关于 Kolmogorov 叠加定理与正则均值

Miguel de Carvalho

arXiv:2601.09659v1 [math.ST] 14 Jan 2026

摘要

尽管柯尔莫哥洛夫概率公理广为人知，但鲜为人知的是，在一篇常被忽视的 1930年短文中，Kolmogorov 提出了一个公理化框架，用于统一均值这一概念，并将其称为 正则均值。这一框架给出了定义明确的函数形式，涵盖算术平均数、几何平均数和调和平均数等。

本文揭示了 Kolmogorov 的两个关键结果之间一种优雅的联系，证明正则均值类可以直接从 Kolmogorov 叠加定理推导得出。这种联系在概念上颇具吸引力，也表明叠加定理理应在统计学中获得更广泛的关注——不仅因为它与正则均值的联系（如本文所示），还因为它对神经网络模型发展的影响，以及它与其他统计框架可能存在的联系。此外，我们确立了正则均值的一种稳定性，表明当生成元发生小幅扰动时，正则均值会平滑变化。最后，我们对近期一项适用于广泛正则均值类的普适中心极限定理进行了阐释。

关键词：正则均值；叠加；普适中心极限定理。

1 引言

1.1 背景与动机

尽管柯尔莫哥洛夫概率公理广为人知，但他那篇常被忽视的1930年短文 [7] 提出了一个公理化框架，用于统一均值这一概念。借助这一框架，Kolmogorov证明，与这些公理相容的唯一函数形式包含算术平均数、几何平均数和调和平均数等作为特例。

本文以 Kolmogorov 二十多年后建立的一项强大却似乎毫不相关的结果为基础——通常称为 Kolmogorov 叠加定理 [例如， 1, 11]。该定理首次得到证明时意义重大，因为它推翻了一个隐含在希尔伯特第十三问题中的猜想。希尔伯特第十三问题于 1900年 在国际数学家大会上提出。历史回顾请参见 [15]。

M. de Carvalho；数学学院，爱丁堡大学；英国 EH9 3FD。

顾名思义，该定理涉及叠加。用数学术语来说，叠加是一个以函数为自变量的函数；例如，

$$f(x_1, x_2, x_3) = g\big(a(\alpha(x_1), \beta(x_2, x_3)),\, b(x_2, x_3)\big), \tag{1}$$

是单变量函数和双变量函数的叠加。叠加定理指出，任何多元连续函数都可以表示为单变量连续函数的和与叠加。通俗地说，任何多元连续函数因此都可以仅使用有限个单变量连续函数来精确表示。

该定理的影响延续至今，成为广泛神经网络模型的基础，这些模型的设计受其原理启发 [例如， 3, 5, 9, 10, 14, 17, 18]。这些神经网络模型将叠加中的两组不同单变量连续函数解释为两个不同的层：所谓的内函数对应于隐藏层中的映射，而外函数对应于输出层中的映射。叠加定理是此类模型的理论基础，正如多层感知器以通用逼近定理为基础一样。

正如本文所示，Kolmogorov 提出的这两个看似无关的结果——即正则均值的函数形式和叠加定理——之间存在一个重要联系，而这一联系此前并未得到充分重视。具体而言，正则均值的函数形式可以作为 Kolmogorov 叠加定理的推论导出。这一联系在概念上颇具吸引力，并表明叠加定理应在统计学中得到更广泛的认可——不仅因为它与正则均值存在如本文所示的联系，也因为它对神经网络模型发展的影响（如前文所述），以及它与其他统计框架的潜在联系（后文将略作提示）。此外，我们还确立了正则均值的一项稳定性，表明生成元发生小幅扰动时，正则均值会平滑变化。我们将这一结果与正则平均数公理体系以及几何平均数的近似联系起来——后者在现代投资组合理论中发挥着基础性作用 [13,第 3] 章。

最后，我们将深入探讨一项近期提出的普适中心极限定理，该定理适用于广义的正则均值。该结果最初由 de Carvalho [2], 提出，确立了适用于所有类型均值的中心极限定理，包括算术平均数、几何平均数和调和平均数。尽管该定理具有普遍性，但此前人们对其有限样本行为在理论和实证方面都知之甚少，而本研究旨在弥补这一空白。我们通过 Edgeworth展开探讨影响正态近似收敛速度的因素，并通过数值实验展示这些发现。

1.2 正则均值的准备工作

在继续之前，我们先为正则均值奠定基础。令 $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$。正则均值 是满足以下公理的映射 $M : \mathbb{R}^n \to \mathbb{R}$：

- A1) $M(\mathbf{x})$ 对每个变量都是连续且递增的。

- A2) $M(\mathbf{x})$ 是对称函数。

- A3) $M(x\mathbf{1}_n) = x$，即，n 个相同值的均值等于它们的共同值。

- A4) $\mathbf{x}$ 的均值在任一分量被其均值替代后保持不变，$m = M(x_1, \ldots, x_{n_0})$，

  i.e. $M(\mathbf{x}) = M(\underbrace{m, \ldots, m}_{n_0}, x_{n_0+1}, \ldots, x_n)$.

