投稿至《统计年鉴》

# 对数悖论：对数变换下混淆统计显著模式反转的必要且充分条件

BY BEN CARDOEN[1,a], HANENE BEN YEDDER[1,b], SIEUN LEE[2,d], IVAN ROBERT NABI[3,e] AND GHASSAN HAMARNEH[1,c]

[1]计算科学系、西蒙菲莎大学，a*bcardoen@sfu.ca*；b*hbenyedd@sfu.ca*；c*hamarneh@sfu.ca*

[2]精准成像中心、诺丁汉大学，d*sieun.lee@nottingham.ac.uk*

[3]生命科学研究所、不列颠哥伦比亚大学，e*ivan.robert.nabi@ubc.ca*

arXiv:2302.04780v1 [stat.ME] 9 Feb 2023

对数变换是统计分析中的常用工具，可减小极端值的影响、压缩报告值的范围以改善可视化效果、支持使用要求数据服从正态分布的参数统计检验，或支持对非线性数据使用线性模型。研究人员很少意识到，对数变换后的结果可能会反转研究发现：不进行变换的假设检验可能显示负趋势，而进行对数变换后则可能显示正趋势，且二者在统计上均具有显著性。我们使用有限差分记号，推导出导致这种悖论式模式反转的充要条件。我们表明，生物医学图像定量分析极易受到这些条件的影响。通过采用一种最大化反转效果的新型启发式方法，我们证明，只需改变数据集中的少至 5% 的数据，就可以轻易诱导出这种悖论式模式反转的统计显著性。我们举例说明，对比例数据中的对象大小进行定量分析，尤其是在对象大小反映潜在的生成与消亡动态变化时，会满足该悖论的前提条件。我们讨论了正确使用对数变换的建议，并讨论了稳健探索潜在模式的方法；同时强调，任何变换后的结果都应始终附带其未变换的原始对应结果，以排除偶然的混杂发现。

## 1. 引言。

1.1. 研究动机。 对单维正向量进行对数变换是统计分析中的常见技术，可用于缩小可视化范围、减小极端值的影响 [22]，或使数据能够应用仅适用于正态分布数据的参数方法 [36]。对数变换是 Tukey 变换阶梯中的一种，也称为“鼓胀规则”，指的是它对分布产生的图形重塑效果。对数及相关变换会在统计学入门课程中讲授 [19, 23]，，而文献 [11]对这些变换作了通俗易懂的介绍。分布的尾部是其概率密度函数（pdf）的极端部分。对数变换的常见应用场景之一是处理服从长尾分布的数据 [19, 23, 11]。直观而言，这类分布中的元素随着数值大幅增加，其出现频率会迅速降低。换句话说，越接近最大值，数值不仅越大，各数值之间的间距也越大。经验数据是否具有长尾，通常取决于两个因素：离群值和极端值。虽然二者都很少见、幅度极端，且在数值上无法区分，但在解释上存在细微却至关重要的差异。

*MSC2020* 学科分类：主要分类 2-02，研究性论述；次要分类 92C55。关键词与短语：对数、对数变换、长尾分布、几何平均数、图像分析。

离群值虽然并非所有人都认同这一术语的用法，但通常指不可能出现或由意外、噪声、测量误差等因素导致的数值。它们的存在源于对“真实”数据的观测受到破坏或噪声干扰；它们不会出现在理想域中，只会出现在含噪的观测域中。理想情况下，移除这类数值不会改变人们希望拟合“真实”数据所采用的模型。因此，移除它们是一项合理且往往必不可少的预处理步骤。如果观测数据的长尾完全是由离群值造成的，那么相比应用对数变换，移除离群值更可取，因为“真实”数据并无长尾，进行变换反而会不适当地改变其形态。感兴趣的读者可参阅 Zimek 等人的综述 [41]，其中全面概述了离群值检测。

另一方面，极端值非常重要，因此专门有一个数学领域 [31, 7] 致力于对极端值事件进行准确建模和预测。地震、股市崩盘、流行病等都是例子。与离群值不同，准确建模和预测极端值至关重要。例如，在地震仪记录中，地震发生的频率极低，但数值却极大。若将地震视为离群值而非极端值，继而将其从数据集中删除，就会严重限制对该数据集进行后续分析的价值。在这个例子中，与其删除数据，不如采用对数变换。

Xiao 等人 [40] 认为，是否使用对数变换应取决于似然检验：如果加性误差和服从正态分布的源数据（$X = N(\mu,\ \sigma) + \epsilon$）能更好地解释数据，那么最好不进行变换。然而，如果统计检验表明乘性误差和对数正态数据（$X = \exp(N(\mu,\ \sigma)) * \epsilon$）拟合得更好，那么采用对数变换，从而使用几何平均数，则更为合适。模型拟合数据后，所得系数用于对底层数据生成函数或总体得出结论。类似地，Feng 等人 [10] 报告了对数变换结果可能被误解的问题，并认为，只有在已知源数据取自对数正态分布时，才应进行变换。如果没有这类信息，他们主张采用广义估计方程等方法。然而，这些方法并非完全非参数；它们仍依赖于对测量数据前两个矩的估计：均值和标准差。我们将表明，依赖这一一阶矩可能导致悖论性结论，即对数悖论。本文将阐述导致这种看似混淆的模式反转的条件；这种模式反转是对长尾数据进行对数变换所引发的。我们还将介绍在若干条件下识别、正式描述和处理对数悖论的方法。

1.2. 定义。 更形式化地说，对实标量 $y$进行对数变换，是 Box-Cox [3]变换的一种特殊情况，给定参数 $\lambda$，其形式如下：

$$f(y) = \begin{cases} \frac{(y^\lambda - 1)}{-\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases},$$

其中 $f(y)$ 分别在取值为 $\lambda \in$[-1、1 和 1/2, 0] 时，返回 $y$ 的倒数、原值、平方根和对数。请注意，$\lambda$ 是指定变换方式的超参数，而不是操作数（y）。如果 $y = 0$，对数变换会导致未定义的值。请注意，在本文余下部分中，表达式 $Y = \log(X)$ 是指对向量 $X$：$Y = \log(X) = [\log(x_i) \quad \forall x_i \in X]$逐元素应用对数的简写。在本研究中，我们聚焦于对数变换，Keene [16] 将其视为常用变换中的一种“特殊”或偏好的变换。给出的一个理由是，对于正值数据，其取值范围（$-\infty, +\infty$）与其最常见的目标分布——正态分布——的取值范围相匹配。该变换可以通过线性偏移量进行参数化：

$$y(x) = \log(x - c)\ \forall x \in X, c \in \mathbb{R}^+, c < \min(X), \tag{1}$$

其中 $c$ 是 $x$ 的下限，用于调整较小数值的频率加权。或者，根据 $X$ 的分布设置对数的底数 $b$，也可以成为一种有用的特定领域选择。另一种变体是截断：

$$y = \log(\max(x, \epsilon)), \quad \epsilon > 0, \tag{2}$$

在计算中避免定义域错误，而无需移除使用 $\epsilon$ 接近机器精度的零值数据（$\sim 1.0e-12$）。当 $0 < x < 1$ 时，如贝叶斯函数中的概率所示，$y = -\log(x)$ 的变化并不少见。

应用对数变换并非没有风险。Feng 等人 [4] 详细说明了对数变换可能导致混杂结论的几种方式，并指出，对一系列数值进行压缩并不总能得到保证，且只有在变换后的定义域同样具有可解释性时，才应应用对数变换。一个对数变换后的定义域具有可解释性的例子，是使用分贝（dB）等尺度测量声音。相比之下，对数尺度的温度、速度或表面进行解释，可能会产生难以理解的单位，尤其是考虑到对数变换由适当选择的底数参数化这一事实 $b$，通常为 $e$ 或 10。我们将在第 $b$ 节进一步讨论B中的影响。

Feng 等人利用计算机模拟数据 [4] 展示了几种情形，在这些情形下，对数变换可能导致难以解读的发现。最近，Gao 等人 [13] 在细菌计数的背景下，报告了一种修正后的方法，用于定义和计算对数变换下的方差和不确定性。从这些研究中可以清楚看出，尽管对数变换十分常见且实用，但若缺乏更深入的理解，就应谨慎并由具备专业知识的人士加以应用。

1.3. 对数变换与几何平均数之间的联系。 常见的假设检验会比较样本或总体分布的期望值，以检验其是否相等或排序。对于离散数据，期望值通常以样本算术平均数近似。然而，若对数据进行对数变换，变换后数据的算术平均数就与几何平均数直接相关 [13]。令 $X \in \mathbb{R}^{|X|}, X = [x_i > 0 | \forall x_i \in \mathbb{R}], b > 1, N = |X|$:

$$\mathbb{E}[\log(X)] = \frac{\sum^{N} \log_b(x_i)}{N}, \tag{3}$$

$$\mu^*(X) = \sqrt[N]{\prod} x_i = \exp_b(\frac{\sum_{i=1}^{N} \log_b(x_i)}{N}). \tag{4}$$

符号 $\mu^*$ 指的是几何平均数，而不是算术平均数。几何平均数是一种非常强大的指标，通常用于以比率而非绝对标量值表示数量的领域，例如金融中的复利；不过，这并不妨碍将几何平均数用于绝对标量值。

在其开创性文章中，Fleming 等人 [12]指出，几何平均数可以避免混杂结论和陷阱，尤其是在比较计算系统的基准测试结果时，因为这类结果通常是标准化数据。这里的标准化，是指将数值数据表示为比率，或相对于共同尺度或参考分布来表示。利率或标准分数本身就是“标准化”数据的例子 $z = {}^{x-\mu(X)}\,\sigma(X)$。Fleming 等人 [12] 认为，当数据以比率形式表示、因而具有共同尺度时，与算术平均数相比，几何平均数是更合适的汇总指标。记录的标量值实际上隐含着比率这一点，并不总是显而易见。例如，测量细胞器在细胞中的大小（体积）时，这些大小的分布会受到每个细胞大小以及其他生物物理约束的限制。换句话说

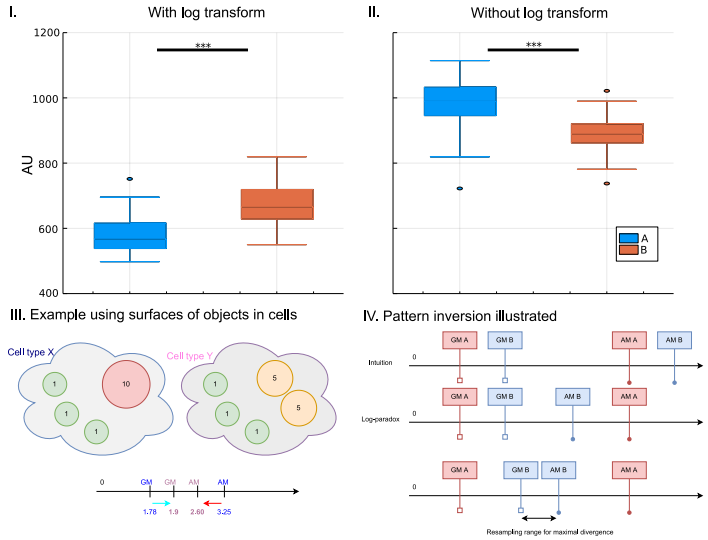
，体积存在一个隐含的上限 $M$，因此每个测得的细胞器实际上都是相对于该上限的比率。

最近，Vogel 等人 [37]针对一组常见分布推导了几何平均数的闭式表达式，并讨论了在若干常见长尾分布下使用几何平均数的影响。是否使用对数变换，并不总是显而易见。例如，Proietti 等人 [26]指出，在一项金融预测应用中，几何平均数有益，但其效果取决于预测的时间跨度。St-Pierre 等人 [33] 认为，尽管他们在现实世界数据集的比较中没有发现悖论式结果，但应用对数变换可能会导致模型系数大幅变化，从而造成解释困难。因此，他们主张改变模型而非改变数据，以获得更稳健的结论。迄今为止，相关工作将是否使用对数变换视为非此即彼的选择。然而，有些应用会同时使用变换后的数据和未变换的数据。例如，“信号平坦度”指标 [28] 就是一个例子：在噪声图像分割的背景下，该指标利用变换后的数据与未变换数据之比，将二者结合起来作为信息性指标。尽管这种做法可能很有力，但对于非专业终端用户而言，是否采用几何平均数，进而是否使用对数变换，并非简单的决定。

1.4. 聚合和重采样语境下的转换。在报告生物医学图像中捕获的结构的统计数据时，聚合（例如将体积聚合为平均体积）通常是必要的，以避免违反抽样独立性的前提（辛普森悖论 [38]）。考虑这样一种情况：我们有来自两组细胞的二维图像，这两组分别接受了“C”和“D”处理。每幅图像包含 $N$ 个细胞，而每个细胞包含数量不定的对象。我们希望检验这样一个假设：在给定处理条件下，二维图像中对象的表面积存在差异。假设实际的生物学动态支持该假设：$\mathbb{E}[\text{表面积}|\text{处理=C}] > \mathbb{E}[\text{表面积}|\text{处理=D}]$。辛普森指出，如果不先在细胞层面进行聚合，而是报告所有观测结果，所得分布可能会发生偏斜，分析结果也可能**反转**该假设。此外，观测数量的膨胀很容易将统计显著性值推过阈值。抽样独立性是假设检验的常见前提，但当人们将不同细胞中的对象混合在一起时，这一前提不一定成立。关于辛普森悖论的更深入介绍和讨论，请参阅 Wagner 等人 [38] 和 Spirtest 等人 [32] 。

一种纠正方法是在细胞层面进行汇总，例如估算均值，然后将汇总值的分布作为未受混杂影响的近似值报告。人们也常使用重采样来计算稳定或更具代表性的汇总值。其中一种方法是“自助法” [8], ；假设底层数据的方差有限 [1], , 可以证明，与未经重采样的统计量相比，自助法能更稳健地收敛到总体均值。重采样的必要性随之引出了一个问题：在重采样过程中，对数变换究竟会产生什么影响?

