者。 doi.org/10.1038/d41586-019-00857-9

13. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., et al. (2013). "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14(5): 365–376. doi:10.1038/nrn3475；功效中位数约 21%。 doi.org/10.1038/nrn3475

14. Nord, C. L., Valton, V., Wood, J. & Roiser, J. P. (2017). "Power-up: A Reanalysis of 'Power Failure' in Neuroscience Using Mixture Modeling." *The Journal of Neuroscience* 37(34): 8051–8061. doi:10.1523/JNEUROSCI.3592-16.2017；关于神经科学低功效是否均匀分布的限定。 doi.org/10.1523/JNEUROSCI.3592-16.2017

15. Center for Open Science. "Registered Reports." COS 官方说明和参与期刊列表；「300 多家期刊」说法的来源。 cos.io/initiatives/registered-reports

16. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5): 702–712. doi:10.1177/1745691616658637。 doi.org/10.1177/1745691616658637

17. Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). "Specification curve analysis." *Nature Human Behaviour* 4(11): 1208–1214. doi:10.1038/s41562-020-0912-z 。 doi.org/10.1038/s41562-020-0912-z

18. Silberzahn, R., Uhlmann, E. L., Martin, D. P., et al. (2018). "Many Analysts, One Data Set." *Advances in Methods and Practices in Psychological Science* 1(3): 337–356.

doi:10.1177/2515245917747646；29 个团队，OR 0.89–2.93。doi.org/10.1177/2515245917747646

19. Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., et al. (2020). "Variability in the analysis of a single neuroimaging dataset by many teams." *Nature* 582(7810): 84–88. doi:10.1038/s41586-020-2314-9；70 个团队。doi.org/10.1038/s41586-020-2314-9

20. Breznau, N., Rinke, E. M., Wuttke, A., et al. (2022). "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty." *PNAS* 119(44): e2203150119. doi:10.1073/pnas.2203150119；73 个团队。doi.org/10.1073/pnas.2203150119

21. Mathur, M. B., Covington, C. & VanderWeele, T. J. (2023). "Variation across analysts in statistical significance, yet consistently small effect sizes." *PNAS* 120(3): e2218957120. doi:10.1073/pnas.2218957120；提醒读者用效应量和区间、而不只是显著性标签来理解「多人同数」分歧。doi.org/10.1073/pnas.2218957120

22. Menkveld, A. J., Dreber, A., Holzmeister, F., et al. (2024). "Nonstandard Errors." *The Journal of Finance* 79(3): 2339–2390. doi:10.1111/jofi.13337；164 个团队。doi.org/10.1111/jofi.13337

23. Girardi, P., Vesely, A., Lakens, D., et al. (2024). "Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test." *Psychometrika* 89(2): 542–568. doi:10.1007/s11336-024-09973-6。doi.org/10.1007/s11336-024-09973-6

24. 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；e-process、testing by betting 与可选停止下仍有效的推断。doi.org/10.1214/23-STS894

25. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum. – 效应量分级惯例，作者本人也承认是刻意随意的。

第 06 日结束 · 还剩 174 次深入