[7] 证明了，如果满足条件 A1 至 A4，则函数 $M(\mathbf{x})$ 具有如下形式

$$M(\mathbf{x}) \equiv M_g(\mathbf{x}) = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right), \tag{2}$$

其中 $g : \Omega \subseteq \mathbb{R} \to \mathbb{R}$ 是一个连续单调函数，g^{-1} 表示其反函数。出于稍后将会明了的原因，我们将 g 称为生成元（或激活）函数，并将 Ω 称为输入范围。例如，若 $g(x) = x$ 且 $\Omega = \mathbb{R}$，则得到算术平均数，

$$M_x(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} x_i.$$

此外，$g(x) = \log x$ 和 $g(x) = 1/x$，其中 $\Omega = (0, \infty)$，分别对应几何平均数和调和平均数：

$$M_{\log x}(\mathbf{x}) = \left(\prod_{i=1}^{n} x_i\right)^{1/n}, \qquad M_{1/x}(\mathbf{x}) = \frac{n}{\sum_{i=1}^{n} x_i^{-1}}.$$

方程 (2) 还包含其他例子作为特例，例如所谓的幂平均：例如，若 $g(x) = x^p$，对于 $p \in (0, \infty)$，则得到 $M_{x^p}(\mathbf{x}) = \{1/n \sum_{i=1}^{n} x_i^p\}^{1/p}$。

Kolmogorov 对 (2) 中结果的证明并非简单直接，需要使用辅助函数，并对有理数和无理数采用不同的方法。另一种证明方法见 Leinster [8,第 5] 章。

1.3 概述

第 2.1 节介绍叠加定理，并推导出两个推论，其中一个确立了（2）。鉴于这一联系，第 2.2 节讨论生成元的性质，第 2.3 节则考察正则均值的连续性。第 3 节陈述普适中心极限定理，通过 Edgeworth展开分析收敛速率，并以数值示例说明该结果。

2 正则均值作为叠加

2.1 均值是叠加

下面给出 Kolmogorov 叠加定理的一个现代版本，它与[11]中的版本等价。

定理 1. 设 $f:\ A \to \mathbb{R}$ 是定义在紧集 $A \subset \mathbb{R}^n$上的多元连续函数。那么，它可以表示为

$$f(\mathbf{x}) = \sum_{j=1}^{2n+1} l\left(\sum_{i=1}^{n} \lambda_i g_j(x_i)\right), \tag{3}$$

其中外函数和内函数 l 与 g_j均为连续的一元函数， g_j 是递增的，且$\sum_{i=1}^{n} \lambda_i \leq 1$ 满足 $\lambda_i \geq 0$, 对所有 i均成立。

简单来说，任何连续多元函数都可以表示为单变量函数的函数。用数学术语来说，这种由函数构成的函数称为叠加。因此，定理 1 可以概括为：任何连续多元函数都可以用单变量函数的和与叠加来表示。通常取 $A = [0,1]^n$，但众所周知，叠加定理对于紧致度量空间[16]也更一般地成立。

正则均值 (2) 的函数形式与 (3) 中叠加定理的函数形式相似。事实上，在 (3) 中令

$$\lambda_i = \frac{1}{n}, \quad g_j(x) = g(x), \quad l(x) = \frac{1}{2n+1} g^{-1}(x),$$

我们得到式 (2) 中的正则均值。

正则均值需要满足的所有公理中，定理 1 只要求连续性。在定理 1 的假设中加入对称性（即 A2），可得以下结果，其证明见附录； [20]中也提出了类似结果。

推论 2。 设 $f:\ A \to \mathbb{R}$ 是定义在紧集$A \subset \mathbb{R}$上的对称多元连续函数。那么它可以表示为

$$f(\mathbf{x}) = l\left(\lambda \sum_{i=1}^{n} g(x_i)\right), \tag{4}$$

具有连续的一维外函数和内函数 l 以及 g，其中 g 递增，且$\lambda n \leq 1$ 满足 $\lambda \geq 0$。

方程 (4) 已经更接近 (2) 中正则均值的函数形式。接下来，我们将证明 A1、A2、A3 和 A4 最终为我们提供了达到 (2) 所需的一切条件。

推论 3。令 $f:\ A \to \mathbb{R}$ 为一个满足 *A1* 至 *A4* 的多元函数。则它具有如下表示形式

$$f(\mathbf{x}) = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right), \tag{5}$$