1.5. 贡献。 本文介绍并探讨了对数悖论，即两个向量在进行对数变换后，其均值出现悖论差异、模式反转或模式倒置：$\mathbb{E}[A] > \mathbb{E}[B]$ **和** $\mathbb{E}[\text{对数}(A)] < \mathbb{E}[\text{对数}(B)]$。为直观展示这一悖论，并凸显这项贡献的重要性，我们在图 1 中以图示说明这一概念。在面板I-II中，我们绘制了抽样指数分布均值的箱线图，展示了比较结果如何在进行对数变换后**发生反转**。面板III说明，即使样本量很小，也可能出现这种情况；在真实数据中，这可能源于分割错误或实际的长尾效应。面板IV对比了人们对均值预期排序的直觉与悖论出现时的实际情况。最后，面板V给出了一个最小数值示例，其中只需对一个向量稍作改动




V. 最小数值示例

| $F$ | $\triangle_Y F(X)$ | $\nabla_R F(X)$ | $\delta_{Q,Z} F(X)$ |
|---|---|---|---|
| $\mu^+$ | - 0.150 | + 0.083 | - 0.250 |
| $\mu^*$ | + 0.097 | - 0.295 | + 0.307 |

图 1：I–II：对数悖论示意图：A 和 B 是从两个不同的计算机模拟长尾指数分布中抽取的均值向量；在聚合之前进行对数变换，似乎会使“更大”模式反转。曼-惠特尼 U 检验表明，这些抽样均值始终会产生统计显著的反转模式。III：生物医学成像中测量对象大小的悖论最小示例。两个细胞各自包含一个物体分布，并以其均值进行量化。若在对数变换后报告结果，就会得出 $\mathbb{E}[\log(A)] < \mathbb{E}[\log(B)]$，的结论，然而算术平均数显示的趋势恰好相反$\mathbb{E}[A] > \mathbb{E}[B]$。实际上，潜在模式体现的是聚合动力学：较小的对象会自组织成较大的对象，或解聚成较小的对象。由于空间有限，且大小报告为二维或三维，这类数据很容易形成长尾分布。这一潜在特征足以导致对数悖论出现。IV：在正实数轴上对对数悖论的图示。对于选定的两个向量（A、B），模式反转可能会因对数变换而发生。V. 展示最小数值示例的表格。设 $X = [2, 4, 6, 13]$, $Y = [5.5], R = [6], Q = [3, 11], Z = [2, 13]$，且 $\mu^*(X) = 4.99, \mu^+(X) = 6.25$。各列分别表示串接、删除和替换时均值的变化。各行表示算术平均数和几何平均数。对于数字很少的每一列，我们可以选择 $Y, Z$，使几何平均数（$\mu^*$）与算术平均数（$\mu^+$）的变化符号相反 。

$X$，通过串接（$\triangle_Y$）、删除（$\nabla_Y$）或替换（$\delta_{Q,\ z}$），我们可以诱发均值的悖论性变化。本文其余部分结构如下：

• 我们引入算术均值和几何平均数的有限差分记号，以便在第2节中严格描述这一模式。

• 接下来，我们推导出充要条件，并提出一种使反转最大化的启发式方法。

• 我们将在第3 节使用计算机模拟数据展示反转何时以及如何发生，并在第3.2节具体探讨重采样下对数悖论的敏感性。

• 第3.3节将详细介绍一项受真实生物医学图像分析启发的计算机模拟实验，突出说明这种反转与该领域密切相关。

我们在第 4 节总结了在可能发生悖论性反转时如何开展数据分析的建议。

## 2. 方法：悖论性均值差异的解析表达式。

2.1. 预备知识与定义。 为避免歧义，我们首先严格定义本文所使用的数学术语。在本节末尾，我们在表 1 中列出所有术语，以便查阅。设 $G$ 是一个采样函数 $\mathbb{N}_{>0} \to \mathbb{R}$。该函数的参数是其样本的索引，例如该函数生成样本的第 i 次。此类函数的例子包括来自正态分布、泊松分布、指数分布或经验分布的随机样本。本文其余部分处理记录的实数样本$x$，这些样本存储在随机变量向量 $X$ 中：

$$X = [x_1, ..., x_i, ..., x_N] \in \mathbb{R}^N, x_i = G(i), x_i >= 0, \forall i \in \mathbb{N}_{>0}. \tag{5}$$

定义 2.1（向量变换函数）。在本文中，当一个值向量被变换时，将实值函数 $f$ 逐元素应用于向量 $X$，记作 $F(X)$ 或 $f.(X)$，这是向量化编程语言中的常见记法。令$f$： $\mathbb{R} \to \mathbb{R}$：

$$f.(X) = F(X) : \mathbb{R}^{|X|} \to \mathbb{R}^{|X|} = [f(x_1), ..., f(x_N)] \; \forall x_i \in X, \tag{6}$$

其中 $|X| =$ 维$(X)$，即向量 $X$ 中的元素个数。

例如，对数$_{10} \cdot([10, 100]) = [1, 2]$。

定义 2.2（向量拼接）。两个向量可以串接：

$$X \| Y : \mathbb{R}^{|X|} \times \mathbb{R}^{|Y|} \to \mathbb{R}^{|X|+|Y|} = [x_1, ..., x_N, y_1, ..., y_M] \;\; \forall n, m >= 1. \tag{7}$$

例如： $[1, 2, 3] \| [3, 4] = [1, 2, 3, 3, 4]$。

定义 2.3（向量差）。两个向量的差可以定义为串接的逆运算：

$$X | \setminus Y : \mathbb{R}^{|X|} \times \mathbb{R}^{|Y|} \subset X \to \mathbb{R}^{C=|X|-|Y|} = Z = [z_1, ..., z_C] | \forall z_i \in X \;\; \wedge Z \| Y = X. \tag{8}$$

例如： $[1, 2, 3, 3, 4] | \setminus [3, 4] = [1, 2, 3]$。然而，由于元素 4 并未出现在 $X$ 中，$[1, 2, 3] | \setminus [3, 4]$ 未定义。此处选择使用“向量”一词，是因为在解释上存在细微差别：“集合”、“序列”甚至“级数”有时也用于表示相同的概念，但它们的语义各不相同。集合可能意味着给定的标量元素基数为 1，但此处一个标量元素可以被抽样多次。序列意味着存在排序，但我们所应用的运算与排序无关。因此，在本研究中，用“向量”表示从生成函数中抽样得到的一组数字最为准确。

定义 2.4（聚合函数）。我们将聚合函数定义为

$$f : \mathbb{R}^N \to \mathbb{R} \;\; \forall N >= 1. \tag{9}$$

一个简单的例子是向量的求和： $f(X) = \sum x. x \in X$

定义 2.5（算术平均数）。这恰好引出了算术平均数：

$$\mu^{+}(X):\mathbb{R}^{|X|}\to\mathbb{R}:\frac{\sum X}{|X|}. \tag{10}$$

$\mu$ 符号表示“均值”， + 上标用于指代算术平均数，以区别于几何平均数。

定义 2.6（几何平均数）。

$$\mu^{*}(A):\mathbb{R}^{|X|}\to\mathbb{R}:\sqrt[|X|]{\prod_{x_i\in X}x_i}\quad\forall x_i>0. \tag{11}$$

出于数值精度、溢出和性能方面的考虑，通常通过计算对数变换后的向量的算术平均数来求几何平均数：

$$\mu^{*}(A)=\exp_b\log_b\sqrt[|A|]{\prod_{a_i\in A}A}=\exp_b\frac{\sum\log_b.(A)}{|A|}=\exp_b\mu^{+}(\log_b.(A)), \tag{12}$$

。底数的选择 $b$ 对分布子集重新加权的影响往往被低估，但实际上可能相当显著。我们将在第B节中进一步讨论。

注意，如果向量中至少有一个元素为零，其几何平均数就会变为零；而在最常见的实现方式中，使用对数实际上会导致定义域错误。本文余下部分仅考虑严格正向量$X=[x_1,...,x_N]\ \ x_i\in\mathbb{R}_{>0}\forall x_i$。对于含有零项或负值的数据，可以使用若干变换（式1、2）来避免定义域错误，但需要谨慎考虑，因为变换后的定义域会变得更加难以解释。

事实（算术-几何平均不等式）。几何平均数和算术平均数分别构成对方的下界和上界，这由“算术-几何平均不等式”给出，它是柯西-施瓦茨不等式的一种特殊情况。也就是说，对于向量 $X\in\mathbb{R}^{N}_{>0}$：

$$\mu^{*}(X)\leq\mu^{+}(X). \tag{13}$$

定义 2.7（均值间距离（ID））。我们将均值间距离定义为两个均值之差的绝对值：

$$ID(X):\mathbb{R}^{|X|}\to\mathbb{R}_{>0}:\text{abs}(\mu^{+}(X)-\mu^{*}(X))=\text{abs}(\mu^{*}(X)-\mu^{+}(X)). \tag{14}$$

根据算术-几何平均不等式（式 13），我们可以将其简化为：

$$ID(X):\mathbb{R}^{|X|}\to\mathbb{R}_{>0}:\mu^{+}(X)-\mu^{*}(X). \tag{15}$$

定义 2.8（悖论式比较）。最后，我们将变换 $f$ 下两个向量 $A$、$B$ 的悖论式比较定义为：

$$\text{Paradoxical comparison}(A,B,f)=\mathbb{E}[A]>\mathbb{E}[B]\wedge\mathbb{E}[f.(A)]<\mathbb{E}[f.(A)]. \tag{16}$$

本文聚焦于向量的悖论式比较，其中 $f$ =对数。

2.2. 使用有限差分算子对均值变化建模。 本文研究两个向量的均值如何以及何时会 $X$、 $Y$ 产生悖论式差异。为了对此进行解析描述，我们需要列举将 $X$ 变换为 $Y$的各种运算，然后描述均值如何随一系列此类运算而变化。对于标量实值函数，有限差分算子（例如前向、后向和中心差分） [39] 用于近似类似的差异。对于标量实值函数 $f$： $\mathbb{R} \to \mathbb{R}$ 和 $h \in \mathbb{R}$，有限差分算子定义如下：

定义 2.9（前向差分）。

$$\triangle_h f(x) = f(x+h) - f(x). \tag{17}$$

定义 2.10（后向差分）。

$$\nabla_h f(x) = f(x) - f(x-h). \tag{18}$$

定义 2.11（替换差）。

$$\delta_h f(x) = f(x-h) - f(x+h). \tag{19}$$

对于我们的使用场景，我们关注的是数据聚合后（即用单个标量表示后）的差异，例如均值。设 $F$ 为一个聚合函数，使得$F$： $\mathbb{R}^{\mathbb{N}} \to \mathbb{R}$，其输入向量为 $X$。当使用运算符 $G$： $\mathbb{R}^N \times \mathbb{R}^M$： $\to \mathbb{R}^K$及参数 $X$，$Y$对 $X$ 进行修改，得到向量 $X'$ 时，我们希望量化 $F$（$X'$）与$F$（$X$）的比较结果：

$$F : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}, D_{(G,Y)} F(X) = F(G(X,Y)) - F(X). \tag{20}$$

