arXiv:2305.00965v1 [stat.ME] 1 May 2023

存在并列值时的无偏非参数相关估计量

Landon Hurley，博士

2023年5月2日

摘要

在扩展实数轴上所有含并列值排列的定义域上，定义了 Kemeny 距离的内积希尔伯特空间表述，并由此在齐性独立同分布样本上得到一个无偏最小方差（高斯–马尔可夫）相关性估计量。本文构造并证明了将这一线性拓扑扩展至 Spearman 的 ρ 和 Kendall 的 τ_b所需的条件，表明这两个空间在实际数据定义域上均存在偏差且效率低下。本文还为 Kemeny τ_κ 估计量定义了概率分布，并提供了有限样本的学生化调整。本文由此识别出一种通用线性模型对偶性，它可作为许多有偏和无偏估计情形下唯一的一致解。

相关性估计量一直是线性模型空间数学的基础。基于完备度量空间上负定距离矩阵的构造性基础，人们已经为相关性估计量的许多理想性质提供了保证，例如中心极限定理、无偏性、有限样本可推广性以及概率分布的保证。然而，Pearson相关以研究充分的欧几里得距离或弗罗贝尼乌斯范数函数空间为基础，几乎无处不在；Spearman和Kendall相关性估计量却仍未得到充分理解，研究者也较难获得。本文认为，这是将定义域不恰当地限制为阶数为 n 的对称群 (S_n) 直接导致的；该群表示 n 个数据项在每个变量上均无并列值时，由排列向量给出的排序。

对于从连续随机变量出发的数据分析，以及关于 n 的渐近分析，这一假设都相当有用；在大数弱律下，秩分数双射得以保证。借助生日悖论，可以轻易得出结论：并列值出现的概率——即两个数据元素在某个随机变量上得分相同、但在第二个变量上取值不同的情形——即便样本量很小，也会迅速趋近于0。这一结果直接源于原始双变量定义域具有无限基数，这等价于假设随机变量是连续的。正如

前面已经指出，Wilcoxon 秩和检验和 Mann-Whitney U检验，分别等价于二项分组变量与连续内生响应变量之间的 Spearman 和 Kendall相关估计；实际上，这些检验构成了点二列相关模型框架，并采用不同的度量拓扑（函数空间上的距离函数度量）。本文提出一个统摄性框架，用于构造无分布假设的估计量和检验统计量。这些估计量和检验统计量在结构上是参数性的（随机意义上具有正则性），Hollander 等人（2014，第 1 章）对此亦有阐述。

这些具有高度规律性的无偏最小方差估计量性质，使我们得以确立若干值得特别关注的观察结果，包括：在有限样本的离散数据分析中，常见的非参数估计量既有偏，也不是最小方差估计量；这削弱了 Diaconis 和 Graham (1977) 所提出的部分 Wald 检验的普遍适用性。在新型估计量之间，我们建立了一种与 Pearson 的 r 直接相关的显式几何对偶性，并证明在特定条件下，Kemeny 的 τ_κ 与 Pearson 的r 之间的非线性双射关系仍然成立。这种对偶性极为强大，能够使非正定（n.p.d）矩阵以及相应的、基于 Pearson Σ 矩阵的多变量有偏估计方程，依据满足高斯-马尔可夫条件的经验条件得到唯一解，并如同一个季洪诺夫正则化线性方程组那样收敛至无偏的总体解。我们通过应用验证了这些方法的实用价值，并取得了积极成果，应用领域包括线性因子分析、非随机缺失问题的求解，以及利用高维核估计框架求解一般的 Karush-Kuhn-Tucker 约束优化问题。

1 一种分布无关但具有参数形式的广义线性估计函数

标准广义线性模型的理论要求空间是希尔伯特内积度量空间。这些性质使得在有限个独立同分布样本（即稳定性）下，所有可观测元素都能建立收敛性和正则性。然而，Spearman 的 ρ 和 Kendall 的 τ 估计量并不具备这种正则的有符号内积距离形式，只满足巴拿赫空间（即完备度量空间）的必要要求，而该空间具有渐近收敛性。假设存在一个样本矩阵，其样本独立同分布地取自扩展实数轴，$\mathbf{X}^{n\times p} \in \overline{\mathbb{R}}$，索引为 $i = 1, \ldots, n$ 和 $j = 1, \ldots, p$，且满足限制条件 $n \gg p$。目前，这些假设符合标准惯例；不过，我们要指出，实证研究表明，Kemeny 度量要保持正定且唯一，并不需要较大的比率 $\frac{n}{p}$。用小写向量（如 $x_{i,j=1}, y_{i,j=2} \in X^{n\times p=2}$ ）表示长度为 $n \times 1$、在扩展实数轴上均匀采样，并以矩阵 $\mathbf{x}$ 为索引的单变量向量。