其中 g为连续的一元递增函数。

尽管这一结果广为人知 [例如， 2], 但从叠加定理（定理 1）推导出该结果完全是全新的，并揭示了 Kolmogorov 的两个基本结果之间的优雅联系。换言之，推论 3 为正则均值提供了一个叠加定理，表明满足 A1–A4 的唯一函数形式是（5）；也就是说，满足 A1–A4 的唯一可接受函数是正则均值。

2.2 正则均值是简单神经模型吗?

由第 2.1 节可知，正则均值与式 (3) 具有相同的结构，因此接下来我们将根据这种表示与神经网络模型进行类比。在继续之前，我们强调，正则均值不 应被视为实际的神经网络模型，尽管这种类比很有启发性。在这种理解下， g 和 g^{-1} 扮演激活函数的角色，而 g仅允许取连续单调函数——因此，GELU（高斯误差线性单元）等非线性函数不在允许之列。表 1 列出了常见均值及其对应的激活函数。虽然其中有些函数在深度学习中并不常见，但它们自然地出现在 Kolmogorov 的正则均值这一语境中。

表 1：常见均值、其对应的激活函数，以及式（6）中定义的 Kolmogorov 期望值。

激活函数 (g)	**输入极差 (Ω)**	**结果 mean**	**表达式** (M)	**Kolmogorov 期望值**
x	$\mathbb{R}$	算术平均数	$1/n\sum_{i=1}^{n} x_i$	$\mathrm{E}(X)$
$\log(x)$	$x>0$	几何平均数	$(\prod_{i=_1}^{n} x_i)^{1/n}$	$\exp\{\mathrm{E}(\log\ X)\}$
$1/x$	$x>0$	调和平均数	$n/\sum_{i=1}^{n} x_i^{-1}$	$1/\,\mathrm{E}(1/X)$
x^p	$x>0$	幂	$\{1/n\sum_{i=1}^{n} x_i^p\}^{1/p}$	$\{\mathrm{E}(X^p)\}^{1/p}$
$\exp(x)$	$\mathbb{R}$	指数	$\log\{1/n\sum_{i=1}^{n}\exp(x_i)\}$	$\log[\mathrm{E}\{\exp(X)\}]$

2.3 正则均值的连续性

本文的主要理论洞见在第 2.1 节中确立：式 (2) 中所有均值的函数形式，都是 Kolmogorov 叠加定理的推论。尽管次要，下面的结果本身也很有意义。令 $C(A)$ 表示紧集 $A \subset \mathbb{R}^n$ 上连续函数构成的空间，令 $C_{\mathrm{LI}}(B)$ 表示紧集 $B \subset \Omega$ 上利普希茨连续、严格单调函数构成的空间。两个空间均赋予上确界范数

$$\|f\|_\infty = \sup_{x \in X} |f(x)|, \qquad X \in \{A, B\}.$$

以下结果成立。

定理 4. 设 $M : C_{\mathrm{LI}}(B) \to C(A)$ 是由 $M_g(\mathbf{x}) = g^{-1}(\sum_{i=1}^n g(x_i))$ 定义的算子，其中A 是 $\mathbb{R}^n$的紧子集， B 是Ω的紧子集。那么， M_g 在 g中连续。

A1 要求正则均值作为数据$\mathbf{x}$的函数具有连续性，而定理 4 则证明， M_g 本身作为生成元 g的函数也是连续的。这一性质保证了正则均值的稳定性，确保正则均值不会因 g的微小扰动而突然变化。如果不具备这种连续性，两个几乎相同的激活函数 g 和 h可能会产生明显不同的均值，使 $M_g(\mathbf{x})$变得不稳定或不可靠。推导该结果所需的条件较为宽松；例如，表 1 中列出的所有激活函数在任意紧集 $B \subset \Omega$上都是利普希茨连续的。

定理 4 表明，可以从基于某个生成元的均值转向由相近生成元定义的另一个均值，同时两个均值仍然保持接近。这一稳定性原理在金融领域得到了广泛应用，如以下示例所示。

示例 1（投资组合理论）。 对于期间收益率为 r_1，……， r_n 的投资组合，几何平均数是适当的平均收益率概念。事实上，如果 w_0 是初始财富，那么

$$w_n = w_0(1 + r_1) \cdots (1 + r_n) = w_0 \left\{ M_{\log x}(1 + r_1, \ldots, 1 + r_n) \right\}^n,$$