我们关注三种运算：添加元素（串接）、移除元素（删除）或替换元素（替换）。

2.2.1. 串接下的有限差分。

定义 2.12（串接下的有限差分）。

$$D_{\|,Y} F(X) = \triangle_Y F(X) = F(X \| Y) - F(X). \tag{21}$$

定义 2.13（删除下的有限差分）。

$$D_{|\setminus Y} F(X) = \nabla_Y F(X | \setminus Y) - F(X). \tag{22}$$

定义 2.14（替换下的有限差分）。

$$D_{\|Y|\setminus Z} F(X) = \delta_{Y,Z} F(X) = F((X \| Y) | \setminus Z)) - F(X). \tag{23}$$

在继续推导算术平均数和几何平均数的有限差分表达式之前，我们先在表中总结迄今定义的术语 1。

2.2.2. 算术平均数的有限差分。设 $X \in \mathbb{R}^N$，$Y \in \mathbb{R}^M$，$Z \in \mathbb{R}^K$，$|X| = N$，$|Y| = M$，$|Z| = K$，其中 $N$，$M$，$K \in \mathbb{N}_{>0}$，且 $K < N, M < N$。

定理 2.15（拼接下算术平均数的有限差分）。

$$\triangle_Y \mu^+(X) = \frac{M}{N+M}(\mu^+(Y) - \mu^+(X)). \tag{24}$$

证明 定理 2.15。

$$\begin{aligned}
\triangle_Y \mu^+(X) &= \mu^+(X\|Y) - \mu^+(X) \\
&= \frac{\sum X + \sum Y}{N+M} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{N}{N+M} + \frac{\sum Y}{M}\frac{M}{N+M} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{-M}{N+M} + \frac{\sum Y}{M}\frac{M}{N+M} \\
&= \mu^+(X)\frac{-M}{N+M} + \mu^+(Y)\frac{M}{N+M} \\
&= \frac{M}{N+M}(\mu^+(Y) - \mu^+(X)).
\end{aligned} \tag{25}$$

□

直接推论如下：

$$\triangle_Y \mu^+(X) < 0 \Leftrightarrow \mu^+(Y) < \mu^+(X). \tag{26}$$

定理 2.16（删除下的有限差分）。

$$\nabla_Y \mu^+(X) = \frac{M}{N-M}(\mu^+(X) - \mu^+(Y)). \tag{27}$$

P定理证明 2.16。

$$\begin{aligned}
\nabla_Y \mu^+(X) &= \mu^+(X|\setminus Y) - \mu^+(X) \\
&= \frac{\sum X - \sum Y}{N-M} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{N}{N-M} - \frac{\sum Y}{M}\frac{M}{N-M} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{M}{N-M} - \frac{\sum Y}{M}\frac{M}{N-M} \\
&= \mu^+(X)\frac{M}{N-M} - \mu^+(Y)\frac{M}{N-M} \\
&= \frac{M}{N-M}(\mu^+(X) - \mu^+(Y))
\end{aligned} \tag{28}$$

□

直接推论如下：

$$\nabla_Y \mu^+(X) < 0 \Leftrightarrow \mu^+(X) < \mu^+(Y). \tag{29}$$

这两个推论符合这样的直觉：如果向 $Y$中加入均值（$\mu^+(Y)$）小于均值（$\mu^+(X)$，拼接差）的数值，或者删除均值（$\mu^+(Y)$）大于均值（$\mu^+(X)$，删除差）的数值，则均值会下降。均值变化的幅度取决于相对于原始数据大小（$N = |X|$）而言所添加或删除的数据比例（$M = |Y|$）。另一个推论是，在添加或删除元素时计算均值，只需要知道 $X$ 和 $Y$ 的均值，而不需要知道 $X$ 和 $Y$ 中的各个元素。

定理 2.17（替换下算术平均值的有限差分）。

$$\delta_{Y,Z}\mu^+(X) = \mu^+(X)\frac{-(M-K)}{N+M-K} + \mu^+(Y)\frac{M}{N+M-K} - \mu^+(Z)\frac{K}{N+M-K}. \tag{30}$$

当我们替换 $X$ 中的元素（例如 $M = K$）时，该表达式可简化为：

$$\delta_{Y,Z}\mu^+(X) = \frac{M}{N}(\mu^+(Y) - \mu^+(Z)). \tag{31}$$

定理证明 2.17。

$$\begin{aligned}
\delta_{Y,Z}\mu^+(X) &= \mu^+((X\|Y)| \setminus Z) - \mu^+(X) \\
&= \frac{\sum X + \sum Y - \sum Z}{N+M-K} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{N}{N+M-K} + \frac{\sum Y}{M}\frac{M}{N+M-K} - \frac{\sum Z}{K}\frac{K}{N+M-K} - \frac{\sum X}{N} \\
&= \frac{\sum X}{N}\frac{N-(N+M-K)}{N+M-K} + \frac{\sum Y}{M}\frac{M}{N+M-K} - \frac{\sum Z}{K}\frac{K}{N+M-K} \\
&= \mu^+(X)\frac{-(M-K)}{N+M-K} + \mu^+(Y)\frac{M}{N+M-K} - \mu^+(Z)\frac{K}{N+M-K}. \qquad (32)
\end{aligned}$$

□

直接推论如下：

$$\delta_{Y,Z}\mu^+(X) < 0 \Leftrightarrow \mu^+(Y) < \mu^+(Z). \tag{33}$$

注意到该表达式中没有 $\mu^+(X)$，变化只取决于 $Y$ 和 $Z$，尽管变化的幅度仍与（$M/N$）成正比。替换差与“滑动”“运行”或“窗口化”均值有关，这些均值通常用于估计不同时间段内的趋势，例如流行病期间的感染病例、机场到达人数等。$\delta_{Y,\ Z}\mu^+(X)$ 表示窗口大小为 $N$ 、步长为 $K$ 时滑动、运行或窗口化均值的变化。例如，使用每日记录计算每周运行平均值：$|X|$=7、$K$=1。快速传播的流行病可能导致在对数尺度上报告运行平均值，在这种情况下，悖论可能会被诱发。

2.2.3. 几何平均数的有限差分。

定理 2.18（串接下几何平均值的有限差分）。

$$\triangle_Y \mu^*(X) = \mu^*(X)(\mu^*(X)^{\frac{-M}{N+M}}\mu^*(Y)^{\frac{M}{N+M}} - 1). \tag{34}$$

因此：

$$\begin{aligned}
&\triangle_Y \mu^*(X) > 0 \Leftrightarrow \\
&\mu^*(X)^{\frac{-M}{N+M}} \mu^*(Y)^{\frac{M}{N+M}} - 1 > 0 \Leftrightarrow \\
&\mu^*(Y)^{\frac{M}{N+M}} > \mu^*(X)^{\frac{M}{N+M}} \Leftrightarrow \\
&\mu^*(Y) > \mu^*(X). \qquad (35)
\end{aligned}$$

定理 2.18的证明。

$$\begin{aligned}
\triangle_Y \mu^*(X) &= \mu^*(X\|Y) - \mu^*(X) \\
&= \sqrt[N+M]{\prod X \prod Y} - \sqrt[N]{\prod X} \\
&= (\sqrt[N]{\prod X})^{\frac{N}{N+M}} (\sqrt[M]{\prod Y})^{\frac{M}{N+M}} - \sqrt[N]{\prod X} \\
&= (\mu^*(X))^{\frac{N}{N+M}} (\mu^*(Y))^{\frac{M}{N+M}} - \mu^*(X) \\
&= \mu^*(X)(\mu^*(X)^{\frac{-M}{N+M}} \mu^*(Y)^{\frac{M}{N+M}} - 1). \qquad (36)
\end{aligned}$$

□

换句话说，当且仅当新增元素的“几何平均数”较大时，串接后几何平均数才会增加。注意，我们最初指定 $\mu^*(X) > 0$，因此等式两边仍然可以相除。

定理 2.19（删除时几何平均数的有限差分）。删除差分定义如下：

$$\nabla_Y \mu^*(X) = \mu^*(X)^{\frac{N}{N-M}} \mu^*(Y)^{\frac{-M}{N-M}} - \mu^*(X). \qquad (37)$$

因此：

$$\begin{aligned}
&\nabla_Y \mu^*(X) > 0 \Leftrightarrow \\
&\mu^*(X)^{\frac{N}{N-M}} \mu^*(Y)^{\frac{-M}{N-M}} - \mu^*(X)) > 0 \\
&\mu^*(Y)^{\frac{-M}{N-M}} > \mu^*(X)\mu^*(X)^{\frac{-N}{N-M}} \\
&\mu^*(Y)^{\frac{-M}{N-M}} > \mu^*(X)^{\frac{-M}{N-M}} \\
&\mu^*(Y) < \mu^*(X). \qquad (38)
\end{aligned}$$

换句话说，移除均值小于均值（$\mu^*(X)$）的值（$Y$），将会使均值增加（$\nabla_Y \mu^*(X) > 0$）。

定理的证明 2.19。

$$\begin{aligned}
\nabla_Y \mu^*(X) &= \mu^*(X|\setminus Y) - \mu^*(X) \\
&= \sqrt[N-M]{\frac{\prod X}{\prod Y}} - \sqrt[N]{\prod X} \\
&= \frac{\mu^*(X)^{\frac{N}{N-M}}}{\mu^*(Y)^{\frac{M}{N-M}}} - \mu^*(X)
\end{aligned}$$

$$= \mu^*(X)^{\frac{N}{N-M}} \mu^*(Y)^{\frac{-M}{N-M}} - \mu^*(X). \tag{39}$$

□

定理 2.20（替换下几何均值的有限差分）。我们仅考虑 $X'$ 的大小保持不变的情况：$|Y| = |Z| = M$。

$$\delta_{Y,Z}\mu^*(X) = \mu^*(X)(\frac{\mu^*(Y)^{M/N}}{\mu^*(Z)^{M/N}} - 1). \tag{40}$$

因此：

$$\delta_{Y,Z}\mu^*(X) > 0 \Leftrightarrow \frac{\mu^*(Y)}{\mu^*(Z)} > 1. \tag{41}$$

定理证明 2.20。

$$\begin{aligned}
\delta_{Y,Z}\mu^*(X) &= \mu^*((X\|Y)|\setminus Z) - \mu^*(X) \\
&= \sqrt[N]{\prod X * \prod Y * (\prod Z)^{-1}} - \sqrt[N]{\prod X} \\
&= \sqrt[N]{\prod X} \sqrt[N]{\frac{\prod Y}{\prod Z}} - \sqrt[N]{\prod X} \\
&= \sqrt[N]{\prod X}(\sqrt[N]{\frac{\prod Y}{\prod Z}} - 1) \\
&= \mu^*(X)(\frac{\mu^*(Y)^{M/N}}{\mu^*(Z)^{M/N}} - 1).
\end{aligned} \tag{42}$$

□

2.2.4. 均值间距离的有限差分。有限差分算子作用于 $X$时，例如进行串接、删除、替换（方程21、22、23），均值间距离（均值间距离）会发生变化（方程 15）。通过重新排列各项，很容易用均值的有限差分表示这些均值间距离的差分：

$$\begin{aligned}
\triangle_Y ID(X) &= (\mu^+(X\|Y) - \mu^*(X\|Y)) - (\mu^+(X) - \mu^*(X)) \\
&= \triangle_Y \mu^+(X) - \triangle_Y \mu^*(X).
\end{aligned} \tag{43}$$

$$\begin{aligned}
\nabla_Y ID(X) &= (\mu^+(X|\setminus Y) - \mu^*(X|\setminus Y)) - (\mu^+(X) - \mu^*(X)) \\
&= \nabla_Y \mu^+(X) - \nabla_Y \mu^*(X).
\end{aligned} \tag{44}$$

$$\begin{aligned}
\delta_{Y,Z} ID(X) &= (\mu^+((X\|Y)|\setminus Z) - \mu^*((X\|Y)|\setminus Z)) - (\mu^+(X) - \mu^*(X)) \\
&= \delta_{Y,Z}\mu^+(X) - \delta_{Y,Z}\mu^*(X).
\end{aligned} \tag{45}$$

表 3 和图 2 总结了均值发生悖论性变化的必要条件。这些条件也可以用均值间距离来表示，例如，

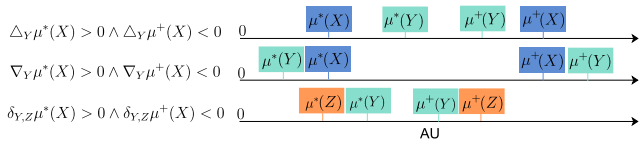



图 2：左侧的表达式列出了通过扰动论证改变 $X$ 和 $Y$， $Z$ 以引发悖论所需满足的三个条件。右侧相应的图示展示了 $X$、 $Y$和 $Z$ 的几何平均数与算术平均数在数轴上的相对排序，数轴单位为任意单位（AU）。要满足这些条件，必须注意替换差 $\delta$ 的表达式仅取决于 $Y$， $Z$而不取决于原向量 $X$。请回想，对于任意向量 $A$： $\mu^*(A) \leq \mu^+(A)$。这些结果由表 2、3以及算术-几何平均不等式（公式 13）结合得出。所有符号见表 1。

替换差的条件确实会导致 $ID[Y] < ID[Z]$。不过请注意，均值间距离是必要条件，但并非充分条件；显然可以找到 $Y$， $Z$ 使得 $ID[Y] < ID[Z]$，但仍不会引发悖论差异。均值间距离并不限定 Y 和 Z 的均值的相对排序。例如，在 $\mu^*(Y) < \mu^+(Y) < \mu^*(Z) < \mu^+(Z)$ 且 $ID[Y] < ID[Z]$ 的情况下，该条件不成立。

*2.2.4.1.* 一个示例。我们将通过一个非常简单的例子（见表 1-V），说明如何使用有限差分算子来计算算术平均数和几何平均数的变化。令 $X = [2, 4, 6, 13]$， $Y = [5.5]$， $R = [6]$， $Q = [3, 11]$， $Z = [2, 13]$，其中 $\mu^*(X)$ =4。99， $\mu^+(X)$ = 6.25。该例遵循算法 1、2中的步骤，选择均值间区间内的元素或最小值/最大值这一对元素；结论可直接从表 3、图 2的观察结果得出。这个最简示例清楚地展示了在三种操作——串接、差分和替换——下，如何轻易诱发符号相反的差值。

2.3. 最大化均值悖论性差异。 至此，我们已推导出两个向量之间何时会出现如公式16所定义的悖论性比较（表3）。这里我们将考察悖论差异的大小。该差异的大小直接决定了悖论对常见统计检验的混淆程度。给定两个实值向量 $X$ 和 $X'$，比较它们的两种常用方法是假设检验和效应量。常用的假设检验涉及一个统计显著性阈值，该阈值直接受 $\mathbb{E}[X] - \mathbb{E}[X']$之差的影响。同样，效应量 [9] 通常根据这一差异并以标准差为尺度计算。回想一下，给定变换函数 $f$，两个向量 $X$ 和 $X'$之间的悖论性差异为：

$$\mathbb{E}[X] > \mathbb{E}[X'] \quad \wedge \quad \mathbb{E}[f(X)] < \mathbb{E}[f(X')]. \tag{46}$$

或者等价地：

$$\mathbb{E}['X] - \mathbb{E}[X] < 0 \quad \wedge \quad \mathbb{E}[f.(X')] - \mathbb{E}[f.(X)] > 0. \tag{47}$$

我们研究 $f$ 为对数变换的情形，此时条件变为：

$$D_A(X, X') = \mu^+(X') - \mu^+(X) < 0 \quad \wedge \quad D_G(X, X') = \mu^*(X') - \mu^*(X) > 0. \tag{48}$$

换言之，$D_G$ （几何平均数随 $X$ 变为 $X'$ 而增加）和 $D_A$ （算术均值随 $X$ 变为 $X'$ 而增加）的符号相反。一个简单的判据是

使符号相反的差值 $D_A$ 和 $D_G$ 最大化的条件是：

$$\underset{X'}{\operatorname{argmax}} \; -(D_A(X,X') \times D_G(X,X')) = -(\mu^+(X') - \mu^+(X)) \times (\mu^*(X') - \mu^*(X)). \tag{49}$$

注意，使用加法（$-D_A + D_G$）来构造优化准则会产生多个次优解。一个显而易见的次优解是 $D_G = \mu^+(X) - \mu^*(X)$ 和 $-D_A = 0$。在此只有几何平均数发生变化，因此算术平均数不存在模式倒置，也不会产生悖论式比较（方程16）。乘法形式确保两个项同时达到最大值。

定理 2.21. 假设我们在选择 $X'$ 时拥有完全自由度，那么对于任意 $X$，使方程 *49*达到最大值的表达式为：

$$\mu^*(X') = \mu^+(X') = \frac{\mu^*(X) + \mu^+(X)}{2} = Q, \tag{50}$$

因此，

$$D_G(X,X') = Q - \mu^*(X) = \frac{\mu^+(X) - \mu^*(X)}{2} \tag{51}$$

$$D_A(X,X') = Q - \mu^+(X) = \frac{\mu^*(X) - \mu^+(X)}{2} = -D_G(X,X'). \tag{52}$$

由此，$D_A = -D_G$ 确保悖论差异达到最大值。

证明分为两部分：首先给出一个关于优化准则取值的引理，随后陈述实际的最优解。

引理 2.22. 令 $\forall x,\ y,\ k \in \mathbb{R} \setminus \{-\infty, +\infty\}$：$y < 0 < x,\ k > 0 \wedge x + (-y) = k$，我们提出：

$$\underset{x,y}{\operatorname{argmax}} \; x * (-y) = \{(x,y) \mid (x = -y) \wedge (x,y) = \underset{x,y}{\operatorname{argmax}} \; \mathit{abs}(x * y)\}. \tag{53}$$

方程 *53* 是方程 *49* 的广义表示，其中 $x = D_G(X,X')$、$y = D_A(X,X')$和 $k = ID(X)$。

引理 2.22 的证明。使用反证法，假设方程 53 不成立，并且假定的 $x = -y$解不是最优解，则 $\forall x$、$y$、$x' \in \mathbb{R} \setminus \{\infty\}$ 满足 $k = x - y$，$\exists\, y' \in \mathbb{R}$：

$$\begin{aligned} & y' < y = -x < 0 < x < -y' \\ & \wedge (x' * (-y') > x * (-y)) \\ & \wedge \ \ k = x' - y' \end{aligned} \tag{54}$$

这意味着存在 $\exists\, \epsilon > 0$，使得：$y' = y - \epsilon \wedge x' = x - \epsilon$。为简洁起见，令 $ab = a \times b$。

$$\begin{aligned} (-y')x' &> (-y)x \\ (y-\epsilon)(x-\epsilon) &< yx \\ yx - \epsilon x - \epsilon y + \epsilon^2 &< yx \\ -\epsilon x - \epsilon y + \epsilon^2 &< 0 \\ -x + \epsilon &< y \\ x - \epsilon &> -y \end{aligned} \tag{55}$$

这与给定的 $x = -y$相矛盾。因此，公式 49 中的判据在 $x = D_G(X,\ X') = -D_A(X,\ X') = y$ 时具有一个最小解。 □

定理 2.21 的证明（$Q$ 为最优值）。已知 $Q$ 是区间 $[\mu^*(X)\ \mu*(X)]$的中点， $Q$ 是唯一使 $D_G(X,\ X')\ = -D_A(X,\ X')$的值，并且在不降低另一个值的情况下，二者都无法在绝对值上增大。假设 $\exists\ J \neq X'$具有最优解 $Q' = D_G(X,\ J)\ = -D_A(X,\ J)$，满足公式 49： $Q' > Q$。于是有$D_G(X,\ X') < D_G(X,\ J)$。换言之，假设我们可以增大几何平均数：

$$\mu^*(X) < Q = \mu^*(X') < Q' = \mu^*(J). \tag{56}$$

算术-几何平均不等式（公式 13）以及所述 $J$ 和 $Q'$ 对于公式49的最优性，进而意味着：

$$\mu^*(X) < \mu^*(X') = \mu^+(X') = Q < \mu^*(J) = \mu^+(J) = Q' < \mu^+(X). \tag{57}$$

因此

$$|D_A(X,J)| < |D_A(X,X')|, \tag{58}$$

但现在

$$|D_G(X,J)|! = |D_A(X,J)|. \tag{59}$$

这就产生了矛盾， $J$和 $Q'$ 不再是最优的，从而证明 Q 是最优的。按照同样的思路，可以对 $Q' < Q$ 作出对称论证。满足这些条件的 $X'$ 的一个简单示例是 $X' = [Q,Q,Q,...]$。 □

2.3.0.1. 近似。$Q$ 是最优解，当 $X$ 和 $X'$ 可以在任意数量的元素上存在差异时。然而，如果只能有少数元素或一个元素存在差异，那么 $Q$ 只能是一个近似解，正如反例图 3所示。在左图（A）中，对于$X =$ [exp(1)、exp(3)], , 我们可以自由选择 $X'$，并轻松收敛到 $Q$处的最优解。然而，在右图（B）中， $X =$ [exp(1)、exp(3)], , 但 $X' = X||[x] =$[exp(1) $x$、exp(3)], $Q$ 虽然接近但并非最优。实际最优值由青绿色线表示。只能添加 1 个元素这一约束，使得梯度 $\triangle_x\mu^*$（$X$）（蓝色）和 $\triangle_x\mu^+$（$X$）（橙色）更多地受到现有元素的限制。图 3-A、B 以图形方式说明，悖论差异受到区间 $[\mu^*$（$X$） $\mu^+$（$X$）]的约束。

如果希望以对数或指数尺度表示差异，可能会得到不同的最优点。在上述条件下，我们可以利用有限差分算子推导出解析表达式，以确定如何求得 $X'$。鉴于比较的通常是大小相同的样本，我们针对常见情形——替换差——进行推导。一个直接而优雅的结果是均值间距离的微分方程（$ID$）：

$$\delta_{Y,Z}ID(X) = ID[X] - ID[(X||Y)| \setminus Z] = ID[X] - 0 = ID[X]. \tag{60}$$

串接差和删除差的表达式也可以用同样的方法推导。

2.3.1. 悖论式替换差。 结合公式50 和表3可得：

$$\begin{aligned}\delta_{Y,Z}\mu^+(X) &= \frac{\mu^*(X) - \mu^+(X)}{2} \\ &= \delta_{Y,Z}\mu^*(X).\end{aligned} \tag{61}$$

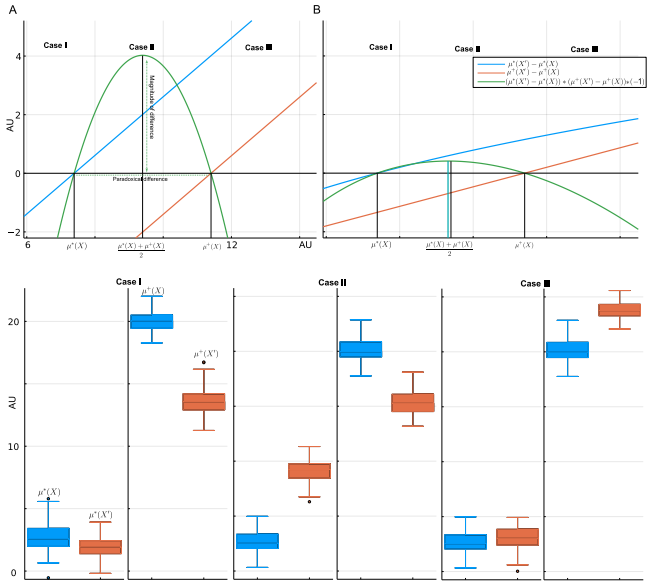



图 3：给定两个向量（$X$、$X'$），比较它们经过和未经变换后的均值时，可能出现 3 种情况。$X$ 和 $X'$ 的差异在于 X 轴上的元素。情况 I–III 对应区间 $(0, \mu^*(X)], (\mu^*(X)\mu^+(X))$, 和 $[\mu^+(X)$ $+\infty)$。在情况 I 和 III 中，$X$和 $X'$ 的均值要么都较小（I），要么都较大（III）。然而，也可能出现悖论差异（情况 II），此时排序取决于变换（对数）。当发生这种悖论性反转时，A-B 面板中的绿色曲线（两个均值梯度的乘积）为负。我们说明，如果可以完全自由地选择 $X'$，那么当 $^{\mu^*}(x)_{+\mu^+}(x)$ 2 时，悖论性反转（公式 16）达到最大值（A）。然而，当 $X' = X \| [x]$时，这个最优值只能被逼近，而无法达到（B）。数值以任意单位（AU）标度绘制。