因此， n 个期间的平均收益率是几何平均数

$$M_{\log x}(1 + r_1, \ldots, 1 + r_n) = \left\{ \prod_{t=1}^n (1 + r_t) \right\}^{1/n}.$$

由于投资组合理论通常是在均值—方差框架下表述的，几何平均数通常通过围绕样本均值 $\bar{r}$ 进行一阶泰勒展开来近似，得到

$$M_{\log x}(1 + r_1, \ldots, 1 + r_n) \approx \exp\{\bar{r} - (\bar{r}^2 + s^2)/2\},$$

其中 $\bar{r}$ 和 s^2 分别表示 r_1、……、 r_n 的样本均值和方差。关于这一近似及相关近似和它们的应用，请参阅 Markowitz [12]。

3 关于普适中心极限定理

3.1 正则均值的大样本行为

标准中心极限定理讨论的是算术平均数的极限行为，而定理5则是一个一般的中心极限定理，适用于整个正则均值类，包括表 1 中列出的正则均值。在下文中，设 $\mathbf{X} = (X_1、\ldots\ldots、X_n)$，并将柯尔莫哥洛夫期望值定义为:

$$E_g(X) = g^{-1}(E_x\{g(X)\}), \tag{6}$$

其中 $E_x\{g(X)\} = \int g(x)\,\mathrm{d}F(x)$，$X \sim F$，且 g 是连续单调函数

定理5 （普适中心极限定理）。假设 $X_1、\ldots\ldots、X_n \overset{\text{iid}}{\sim} F_X$，且 $\mathrm{var}\{g（X）\} < \infty$，其中 $g（x）$ 是严格单调函数，并且在 $x = E_g（X）$ 处的导数为 $g'（E_g（X）$。那么，当$n \to \infty$时，

$$\sqrt{n}\{M_g(\mathbf{X}) - E_g(X)\} \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{\mathrm{var}\{g(X)\}}{\{g'(E_g(X))\}^2}\right), \tag{7}$$

其中 $M_g（\mathbf{X}）$如 （2）所示。

普适中心极限定理归功于 de Carvalho [2]。尽管该结果已得到充分确立，但其收敛性质和数值性能仍仅得到部分理解。在本节剩余部分，我们旨在进一步阐明这些问题。

为了深入了解影响收敛到正态分布速度的因素，我们采用 Edgeworth展开。令 F_n 表示.....的分布函数

$$\sqrt{n}\left(\frac{M_g(\mathbf{X}) - E_g(X)}{\sqrt{\mathrm{var}\{g(X)\}}/g'(E_g(X))}\right),$$

并令 G_n 表示.....的分布函数

$$\sqrt{n}\left(\frac{n^{-1}\sum_{i=1}^n g(X_i) - E\{g(X)\}}{\sqrt{\mathrm{var}\{g(X)\}}}\right).$$

标准的 delta 方法论证 [6, 第 2 节7] 表明，$F_n(x) = G_n(x) + o(1)$，当 $n \to \infty$ 时。由于 $G_n(x)$ 的埃奇沃思展开众所周知 [6,第 2 节6], ，因此可得

$$F_n(x) = \Phi(x) - \phi(x)\left(\frac{\gamma_g p_1(x)}{6\sqrt{n}} + \frac{\kappa_g p_2(x)}{24n} + \frac{\kappa_g^2 p_3(x)}{72n}\right) + o(n^{-1}),$$

其中，$\Phi(x)$ 和 $\phi(x)$ 分别是标准正态分布的分布函数和密度函数，γ_g 和 κ_g 分别表示 $g(X)$ 的偏度和超额峰度，而 p_1、p_2 和 p_3 是埃尔米特多项式（即 $p_1(x) = x^2 - 1$、$p_2(x) = x^3 - 3x$、$p_3(x) = x^5 - 10x^3 + 15x$）。上述

Edgeworth展开表明， $g(X)$ 的尾部越重或偏度越大，收敛到极限正态分布的速度就越慢。正如预期的那样，当 $g(x) = x$时，该展开就化为算术平均数的标准埃奇沃斯展开。