这可简化为

$$\begin{aligned} \delta_{Y,Z}\mu^+(X) &= \frac{\mu^*(X) - \mu^+(X)}{2} \\ \frac{M}{N}(\mu^+(Y) - \mu^+(Z)) &= \frac{\mu^*(X) - \mu^+(X)}{2} \\ \mu^+(Y) - \mu^+(Z) &= \frac{N(\mu^*(X) - \mu^+(X))}{2M} \end{aligned} \tag{62}$$

对于算术平均数，以及

$$\begin{aligned} \delta_{Y,Z}\mu^*(X) &= \frac{\mu^+(X) - \mu^*(X)}{2} \\ \mu^*(X)(\frac{\mu^*(Y)}{\mu^*(Z)}^{\frac{M}{N}} - 1) &= \frac{\mu^+(X) - \mu^*(X)}{2} \\ 2(\frac{\mu^*(Y)}{\mu^*(Z)}^{\frac{M}{N}} - 1) &= \frac{\mu^+(X)}{\mu^*(X)} - 1 \\ \frac{\mu^*(Y)}{\mu^*(Z)} &= (\frac{\mu^+(X)}{2\mu^*(X)} + 1/2)^{\frac{N}{M}} \end{aligned} \tag{63}$$

对于几何平均数，其中 $|Y| = M = |Z|$。方程 63 展示了其非线性特征，与图 3-B 相符；相比之下，方程 62呈现线性行为。“平坦度” [28] 度量（$\mu^*(X)/\mu^+(X)$）的倒数也出现在方程 63中，体现了其信息价值。

2.3.2. 最大化悖论性散度的启发式方法。上述方程使我们能够求出可获得最大悖论性 $X'$的精确最优变换（如果存在）。然而，这类变换可能不止一种，或者实际上并不存在精确变换。我们可以定义一些仍能近似这一结果的启发式方法；正如下文将要展示的，这些方法能够帮助我们理解会引发该悖论的真实数据模式。首先请注意，向 $X$ 添加元素 $Y$，且从均值区间（$\mu^*(X)$ $\mu^+(X)$）中选取 $Y$，根据公式 43,26,34，将直接导致几何平均数增大、算术平均数减小。因此，$ID[X\|Y] < ID[X]$。反过来，从均值区间之外的 $X$ 中移除元素可以增大 $ID[X']$，其中最显著的变化由最小值(X)和最大值(X)引起。这是公式 39 和 27 的推论，其中 $|Y| = 1$:

$$\underset{[x]}{\operatorname{argmin}} \nabla_{[x]}\mu + (X) = \min(X) \tag{64}$$

$$\underset{[x]}{\operatorname{argmin}} \nabla_{[x]}\mu * (X) = \min(X) \tag{65}$$

$$\underset{[x]}{\operatorname{argmax}} \nabla_{[x]}\mu * (X) = \max(X) \tag{66}$$

$$\underset{[x]}{\operatorname{argmax}} \nabla_{[x]}\mu + (X) = \max(X) \tag{67}$$

证明可直接由以下显而易见的观察得出：$[x]$ 的几何均值和算术均值都等于 min([x]) 和 max([x])。据此，现在假设 $\exists\ x' \neq \min(X)$ 是公式 65 的一个最小解。为了满足公式， 39 $x'$ 必须使得 $\mu^*([x']) < \mu^*([\min(X)])$，这意味着 $x' < \min(X)$，而这不可能成立，从而证明了公式 65。其他结论可用对称论证证明。我们将替换的启发式方法定义为添加 $Y = [q, q]$，以替换 $Z = [m, M]$，其中 $q = \frac{\mu^+(X)+\mu^*(X)}{2}$，$m = \min(X)$ $M = \max(X)$。为了在替换过程中保持元素数量不变，我们使用两次 $q$。回想公式 33、41，若在 $X$ 中用 $Y$ 替换元素，则会产生悖论差异，$Z$ 如果

$$\begin{aligned} \delta_{Y,Z}\mu^*(X) &> 0 \\ \frac{\mu^*(Y)}{\mu^*(Z)} &> 1 \\ \mu^*(Y) &> \mu^*(Z) \end{aligned} \tag{68}$$

and

$$\begin{aligned} \delta_{Y,Z}\mu^+(X) &< 0 \\ \mu^+(Y) &< \mu^+(Z). \end{aligned} \tag{69}$$

$Z = [m, M]$ 的几何平均数和算术平均数分别为 $\sqrt{m * M}$和 $\frac{m+M}{2}$，因此我们得到

$$\sqrt{m * M} < \frac{\mu^*(X) + \mu^+(X)}{2} < \frac{m + M}{2}. \tag{70}$$

方程 70 确定了启发式方法何时会导致悖论差异。令 $Q = {}^{\mu^*(X)+\mu^+(X)}\,{}_2$，代入表 2中的结果后，启发式方法将在以下条件下导致最大的悖论差异：

$$\delta_{[Q,Q],[m,M]}\mu^+(X) = Q - \mu^+(X)$$
$$\frac{2}{N}(Q - \frac{m+M}{2}) = Q - \mu^+(X) \tag{71}$$

and

$$\delta_{[Q,Q],[m,M]}\mu^*(X) = Q - \mu^*(X)$$
$$\mu^*(X)(\frac{Q}{\sqrt{m*M}}^{\frac{2}{N}} - 1) = Q - \mu^*(X)$$
$$(\frac{Q}{\sqrt{m*M}}^{\frac{2}{N}} - 1) = \frac{Q}{\mu^*(X)} - 1. \tag{72}$$

对于任意实数集合 $|X| \geq 1$，$m$ 和 $M$ 始终存在，$Q$也始终存在；我们将其插入两次，以使 $X'$ 的大小与 $X$相同。为了评估启发式方法的最小条件（方程70），现在定义：

$$d(m, M) = \frac{m+M}{2} - \sqrt{m*M}. \tag{73}$$

如图 4所示，我们注意到，$d(m,\ M)$的值越大，$m$越小或 $M$越大，就越可能导致悖论式比较（方程 16）。这对从业者而言十分重要，因为移除离群值、存在极端值、重新抽样数据，甚至对象分割错误地将对象合并或拆分，都可能导致 $m$ 减小或 $M$ 增大，进而使均值发生意料之外的显著变化，并有引发悖论的风险。长尾分布通常是不对称的，但图 4的结果应能让从业者确信，即便在对称分布中也可能出现这一悖论。不过，这一悖论也可能源于数据的细微变化，而这些变化可能对应着有价值的领域特定发现，正如我们将在第 3节中探讨的那样。

**3. 实证结果。** 我们首先在第 3.1 节中使用指数分布构造了一个模拟现实世界数据的示例。鉴于重采样广泛用于图像数据量化，接下来我们将在第 3.2 节中回顾悖论在重采样下的表现。最后，我们以一个受细胞生物学启发的应用案例作结，说明即使是由马尔可夫链建模的蛋白质结构频率发生微小变化，也可能诱发该悖论。

**3.1. 经验分布中悖论的普遍性。** 现在我们着手回答实践者的问题：“在现实数据上，悖论会多快出现，其统计显著性又能达到多大？”我们创建了一个由指数分布产生的简单数值向量 $A$，该向量从经过缩放的概率密度函数（pdf）中采样得到

$$f(x) = 10 + e^{-x} * 1000. \tag{74}$$

分布 $f$包含 2000 个样本，如图 5-I.a-b 所示。偏移量设为 10、缩放系数设为 1000，可确保对数底数为 10 不会引入不良影响（第 B 节），同时反映现实世界指数分布的长尾特性。生物医学二维和三维图像中记录的物体大小通常呈长尾分布，这是由于表面积/体积呈二次（二维）或三次（三维）增长所致。

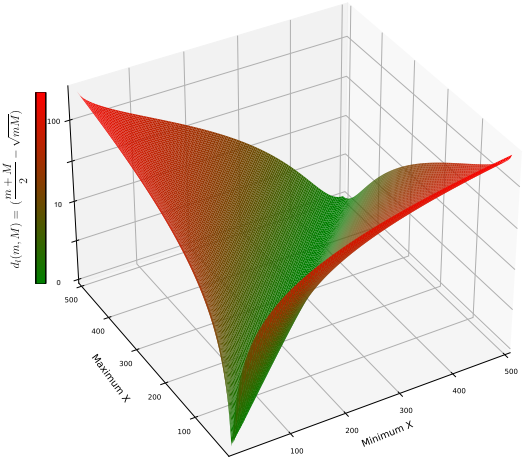



图4：说明向量$m$的最小值（$M$）和最大值（ $X$ ）的变化如何导致 $d$（$m$， $M$）（Z 轴、色条、对数10 刻度）迅速增大，进而使悖论式比较（公式16）更有可能发生。

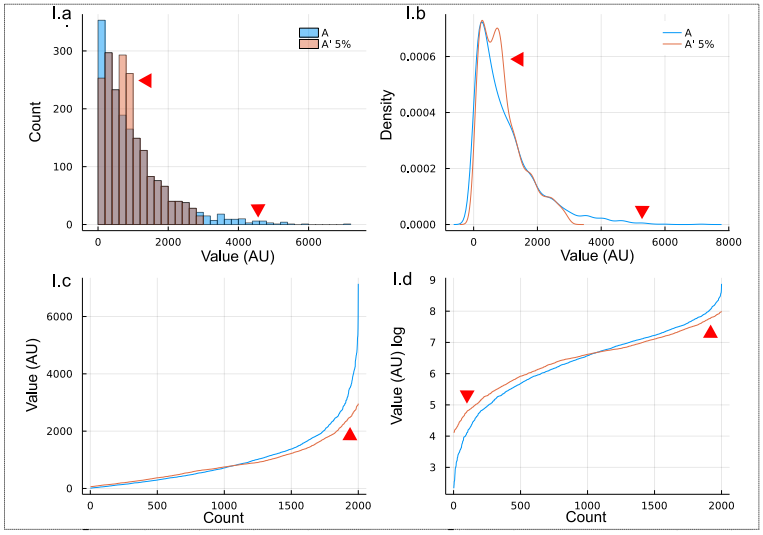



图5：（I）实验中使用的长尾数据分布（$A$， $A'$）（I.a-d）。注意：为避免与 X 轴混淆，图例中使用符号 $A$ 和 $A'$，而非 $X$和 $X'$，来指代向量。随机替换 $A$ 中的 5%，就会显著缩短长尾，无论是在直方图（I.a）还是密度图（I.b）中都是如此，尽管仅有一小部分 $A$被更改。请注意，在排序值图（I.c-d）中，对数变换（I.d）对变化的响应近似对称，但在未变换空间（I.c）中则并非如此，这凸显了算术均值和几何平均数对这类变化的反应为何不同。

长尾现象的成因多种多样：物体可能是不同类别（例如不同的细胞类型、病变、器官）的混合，也可能是由许多小结构聚合成少数大型结构（例如蛋白质亚细胞结构）；分割可能无法将相邻物体分开，导致其大小翻倍；此外还有其他领域特定原因。例如，涉及病理状况表征的研究，如斑块聚集，例如阿尔茨海默病中的淀粉样-$\beta$ [14]，或

如恶性癌症中所体现的失控增长，在大量小对象存在的情况下，很可能会出现频率较低的超大对象。所有这些情形都可能导致超大对象出现频率较低，从而使对象大小测量呈现经典的长尾分布。然而，长尾不一定只出现在分布较大的一端；尾部也可能延伸至最小值，或同时延伸至两端。例如，当噪声（表现为非常小的对象）未被完全滤除时，频率较低的极小对象可能与频率较高的中等大小对象并存。正如我们稍后将看到的（第 3.3 节），在涌现系统中报告对象大小是一个典型例子：较小对象聚集形成较大对象的过程中，可能出现长尾分布，继而产生对数悖论。在非生物医学领域，财富分布（帕累托）也会呈现类似的长尾分布。接下来，我们通过应用算法 $A1$ 来展示 $\triangle_Y f(X)$ 差（公式21），并应用算法 2 来计算替换差 $\delta_{Y,\ Zf}(X)$（公式 23）。图 6展示了移除数值（I.a-b）或插入 $Q$ （II）时的结果，其中插入明确导致了悖论。由于算术——

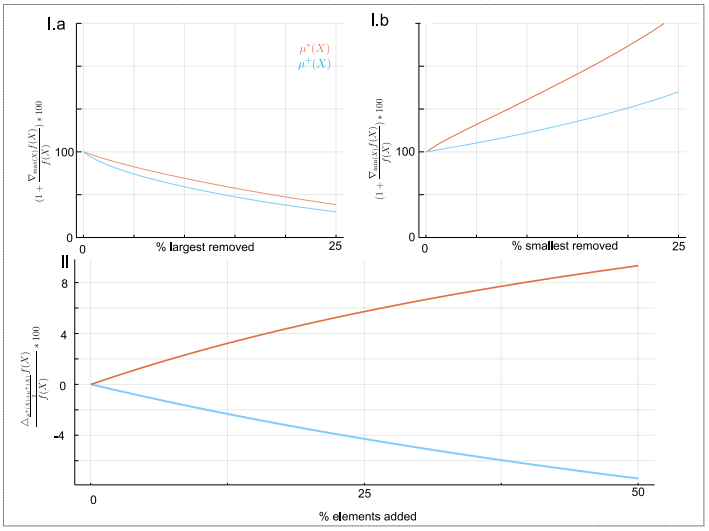



图 6：（I）移除 $k^{th}$ 最大值（I.a）和最小值（I.b）（不进行替换），展示了两个均值的不同响应，以及均值间比率和区间的变化。请注意，移除较小值在数值上等同于添加或重新加权最大值。因此，移除最小值时观察到的更大变化，凸显了极端值对均值的巨大贡献。（II）插入元素时均值发散的变化（算法 1）。每一步中， $x$ （X 轴）都会添加一个新元素 $Q = \frac{\mu^*_{Ax} + \mu^+_{Ax}}{2}$，并计算 $\delta^*$， $\delta^+$ （Y 轴）。 $Q$ 的推导见第2.3.2节。 $A$起始时包含指数分布中的 2000 个元素。均值的发散程度显著增加。

术平均数对极端值非常敏感，因此利用式中的启发式方法，我们可以相当容易地在插入和替换情形下诱发对数悖论。70。即使在随机替换的情况下（图7-A），这种分歧也始终存在，甚至在替换比例达到 10%（2000 个数据点中的 200 个）时，也会出现明显的反转。请注意，仅 5% 就足以显著改变分布形状。分别替换最小值和最大值（如图6-I.a-b 所示），展示了算术平均数对极端值的敏感性。两种均值对极端值的敏感性差异推动了这种分歧（ID(X)）。

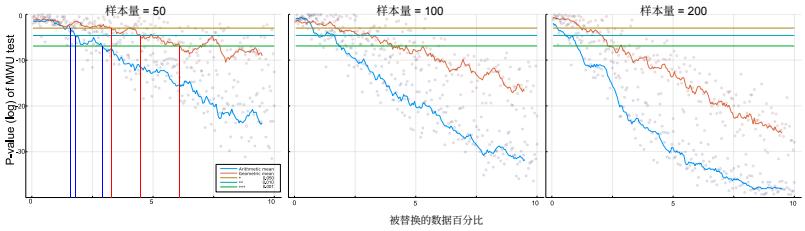



图7：聚合与重采样下对数悖论的普遍性和一致性。“一致性”一词指的是随着参数（x）增加而趋近于已知模式。我们报告曼-惠特尼U检验的p值，用于检验聚合均值是否存在差异（Y轴采用对数刻度）。垂直线表示统计显著性超过三个常用阈值之一的第一个点（水平线分别为0.05-*、0.01-**、0.001-***）。重采样进行50（$N$）次，样本量（$S$）分别为50、100和200。最多替换 $A$ 中的200个元素（10%）以形成 $A'$。替换值从范围 $[\mu^*(A)\ \ \mu^+(A)]$中随机选取。三个标准的统计显著性阈值水平以水平线表示。在数据点之间绘制了一条平滑线（窗口大小为10的均值）。样本量较大时p值下降得更快，并非源于实际样本量；从源数据（$A$）中进行更大规模的重采样意味着更可能捕获极端值，从而使差异最大化。因此，对数悖论在聚合与重采样下始终存在。

3.2. 聚合与重采样情况下的悖论。为检验重采样时是否存在该悖论，我们取生成的数据序列 $A$ （式 74），并按照自助法协议（算法 3）反复进行有放回抽样。我们检验 3 种样本量：50、100 和 200。我们比较对 $A'$ 进行重采样的结果，其中我们将最多 10% 的 $A$ 替换为从区间 $[\mu^*(A)\ \ \mu^+(A)]$（算法 2）中抽取的点。结果见图 7）。对于每对样本（$B_i$、$B_i'$），我们分别计算几何平均数和算术平均数，从而得到两个样本均值分布。我们使用 Mann-Whitney U 检验（MWU）来检验二者是否存在差异。该检验是一种常用的非参数检验，我们报告其 p 值。我们只报告差异呈现悖论的情况，即几何平均数上升而算术平均数下降。请注意，重采样或聚合方法的选择，以及最佳检验方法（例如 MWU）的选择，都取决于研究问题和底层数据分布。在这里，我们更关注的是观察在这些报告条件下是否存在对数悖论；如果存在，不同参数选择下的模式会如何变化。将抽样数据点的数量设为 50 到 200 之间，以兼顾统计效力，避免仅因数据量而导致统计显著性。请注意，增加 $A$ 和 $A'$（$S$）的样本量会迅速带来“统计显著性”。增加样本量会提高纳入极端值的概率，而极端值反过来会使悖论更容易出现。仅随机替换 $A$ 中的 5%（图 7 X=100），就很容易得到 p 值 <g 0.001，这证实了即使是随机替换，也能轻易且稳定地诱发悖论。图中绘制了一条平滑曲线，即对每个数据点取移动平均（窗口为 =10）。与前述结果和已有认知一致，算术平均数（蓝色）明显更容易偏离其初始值。该实验清楚表明，重采样并不一定能避免悖论。

3.3. 生物医学成像启发的应用案例。 在此应用案例中，我们使用马尔可夫链对蛋白质复合物的形成/解体动态变化以及随后对捕捉此类结构的图像进行定量分析建模，如图 8-I 所示。蛋白质结构，

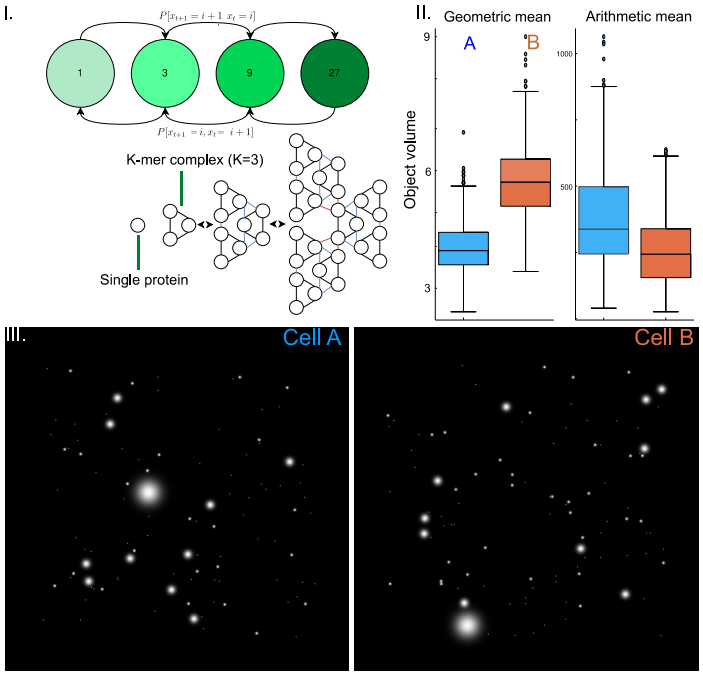



图 8：由马尔可夫链建模的细胞中蛋白质复合物的聚合—解聚动态。单个蛋白质结合形成 k-聚体，其形成取决于环境条件。在此，我们展示了一个简单的 3-聚体三路嵌套聚合。马尔可夫链在节点之间存在有向弧 $i$ 和 $j$，其权重为概率 $P[x_{t+1}=j|x_t=i]$，用于确定类型为 $i$ 的复合物在时间 $t$ 转变为类型为$j$ 的复合物的可能性，时间为 $t+1$。II：我们从表 4（A、B）中的概率分布抽取 1000 个虚拟细胞，并测量每个细胞中复合物的平均体积。箱线图显示，与算术平均值相比，几何平均值呈现出相反的组间差异。III：从表 4中的马尔可夫链模型抽取的两幅图像。这些图像表明，蛋白质尺寸分布的差异在视觉上可能并不明显，但却可能导致显著不同的定量分析。此外，在进行对数变换后，分析人员可能得出悖论式结论。

或复合物，由一个或多个 k-聚体组成，其中 $k$ 表示构成该结构的单个蛋白质数量。物理学和生物环境决定多个蛋白质如何聚集形成 k-聚体，以及多个 k-聚体如何聚集形成更大的蛋白质结构。从节点 $i$ 指向节点 $j$ 的有向弧以概率加权，$P[x_{t+1}=i|x_t=j]$，即复合物 $x$ 类型 $i$ 在时间 $t$ 转变为类型 $j$ 的时间 $t+1$。例如，若弧 $1\to 3$ 的权重为 0.3，则单个蛋白质形成三聚体的概率为 30%。我们通过两个不同的示例细胞系 A 和 B，展示一项简单的计算机模拟实验；在其中，蛋白质以嵌套三聚体的形式组合（图 8-I），两者的分布略有不同，B 更倾向于形成三聚体，而非 27 聚体和单个蛋白质。表 4 列出了每个细胞中的结构数量（按相同蛋白质总量归一化）及其频数分布。接下来，我们记录这些结构的体积，估算方式为 $k^3$。请注意，尽管超大型结构在概率分布中所占比例极小（表4），但由于体积庞大，它们在视觉上十分醒目，如图 8-III 所示。我们从每种分布中抽取 1000 个“细胞”，每个由 525 个 k-聚体组成，以模拟大规模受控实验。图 8-III 展示了每种类型各一个细胞的二维示例图像。图 8-II 绘出了每个细胞的几何均值和算术平均值分布

将“细胞”作为总体均值的估计值。研究问题“蛋白质结构在 A ＞B 中的体积是多少？”现在有两个相互矛盾的答案，因为 $\mathbb{E}[\text{对数。}(A)] < \mathbb{E}[\text{对数 。}(B)]$，而 $\mathbb{E}[A] > \mathbb{E}[B]$。这个计算机模拟示例说明，对数悖论的存在实际上表明，背后存在复杂的动态变化，使细胞系 A 和 B 得以区分。

**4. 讨论**。 对数悖论的存在可能会使实践者倾向于彻底放弃对数变换，这与 Feng 等人 [4]所主张的做法类似，即转而采用广义线性模型等方法 [20]。然而，我们认为，这一悖论具有信息价值，并非放弃对数变换的理由。我们确定了悖论出现所需的必要且充分条件。掌握这些信息后，实践者无需舍弃变换后的结果，而可以根据这些前提条件，选择并应用更复杂的发现方法。我们已说明，隐马尔可夫链模型能够引发这一悖论（图8），有兴趣的读者可参阅 Cowles 等人 [6] 所作的深入综述。马尔可夫链是为抽象的“生成函数”恢复数学模型的一个示例。第 3.3 节中的马尔可夫链抽象及其应用案例，基于亚细胞生物学中的动态变化建模。

例如，线粒体这种大型亚细胞细胞器，会通过融合（合并）和分裂（拆分）来响应饥饿、压力和能量需求。线粒体动态变化功能失调与多种退行性疾病的关联日益受到关注 [21, 27]。另一个例子是小窝，它们是细胞膜上直径为 ∼100 纳米（n m）的球形内陷结构，除了其他功能外，还能帮助细胞膜适应较强的生理应激因素，并具有类似的（解）聚集动态 [2, 34, 18] ，对生物体健康也有重大影响 [5]。细胞会根据物理、化学和遗传触发因素，不断从支架中构建小窝，并将其拆解回支架中；它会由较小的（∼10-20 nm）组分构建小窝，也会将小窝拆解为更小的组分 [17]。在这些情况下，物体（例如 k-mer、线粒体、小窝）的大小与其出现频率呈强烈负相关，即较大的蛋白质/线粒体/小窝比起较小的蛋白质/线粒体/小窝，出现频率更低，因此形成长尾分布。