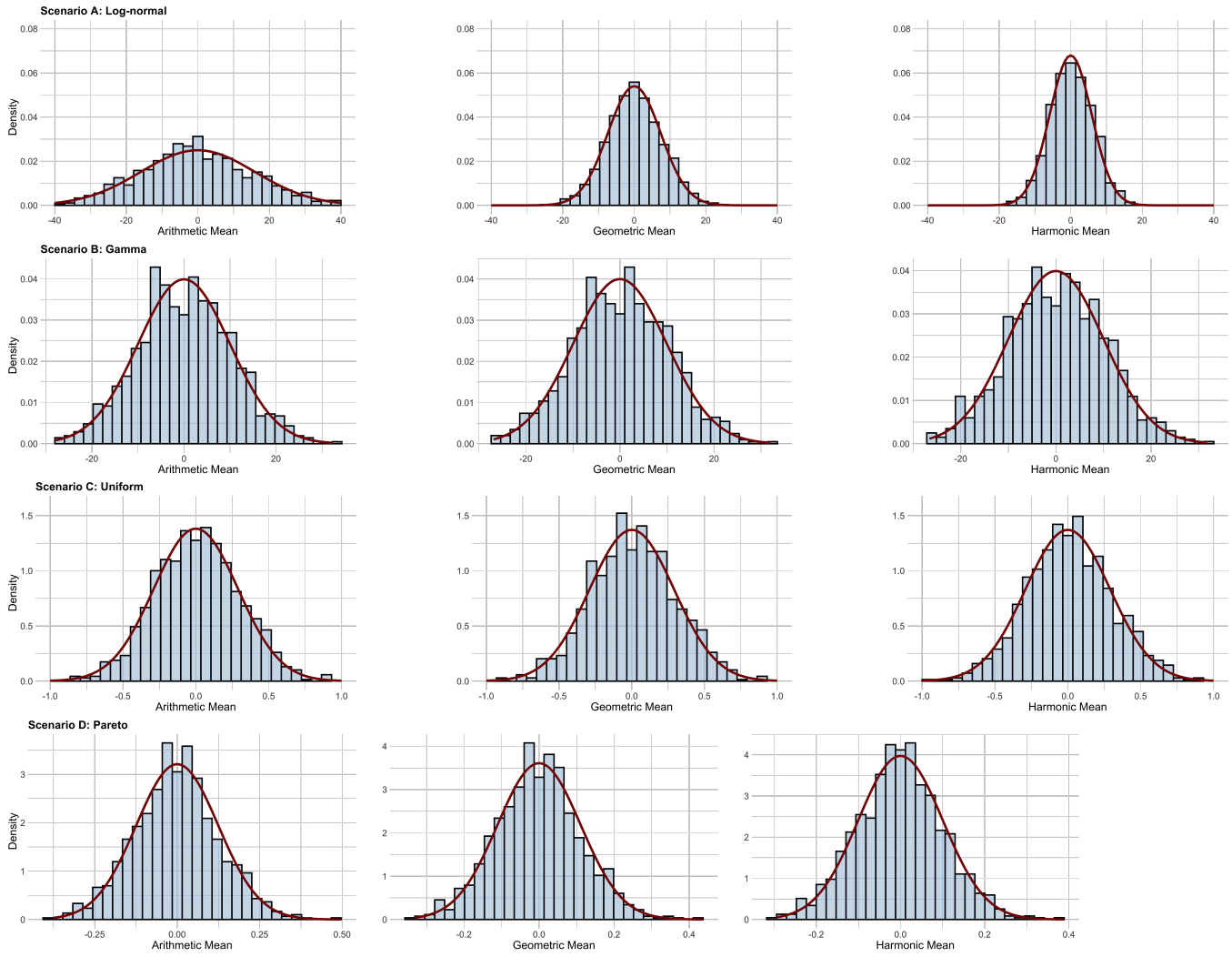

图1：蒙特　　卡洛模拟结果：普适中心极限定理， $n = 1\,000$.

3.2 数值实验

为展示普适中心极限定理在实践中的应用，我们报告一项模拟研究的结果：从以下模型中抽取 1 000 个大小为 $n = 1\,000$ 的数据集：

- 情景 A：对数正态分布(μ, σ^2)，其中 $(\mu, \sigma^2) = (2, 1)$。
- 情景 B：伽马分布(a, b)，其中 $(a, b) = (100, 1)$。
- 情形 C：均匀分布(a, b)，其中 $(a, b) = (1, 2)$。
- 情景 D：帕累托分布(α)，其中 $\alpha = 10$。

结果见图1。左列对应标准中心极限定理的特例（$g(x) = x$），并作为比较之用。中列和右列分别展示几何平均数和调和平均数的特例（分别为 $g(x) = \log x$ 和 $g(x) = 1/x$）。如图所示，经验分布与理论预测高度吻合，与定理5所蕴含的渐近正态分布一致，其渐近方差由式（7）给出。