近年来，人们提出了改进的因果发现算法 [25, 29, 15] ，借助这些算法，可以恢复出能够解释数据生成函数的图因果模型。可解释模型绝非应对这一悖论的唯一途径。例如，近期采用弱监督或无监督方法识别不同亚群体的研究 [24]，展示了针对同一问题陈述的另一种方法。如果所要解决的问题在于量化极端值，那么采用受极值理论启发的模型便是恰当的 [7, 31]。无论如何，这一悖论都能为实践者提供线索，促使他们深入挖掘数据，揭示导致这一悖论的潜在模式。

虽然到目前为止我们讨论的都是右长尾，即极端值大于分布的均值，因而位于数轴右侧，但我们推导出的条件并未将悖论限定为只发生在右长尾中。正如启发式方法的推导（第2.3.2节）所示，悖论同样可能出现在左长尾分布中。对称性也不会排除悖论。不过请记住，只有当均值间距离足够大时，才会诱发悖论；而对于大多数值远大于对数底数、同时最小值和最大值也足够大的分布，这种情况更有可能发生。我们在图10中展示了一个计算机模拟示例。按照定义呈对称性的正态分布，其 $\mu = x$ 参数发生变化，并分别通过添加 $\sqrt{x}$ 和 $x^2$来轻微引入左尾和右尾。我们绘制 $ID(A)$ ，并观察到随着 $\mu$ 增大，尾部的值发生变化，进而使$ID(A)$ 增大，导致对悖论的敏感性增强。本文从一个ID(X)较大的向量 $X$出发，然后寻找 $Y$，以便找到一种悖论式

比较（公式 16），并且总是通过减小 $ID(Y)$来实现。然而，反过来的论证同样成立：给定 $ID(X) = 0$ ，可以通过增大 $ID(Y)$，诱发其与 $Y$ 之间的悖论式比较。

该悖论也可以在预处理和后处理中发挥保障或“金丝雀”的作用。我们记录了即使是简单的分割差异也可能诱发该悖论。随后，可以将该悖论用作检验不同方法或参数值下分割结果稳健性的测试。对数变换中底数的选择在这里同样有所帮助，因为它决定了变换相对于受底数约束的值的敏感性。虽然重点主要放在从长尾分布（通常是不对称的）中抽取的数据上，但启发式方法（第 2.3.2节）的结果应当能够清楚表明，从对称分布中抽取的数据同样可能出现该悖论。最后，可重复性要求始终提供变换后的数据和未变换的数据。

**5. 结论**。 我们表明，长尾不对称分布是对数悖论出现的充分条件，并定义了其效应达到最大的条件。我们推导出了该悖论出现的充要条件，向实践者说明了该悖论可能在何时以及为何出现。我们没有将其描述为一个警示故事，而是指出该悖论揭示了数据中的潜在模式。我们通过实证表明，重采样并不能阻止该悖论。借助一个来自生物医学成像的简化现实用例，我们说明了对亚细胞对象动态变化的（解）构建研究容易受到对数悖论的影响，并就如何超越该悖论、报告关于底层数据的去混杂的研究发现提出了建议。总体而言，按比例分布的长尾数据足以诱发该悖论，这提醒我们应当适当反思数据中的样本独立性，因为资源限制可能会使比例值发生偏斜，例如细胞中的对象大小。对数悖论是一个有用的指标，它能够揭示潜在的复杂模式；更强大、更直观但可能并不总是可用的方法可以捕捉数据生成模型。

## 参考文献

[1] ATHREYA, K. (1987). 无限方差情形下均值的自助法。《统计年鉴》724–731. [2] BASTIANI, M. and PARTON, R. G. (2010). 小窝一览。《细胞科学杂志》123 3831–3836. [3] BOX, G. E. and COX, D. R. (1964). 变换分析。《英国皇家统计学会杂志：B辑（方法学）》26 211–243. [4] CHANGYONG, F., HONGYUE, W., NAIJI, L., TIAN, C., HUA, H., YING, L. et al. (2014). 对数变换及其对数据分析的影响。《上海精神医学》26 105. [5] COHEN, A. W., HNASKO, R., SCHUBERT, W. and LISANTI, M. P. (2004). 小窝和小窝蛋白在健康与疾病中的作用。《生理学评论》84 1341-1379. PMID: 15383654. https://doi.org/ 10 1152 00046 2003 . /physrev. . [6] COWLES, M. K. and CARLIN, B. P. ( ). 马尔可夫链蒙特卡罗收敛诊断：比较综述。《美国统计学会杂志》91 883–904. [7] DE HAAN, L., FERREIRA, A. and FERREIRA, A. (2006). 极值理论：导论 21. 施普林格. [8] EFRON, B. (1979). 自助法：再看刀切法。统计年鉴 7 1-26. https://doi.org/ 10 1214 1176344552 . /aos/ ELLIS, P. D. ( ). 效应量必备指南：统计功效、荟萃分析与研究结果的解释。剑桥大学出版社. [10] FENG, C., WANG, H., LU, N. and TU, X. M. (2013). 对数变换：在生物医学研究中的应用与解释。《医学统计学》32 230–239.[11] FINK, E. L. (2009). 数据变换常见问题。《传播学专论》76 379–397. [12] FLEMING, P. J. and WALLACE, J. J. (1986). 如何避免被统计数据误导：汇总基准测试结果的正确方法。《美国计算机学会通讯》29 218–221. [13] GAO, A. and MARTOS, P. (2019). 对数变换及其对微生物计数中均值和测量不确定性的估计、含义与解释的影响。《AOAC国际期刊》102 233–238.

[14] HART, N. J., KORONYO, Y., BLACK, K. L. and KORONYO-HAMAOUI, M. (2016). 阿尔茨海默病的眼部指标：探索视网膜中的疾病。《神经病理学学报》132 767–787. [15] HUANG, B., ZHANG, K., LIN, Y., SCHÖLKOPF, B. and GLYMOUR, C. (2018). 因果发现的广义得分函数。载于第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集 1551–1560. [16] KEENE, O. N. (1995). 对数变换的特殊性。《医学统计学》14 811–819. [17] KHATER, I. M., LIU, Q., CHOU, K. C., HAMARNEH, G. and NABI, I. R. (2019). 超分辨率模块性分析显示，多面体小窝蛋白-1寡聚体结合形成支架和小窝。《科学报告》9 1–10. [18] KURZCHALIA, T. V. and PARTON, R. G. (1999). 膜微区与小窝。《细胞生物学当前观点》

**11** 424 431 10 1016 0955 0674 99 80061 1 - . https://doi.org/ . /

[19] S - ( ) - LANE, D., SCOTT, D., HEBL, M., GUERRA, R., OSHERSON, D. and ZIMMER, H. (2003). 统计学导论。Citeseer.[20] LIANG, K.-Y. and ZEGER, S. L. (1986). 使用广义线性模型进行纵向数据分析。《生物计量学》73 13–22.[21] LIMA, T., LI, T. Y., MOTTIS, A. and AUWERX, J. (2022). 线粒体在衰老中的多效性作用。《自然·衰老》

**2** 1 15 10 1038 43587 022 00191 2 - . https://doi.org/ . /s - - -

[22] 2008 OLIVIER, J., JOHNSON, W. D. and MARSHALL, G. D. ( ). 实验性免疫球蛋白E结果报告中的对数变换与几何平均数：它们是什么，何时以及为何使用？《过敏、哮喘与免疫学年鉴》100 333–337. [23] PECK, R., OLSEN, C. and DEVORE, J. L. (2015). 统计学与数据分析导论。圣智学习. [24] PENG, X. and WANG, H. J. (2022). 用于亚组识别的广义分位数树方法。《计算与图形统计学杂志》刚刚录用 1–30. [25] PETERSEN, A. H., RAMSEY, J., EKSTRØM, C. and SPIRTES, P. (2022). 使用监督机器学习进行观察科学中的因果发现。[26] PROIETTI, T. and LÜTKEPOHL, H. (2013). Box–Cox变换有助于宏观经济时间序列预测吗？《国际预测杂志》29 88–99. [27] RAEISOSSADATI, R. and FERRARI, M. (2022). 神经退行性疾病中的线粒体-内质网系链。《细胞与分子神经生物学》

**42** 10 1007 10571 020 01008 9 .

[28] 2010 https://doi.org/ . /s - - - RAO, J., ABUGHARBIEH, R. and HAMARNEH, G. ( ). 使用局部图像曲率线索进行图像分割的自适应正则化。载于欧洲计算机视觉会议 651–665. 施普林格. [29] ROHEKAR, R., NISIMOV, S., GURWICZ, Y. and NOVIK, G. (2021). 在可能存在潜在混杂因素和选择偏倚的情况下进行迭代因果发现。[30] SEDRAKYAN, H. and SEDRAKYAN, N. (2018). 调和平均数、几何平均数、算术平均数与均方根不等式。载于代数不等式 21–43. 施普林格. [31] SMITH, R. L. (1990). 极值理论。《应用数学手册》7 437–471. [32] SPIRTES, P., GLYMOUR, C. N., SCHEINES, R. and HECKERMAN, D. (2000). 因果关系、预测与搜索。麻省理工学院出版社. [33] ST-PIERRE, A. P., SHIKON, V. and SCHNEIDER, D. C. (2018). 生物学中的计数数据——数据变换还是模型重构？《生态学与进化》8 3077–3085. [34] STAN, R. V. (2005). 小窝结构。《生物化学与生物物理学报（BBA）—分子细胞研究》1746 334-348. 脂筏：从模型膜到细胞。https://doi.org/

10 1016 2005 08 008 . /j.bbamcr. . . STEELE, J. M. ( ). 柯西-施瓦茨大师

[35] 2004 课：数学不等式艺术导论。剑桥大学出版社. [36] TUKEY, J. W. et al. (1977). 探索性数据分析 2. 马萨诸塞州雷丁. [37] VOGEL, R. M. (2022). 几何平均数？《统计学中的通讯：理论与方法》51 82–94. [38] WAGNER, C. H. (1982). 现实生活中的辛普森悖论。《美国统计学家》36 46–48. [39] WILMOTT, P., HOWSON, S., HOWISON, S., DEWYNNE, J. et al. (1995). 金融衍生品数学：学生入门读物。剑桥大学出版社. [40] XIAO, X., WHITE, E. P., HOOTEN, M. B. and DURHAM, S. L. (2011). 关于使用对数变换与非线性回归分析生物幂律。《生态学》92 1887–1894. [41] ZIMEK, A., SCHUBERT, E. and KRIEGEL, H.-P. (2012). 高维数值数据无监督离群点检测综述。《统计分析与数据挖掘：美国统计学会数据科学期刊》5 363 387–。

## 附录 A：软件

Julia 代码形式的软件、示例输入数据集以及完整文档，均可在开放源代码许可证下获取，网址为https://github.com/bencardoen/LogParadox.jl。与本文版本相对应的 Zenodo 数字对象标识符链接可在此处找到：https://doi.org/10.5281/zenodo.7545842。

## AP附录 B：选择底数的重要性及其后果

在讨论实证结果之前，重要的是要说明底数的选择 $b$ 对对数变换数据会产生何种影响。图 9 绘出了三种常用底数（2、 $e$、10）的对数及其导数（纵轴）。显然，底数为 $b$ 的对数，其导数在 $x < b$处最大。不仅如此，$\lim_{x\to 0} {}^{d}{}_{\text{对数}dx}{}^{b} = +\infty_{(x)}$。直观而言，导数所体现的是，与$x > b$处的值相比，对数变换对 $x < b$处的值变化更快。因此，通过确保 $\forall x \in X: \ x > b$，我们可以避免因数据子集而产生的、取决于所选 $b$ 的非对称效应。

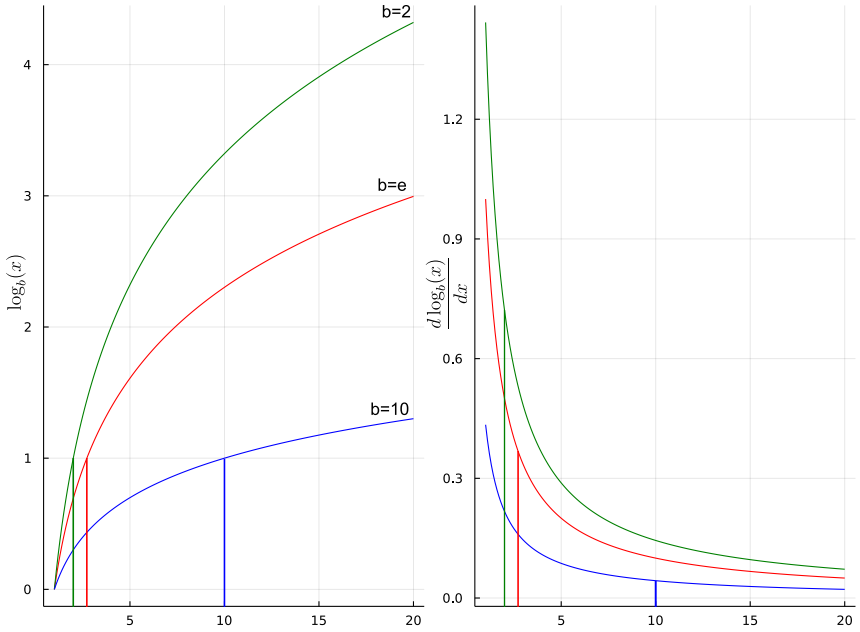



图 9：在分布背景下，选择底数的重要性及其非对称效应示意图。左图显示 $b = [2, e, 10]$ 对应的 对数$_b(x)$，右图显示相对于 $x$ 的相应一阶导数：${}^{d}{}_{\text{对数}dx}{}^{b}{}^{x} = {}^{1}{}_{()\,\text{对数}_e(b)x}$。垂直线标示底数 $b$。请注意，与 $x > b$相比，$x < b$ 和 $x \to 0$ 的导数幅度更大。因此，如果子集 $A \subset X$ 和 $\{a < b | \forall a \in A\}$，那么 $A$ 可能会使变换后的数据 对数$_b \cdot (X)$产生偏差。为此，可以将 $b <$ 最小值$(X)$ 设为，以避免变换引起的偏差。

## 附录 C：对称分布与左尾分布中的悖论 S

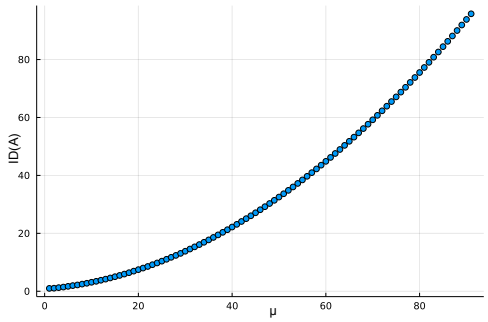



图 10：对称分布在左右两侧具有长尾时对悖论的敏感性。使用一个包含 100 个值的向量 $A$，这些值取样自正态分布 $\mathrm{N}(\mu = x,\ \underline{\sigma} = 2)$。为模拟两侧的长尾，添加 $\sqrt{(x)}$ 和 $x^2$，并将分布参数 $\mu$从 10 到 100 按 1 递增（X 轴）。例如，对于 $x = 16$，$A = [4]||\mathrm{N}(\mu = 16, \sigma = 2||[256]$。从均值间距离 $ID(A)$的图中可以观察到，$ID(A)$ 很容易达到足以引发悖论式比较的程度，因此表明对称分布和左长尾分布同样容易受到影响。

# 附录 D：算法

**算法 1**使用插入启发式方法诱导悖论。

**输入：** 向量 $A$
**输出：** 向量 $A'$
$N \leftarrow |A|$
$\mu^* \leftarrow \exp \frac{\sum \text{对数}(A)}{N}$
$\mu^+ \leftarrow \frac{\sum(A)}{N}$
$A' \leftarrow A \cup \frac{\mu^* + \mu^+}{2}$
**返回** $A'$

**算法 2**使用替换启发式方法诱发悖论。

**输入：** 向量 $A$
**输出：** 向量 $A'$
$N \leftarrow |A|$
$\mu^* \leftarrow \exp \frac{\sum \text{对数}(A)}{N}$
$\mu^+ \leftarrow \frac{\sum(A)}{N}$
$A = A| \setminus$选择$(A)$ ▷ 选择=随机、最小值、最大值、（最小值，最大值）
$A' \leftarrow A||$样本均匀$([\mu^*, \mu^+])$
**返回** $A'$

**算法 3**自助重采样

**输入：** 向量 $A$，样本量 $N$，抽样频率 $M$，统计量 $f$**输出：**
向量 $B$：统计量 $f$ 的抽样分布$B \leftarrow []$ **对于**$\mathrm{i} \in 1 \,..\, M$ **执行**
$B[i] \leftarrow f$(均匀抽样$(A,\ N)$)**结束循环 返回** $B$

# 附录 E：表格

表 1定义参考表。 $X \in \mathbb{R}^N$、 $Y \in \mathbb{R}^M$、 $Z \in \mathbb{R}^K$，以及 $N$、 $M$、 $K \in \mathbb{N}_{>0}$和 $K < N$。请注意，期望值适用于可数有限分布，涵盖本文讨论的数据。

| Name | 符号 | 定义 | 方程 |
|---|---|---|---|
| 向量 | $X$ | $X = [x_1, .., x_n] \in \mathbb{R}^{\lvert X \rvert}$ | 5 |
| 期望值 | $\mathbb{E}(X)$ | $\frac{1}{\lvert X \rvert} \sum_{x_i \in X} x_i$ | - |
| 算术平均数 | $\mu_X^+$ | $\frac{1}{\lvert X \rvert} \sum_{x_i \in X} x_i = \mathbb{E}(X)$ | 10 |
| 几何平均数 | $\mu_X^*$ | $\sqrt[\lvert X \rvert]{\prod_{x_i \in X} x_i} = \mathbb{E}(\text{对数}.(X))$ | 11 |
| 均值间距离（均值间距离） | $ID(X)$ | $\mu_X^+ - \mu_X^*$ | 14 |
| 向量化函数 | $f.(X)$ $F(X)$ | $[f\ (x)\ \ \forall x \in X],\ \ f\colon\ \mathbb{R} \to \mathbb{R}$ | 6 |
| 向量拼接 | $X \| Y$ | $[x_1, .., x_n, y_1, .., y_m]\ \forall x \in X,\ \ y \in Y$ | 7 |
| 向量差 | $X \mid \setminus Y$ | $[x_1, .., x_j]\ \forall x \in X \wedge x \notin Y$ | 8 |
| 悖论式比较 | $X > Y \wedge f.\ (X)\ < f.\ (Y)$ | $\mathbb{E}[X] > \mathbb{E}[Y] \wedge \mathbb{E}[f.\ (X)\ ] < \mathbb{E}[f.\ (Y)\ ]$ | 16 |
| 有限差分（有限差分） | $D_{G,Y} F(X)$ | $F(G_Y(X))\ - F(X)$ | 20 |
| 拼接下的有限差分 | $\triangle_Y F(X)$ | $F(X \| Y)\ - F(X)$ | 21 |
| 删除下的有限差分 | $\nabla_Y F(X)$ | $F(X \mid \setminus Y)\ - F(X)$ | 22 |
| 替换下的有限差分 | $\delta_{Y,Z} F(X)$ | $F((X \| Y) \mid \setminus Z))\ - F(X)$ | 23 |

表2本表回答了以下问题：对于向量 $X$，在由 $Y$和/或 $Z$定义的扰动下，算术均值、几何平均数和均值差如何变化？答案以闭式表达式给出，引入方程链接在最右侧列中。$X$， $Y$， $Z \in \mathbb{R}^N$， $\mathbb{R}^M$， $\mathbb{R}^K$，$\lvert X \rvert = N$， $\lvert Y \rvert = M$， $\lvert Z \rvert = K$。对于替换差，列出了 $\lvert Y \rvert = \lvert Z \rvert$ 的简化情形。有关此处所用符号的完整定义，请参阅表 *1* 。

| 有限差分 | 表达式 | 引入方程 |
|---|---|---|
| $\triangle_Y \mu^+(X)$ | $\frac{M}{N+M}(\mu^+(Y)\ - \mu^+(X))$ | 24 |
| $\triangle_Y \mu^*(X)$ | $\mu^*(X)(\mu^*(X)^{\frac{-M}{N+M}} \mu^*(Y)^{\frac{M}{N+M}} - 1)$ | 34 |
| $\triangle_Y$均值间距离$(X)$ | $\triangle_Y \mu^+(X) - \triangle_Y \mu^*(X)$ | 43 |
| $\nabla_Y \mu^+(X)$ | $\frac{M}{N-M}(\mu^+(X) - \mu^+(Y))$ | 27 |
| $\nabla_Y \mu^*(X)$ | $\mu^*(X)^{\frac{N}{N-M}} \mu^*(Y)^{\frac{-M}{N-M}} - \mu^*(X))$ | 39 |
| $\nabla_Y$均值间距离$(X)$ | $\nabla_Y \mu^+(X) - \nabla_Y \mu^*(X)$ | 44 |
| $\delta_{Y,Z} \mu^+(X)$ | $\frac{M}{N}(\mu^+(Y)\ - \mu^+(Z))$ | 31 |
| $\delta_Y$<style 均值间距离='2'>，$Z^{\mu^*}$（$X$） | $\mu^*(X)(\ \frac{\mu^*(Y)^{M/N}}{\mu^*(Z)^{M/N}} - 1)$ | 42 |
| $\delta_{Y,Z}$均值间距离$(X)$ | $\delta_{Y,Z}{}^{\mu^+}\ (X)\ - \delta_{Y,Z} \mu^*(X)$ | 45 |

表格 3本表回答了以下问题：“给定 $X$，应如何选择 $Y$和 $Z$，以诱导出符号相反的差值，从而产生悖论？”表 *1* 列出了所使用的全部符号。表达式中的颜色编码与图 *2*中使用的颜色编码一致。右栏链接回表达式的推导过程。

| 条件 | 表达式 | 引入方程 |
|---|---|---|
| $\triangle_Y \mu^+(X) < 0$ | $\mu^+(Y) < \mu^+(X)$ | 26 |
| $\nabla_Y \mu^+(X) < 0$ | $\mu^+(X) < \mu^+(Y)$ | 29 |
| $\delta_{Y,Z} \mu^+(X) < 0$ | $\mu^+(Y) < \mu^+(Z)$ | 33 |
| $\triangle_Y \mu^*(X) > 0$ | $\mu^*(Y) > \mu^*(X)$ | 35 |
| $\nabla_Y \mu^*(X) > 0$ | $\mu^*(Y) < \mu^*(X)$ | 38 |
| $\delta_{Y,Z} \mu^*(X) > 0$ | $\frac{\mu^*(Y)}{\mu^*(Z)} > 1$ | 41 |

表4

两个细胞系的示例，每个细胞系的 *k*-聚体比例各不相同。每个细胞中的蛋白质总数相同（*1059*），但转移动力学略有不同，因此与细胞*A*相比，细胞*B*的三聚体（由三个蛋白质组成的复合物）比例更高，而单聚体和*27*聚体的频率则有所降低。最后两行显示了相应的概率（频率）分布。请注意，尽管最大结构的频率非常低，但由于其尺寸远大于其他结构，在实际图像中会主导视觉感知。转移模型（马尔可夫链）如图 *8-I* 所示。就该链而言，*B*中的转移 $\frac{P[x_t=1,x_{t+1}=3]}{P[x_t=3,x_{t+1}=1]}$ 比*A*中的更高。将这两者视为两个细胞系的 *k*-聚体概率分布，并分别从中反复抽样，可得出图 *8-II* 中各细胞蛋白质体积均值（每个细胞的均值 $k^3$）的箱线图。这些箱线图展示了悖论式比较（公式 *16*），即 $\mathbb{E}(A) > \mathbb{E}(B)$，但 $\mathbb{E}(\log.(A)) > \mathbb{E}(\log.(B))$

| **# 蛋白质 / 结构** | 1 | 3 | 9 | 27 | 总数 # 结构 | 总数 # 蛋白质 |
|---|---|---|---|---|---|---|
| 细胞中的结构数量 A | 300 | 100 | 30 | 7 | **437** | **1059** |
| 细胞中的结构数量 B | 240 | 147 | 30 | 4 | **437** | **1059** |
| 细胞中结构的频率 A | 0.283 | 0.094 | 0.028 | 0.006 | **1** | **1** |
| 细胞中结构的频率 B | 0.227 | 0.139 | 0.283 | 0.003 | **1** | **1** |

**致谢**。Ivan Robert Nabi 任职于加拿大不列颠哥伦比亚大学生物医学工程学院。Sieun Lee 任职于英国诺丁汉大学医学院心理健康与临床神经科学系。