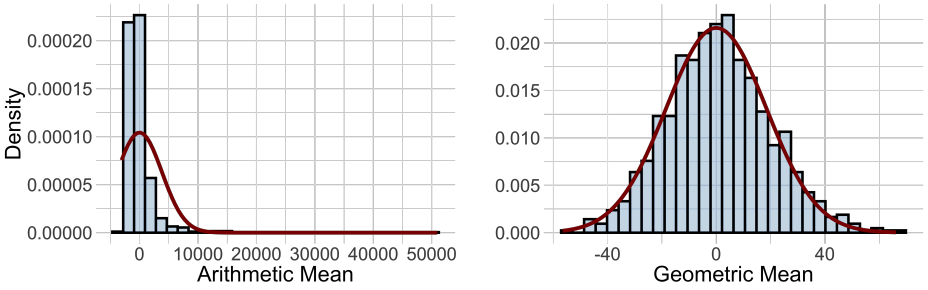

图2： $n = 1\,000$ 服从对数正态分布(μ， σ^2)时的更多模拟结果，其中$\mu = 2$ 且 $\sigma = 2.5$。正如第3.1节中的Edgeworth展开所预示的，算术平均数的收敛速度较慢，而几何平均数的收敛速度较快。

第3.1节中Edgeworth展开的理论见解与我们的数值结果一致。例如，在对数正态分布情形下，当 σ 较大时，标准中心极限定理($g(x) = x$)的收敛速度较慢，而几何平均数($g(x) = \log x$)的收敛速度则快得多，因为此时 $g(X) \sim \mathrm{N}(\mu, \sigma^2)$，因此，正如Edgeworth展开所预示的，正态近似能更快地达到准确。图2展示了这一现象，图中结果基于对数正态分布(μ， σ^2)，其中 $\mu = 2$ 且 $\sigma = 2.5$。

4 证明与技术细节

4.1 推论 2 的证明

对称性（即 A2）表明，对于指标的任意排列 σ， $f(x_1, \ldots\ldots, x_n) = f(x_{\sigma_1}, \ldots\ldots, x_{\sigma_n})$。这与定理 1 一起意味着，对于任意排列，

$$\sum_{j=1}^{2n+1} l\left(\sum_{i=1}^{n} \lambda_i g_j(x_i)\right) = \sum_{j=1}^{2n+1} l\left(\sum_{i=1}^{n} \lambda_i g_j(x_{\sigma_i})\right).$$

进而，这意味着所有 λ_i 都必须相同，才能确保上述等式成立；也就是说，对于所有 i， $\lambda_i = \lambda$ 。因此，

$$\sum_{j=1}^{2n+1} l\left(\lambda \sum_{i=1}^{n} g_j(x_i)\right) = \sum_{j=1}^{2n+1} l\left(\lambda \sum_{i=1}^{n} g_j(x_{\sigma_i})\right). \tag{8}$$

现在，要使 (8) 对任意排列都成立，唯一的可能是 g_j 完全相同，因为任何变化都会破坏对称性；也就是说，对于所有 j， $g_j = g$ 。这进而意味着

$$l\left(\lambda \sum_{i=1}^{n} g(x_i)\right) = l\left(\lambda \sum_{i=1}^{n} g(x_{\sigma_i})\right),$$

因此，我们已证 明，在对称性条件下，内函数和外函数均为常数，

$$f(x_1, \ldots, x_n) = l\left(\lambda \sum_{i=1}^{n} g(x_i)\right),$$

从而得出了最终结果。 □

4.2 推论 3 的证明

首先注意，根据假设 A1 和 A2 f 是对称且连续的，因此可以应用定理 3。接下来，由定理 3 可知 $g(x)$ 是递增的，且 $\lambda > 0$，这意味着

$$h(\mathbf{x}) \equiv \lambda \sum_{i=1}^{n} g(x_i)$$

在每个变量上递增。由于根据假设，A1 $f(\mathbf{x})$ 在每个变量上递增，接下来我们证明 $y \mapsto l(y)$ 必须递增；否则， $f(\mathbf{x}) = l(h(\mathbf{x}))$ 和 $h(\mathbf{x})$ 不可能同时在每个变量上递增。事实上，注意到 f 和 h 对每个变量的单调性意味着，对任意 $\varepsilon > 0$， $a \equiv h(x_1 + \varepsilon, \ldots, x_n) > b \equiv h(x_1, \ldots, x_n)$，且 $l(a) > f(x_1 + \varepsilon, \ldots, x_n) > f(x_1, \ldots, x_n) = l(b)$。由于同样的论证也适用于 x_2、 x_3 等，因此这证明了若 $a > b$，则 $l(a) > l(b)$，从而确立了 l 是递增的，因而是单射。最后，将 A3 和 A4 与 (4) 结合可知，对于 $n_0 = n$ 的情形， $x = f(x_1, \ldots, x_n)$，因此由于 l 是单射，

$$\begin{aligned}
&f(x_1, \ldots, x_n) = f(x, \ldots, x) \\
&\Rightarrow l\left(\lambda \sum_{i=1}^{n} g(x_i)\right) = l\left(\lambda \sum_{i=1}^{n} g(f(x_1, \ldots, x_n))\right) \\
&\Rightarrow \sum_{i=1}^{n} g(x_i) = n\, g(f(x_1, \ldots, x_n)) \\
&\Rightarrow g^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} g(x_i)\right) = f(x_1, \ldots, x_n),
\end{aligned}$$

因此，结论得证。 □

4.3 定理 4 的证明

任取 $h \in C_{\mathrm{LI}}(B)$。由于 g 既严格单调又利普希茨连续，很容易证明 $g^{-1} \in C_{\mathrm{LI}}(B)$ [4, 定理 22。1]。利用中值定理的标准论证可得

$$\|g^{-1} - h^{-1}\|_\infty \le \frac{1}{m}\|g - h\|_\infty, \tag{9}$$

其中，由于 g 和 h 均严格单调，因此 $m = \min\{\inf_{x\in A} g'(x),\ \inf_{x\in A} h'(x)\} \neq 0$。

令 L 为 g^{-1} 的利普希茨常数，令 $G_i = g(x_i)/n$ 且 $H_i = h(x_i)/n$，对所有 i 均成立。方程 (9)、三角不等式以及 $g^{-1} \in C_{\mathrm{LI}}(B)$ 这一事实表明

$$\begin{aligned}
\|M_g - M_h\|_\infty &= \|g^{-1}(\textstyle\sum_{i=1}^n G_i) - h^{-1}(\sum_{i=1}^n H_i)\|_\infty \\
&\le \|g^{-1}(\textstyle\sum_{i=1}^n G_i) - g^{-1}(\sum_{i=1}^n H_i)\|_\infty + \\
&\quad \|g^{-1}(\textstyle\sum_{i=1}^n H_i) - h^{-1}(\sum_{i=1}^n H_i)\|_\infty \\
&\le L\|\textstyle\sum_{i=1}^n (G_i - H_i)\|_\infty + \|g^{-1} - h^{-1}\|_\infty \\
&\le L\|g - h\|_\infty + 1/m\,\|g - h\|_\infty \\
&\le (L + 1/m)\,\|g - h\|_\infty.
\end{aligned}$$

因此，为确保对所有 $\varepsilon > 0$ 均有 $\|M_g - M_h\|_\infty < \varepsilon$，只需考虑 $\|g - h\|_\infty < \delta$，其中 $\delta = \varepsilon/(L + 1/m)$。 □

5 结束语

本文揭示了 Kolmogorov 的两个重要结果之间优雅的联系，指出（2）中的正则均值类可以直接由著名的叠加定理导出；参见推论 3。除此之外，本文还凸显了公理化视角在统计学及相关领域中常被忽视的力量。假设统计量 f(x) 连续，等价于假设叠加定理成立。以此为起点，我们可以进一步探讨统计量 f(x) 还应具备哪些额外性质（例如对称性），从而进一步约束其表示形式，使其更接近一个感兴趣的概念，这与推论 2 和 3 的思路一致。

我们还证明了正则均值光滑性方面的一项结果（定理 4），可将其视为一种稳定性原理，它扩展了 A1 所施加的连续性要求。最后，我们对普适中心极限定理作了探讨；该定理可视为一种涵盖几乎所有类型均值的中心极限定理，包括算术平均数、几何平均数和调和平均数。

上述所用的叠加定理版本有时也称为 Kolmogorov–Arnol'd–Kahane–Lorentz–Sprecher 定理。可以构造其他非单调内函数 [15, 定理 4。2]事实上，该结果的另一个著名版本取消了对内函数单调性的要求，同时要求 $2n+1$ 外函数。具体而言，设$f : A \to \mathbb{R}$ 是一个连续函数，则它可表示为

$$f(\mathbf{x}) = \sum_{j=1}^{2n+1} \Phi_j^{(2)} \left(\sum_{i=1}^{n} \Phi_{j,i}^{(1)}(x_i) \right), \tag{10}$$

对于所有 i 和 j，存在某些连续的单变量函数 $\Phi_{j\,i}^{(1)}$，以及 $\Phi_j^{(2)}$；例如，参见 Braun 和 Griebel [1]。就本文而言，我们更倾向于定理 1 中的版本，因为它与正则均值的联系更直接。

本文最后简要讨论开放问题和未来方向。在我看来，Kolmogorov 叠加定理中的双重加性表示在统计学中的全部潜力仍未得到充分认识。正如本文所示，该定理不仅可以导出整个正则均值类，还能得到其他结果（即推论 2 和 3）。更广泛地说，尽管本文未作探讨，但显然，(10) 与统计建模有直接联系。例如，令 $n = p$，并令 $\mathbf{x} \in A \subset \mathbb{R}^p$ 为协变量向量。如果所有外函数都与恒等函数成比例——也就是说，$\Phi_j^{(2)}(x) = x/(2p+1)$——且 $\Phi_{j\,,\,i}^{(1)} = f_i$，那么 (10) 中的表示就变成加性的，如广义加性模型（GAM） [19]

$$f(\mathbf{x}) = f_1^{(1)}(x_1) + \cdots + f_p^{(1)}(x_p).$$

从这一角度看，基于（10） [例如， 3, 14]的模型可以视为双重加性神经模型，它们自然地扩展了广义加性模型（GAM）的类别。最后，（10）中的叠加定理有望为联合密度、Copula 等其他研究对象的灵活建模奠定基础，这似乎是顺理成章的。

参考文献

[1] Braun, J. 和 Griebel, M.（2009），“关于 Kolmogorov 叠加定理的构造性证明”，*Constructive Approximation*，30，653–675。[2] de Carvalho, M.（2016），“均值，你所说的均值是什么？”，*The American Statistician*，70，270–274。[3] de Carvalho, M.、Ferrer, C. 和 Vallejos, R.（2025），“用于级联极值的 Kolmogorov–Arnold 神经模型”，*arXiv:2505.13370*。[4] Estep, D.（2002），单变量实用分析，纽约：Springer。[5] Fakhoury, D.、Fakhoury, E. 和 Speleers, H.（2022），“ExSpliNet：一种可解释且表达能力强的基于样条的神经网络”，*Neural Networks*，152，332–346。

[6] Hall, P.（2013），自助法与埃奇沃思展开，纽约：Springer。[7] Kolmogorov, A.（1930年），“关于平均值的概念”，意大利林琴国家科学院院刊，12，388–391。[8] Leinster, T.（2021），熵与多样性：公理化方法，马萨诸塞州剑桥：剑桥大学出版社。[9] Lin, J.-N. 和 Unbehauen, R.（1993），“关于 Kolmogorov 网络的实现”，神经计算，5，18–20。[10] Liu, Z.、Wang, Y.、Vaidya, S.、Ruehle, F.、Halverson, J.、Solja˘ci´c, M.、Hou, T. Y. 和 Tegmark, M.（2024），“Kolmogorov–Arnold 网络：Kolmogorov–Arnold 网络”，*arXiv:2404.19756*。[11] Lorentz, G. G.、von Golitschek, M. 和 Makovoz, Y.（1996），构造性逼近：高级问题，第304卷，Citeseer。[12] Markowitz, H.（2014），“期望效用的均值—方差近似”，《欧洲运筹学杂志》，234，346–355。[13] Markowitz, H. M. 和 Blay, K.（2013），风险—收益分析：理性投资的理论与实践（第一卷），纽约州纽约市：麦格劳-希尔。[14] Montanelli, H. 和 Yang, H.（2020），“利用 Kolmogorov–Arnold 叠加定理为深度 ReLU 网络推导误差界”，*Neural Networks*，129，1–6。[15] Morris, S.（2021），“希尔伯特第十三问题：是否存在真正的连续多元实值函数？”，《美国数学会通报》，58，107–118。[16] Ostrand, P. A.（1965），“度量空间的维数与希尔伯特第十三问题”，《美国数学会通报》，71，619–622。[17] Schmidt-Hieber, J.（2021），“重新审视 Kolmogorov–Arnold 表示定理”，*Neural Networks*，137，119–126。[18] Sprecher, D. A. 和 Draghici, S.（2002），“空间填充曲线与基于 Kolmogorov 叠加定理的神经网络”，*Neural Networks*，15，57–67。[19] Wood, S. N.（2006），广义加性模型：*R* 语言入门，佛罗里达州博卡拉顿：Chapman and Hall/CRC。[20] Zaheer, M.、Kottur, S.、Ravanbakhsh, S.、Poczos, B.、Salakhutdinov, R. R. 和 Smola, A. J.（2017），“深度集合”，《神经信息处理系统进展》，30。

经费支持：MdC 获得莱弗休姆信托基金、投资创新中心以及阿威罗大学—CIDMA 的支持；后者由科学与技术基金会 Funda.c˜ao 提供资助（UID/4106/2025 和 UID/PRR/4106/2025）。