1.1 总体上的一阶矩估计方程

我们提出如下距离函数：它由两个独立生成的反对称矩阵的 Hadamard 内积构造而成，这两个矩阵表示基 κ 上的线性置换空间：$\overline{\mathbb{R}}^n \to n \times n$，其索引为$k$, $l = 1, \ldots, n$，对应于数据矩阵的每一列 $j \in \mathbf{X}^{n \times p}, j = 1, \ldots, p$:

$$\rho_\kappa(x, y) = \frac{n^2 - n}{2} + \sum_{k,l=1}^{n} \kappa_{kl}(x) \odot \kappa_{kl}^{\intercal}(y), \; k, l = 1, \ldots, n. \tag{1a}$$

$$\kappa_{kl}(x) = \begin{cases} \sqrt{.5} & \text{if } x_k > x_l \\ 0 & \text{if } x_k = x_l, \\ -\sqrt{.5} & \text{if } x_k < x_l \end{cases} \tag{1b}$$

$$\kappa_{kl}(y) = \begin{cases} \sqrt{.5} & \text{if } y_k > y_l \\ 0 & \text{if } y_k = y_l, \\ -\sqrt{.5} & \text{if } y_k < y_l \end{cases} \tag{1c}$$

κ 函数将每个扩展实向量映射为一个 $n \times n$ 阶的斜对称矩阵，κ： $\overline{\mathbb{R}}^n \to$（$n \times n$）。上述 κ 矩阵中与向量 x, y 对应的每个元素，分别用第 k^{th} 行和第 l^{th} 列中的元素 κ_{kl}（x_i）和 κ_{kl}（y_i）表示。随后对这两个 κ 斜对称矩阵进行阿达玛乘积（$\odot$），每个矩阵均代表一个双变量对，其内积再对所有 $n \times n$ 元素求和。由有序向量空间的线性组合计算得到的距离，源自对向量 x_i上任意两个元素进行 κ 映射（即置换表示），从而实现对所有含并列值的排列的线性表示；这些排列构成一个基数为 M $= n^n - n$ 的完备空间，而非 $n!$，同时排除所有常数 n 退化随机变量。矩阵转置用上标（·）ᵀ表示。

我们证明了期望值周围存在一个邻域，这等价于证明在阿贝尔线性肯德尔函数空间上测量的任意随机变量都具有一致性和收敛性。令 $\rho_\kappa \in [0, n^2 - n]$ 如方程 1a 所测量。M 中任意空间可能达到的最大 Kemeny 距离由恒等置换 $I_{n\times 1} = 1$、.....、n及其逆序 $I'_{n\times 1} = n$、.....、1 得到；对于所有有限的 n, 其距离始终为 $n^2 - n$，并且在 x、y 上表示扩展实数域上的线性距离（作为度量空间）。 $E(\rho_\kappa(\mathbf{M}) = \frac{n^2-n}{2}$，且任何仿射线性变换，包括从方程 1a 中减去期望值，根据度量空间的定义，仍然是一个完备的希尔伯特空间。此外，所有可度量化空间都被认为是完美正规 T_6 豪斯多夫空间。通过从期望值附近的集合中减去所有距离，Kemeny 度量的邻域现在以 0 为中心带有正负号： $U(\mathbf{M}) = [-\frac{n^2-n}{2}, \frac{n^2-n}{2}]$。该邻域作为 $n^n - n \mapsto [0, n^2 - n]$ 的映射几乎必然有限，是确立 Kemeny 距离一致性的等价条件。因此，

引理 1.1. 对任意有限的 n，Kemeny 度量空间都是紧且全有界的。

证明。 根据引理 A.3，Kemeny 度量是完备的，且全有界，也就是说，它不存在有限的 n 实数点，使用有限正实标量 $0 <$

$a \in \mathbb{R}^{1\times 1} < \infty^{+}$，这超出了 $0 \leq \rho_{\kappa}(a\kappa^{*}(x_A)\ a\kappa^{*}(x_B))\ \leq a(n^2-n)$ $\forall\ 0 < a < \infty^{+}$ 的界，依据引理 A.5，其中 a 是任意有限标量。因此，Kemeny 度量空间既紧致又完备，进而也是可分稠密的拓扑空间。■

Kemeny 线性方差或集中度量，是通过对表示所有有限$\kappa(x_{n\times 1})$有限支撑的 n^2-n 自由参数求和得到的：

$$\sigma_{\kappa}^{2}(x) = \tfrac{2}{n(n-1)}\Big(\sum_{k=1}^{n}\sum_{l=1}^{n}\kappa_{kl}(x)\kappa_{kl}^{\intercal}(x)\Big) \equiv \tfrac{2}{n(n-1)}\sum_{l,k=1}^{n}\kappa_{kl}^{2}(x). \tag{2}$$

根据方差的这一表达式，可定义一个从扩展实数域子空间到奇异实数的映射；当且仅当 x 非退化时，该映射为正。缩放空间体现了平方元素的构成（因此始终非负），这由引理 1.1保证。对斜对称矩阵中的所有平方元素求和时，共有 n^2-n 个自由元素；对角线 n 元素严格为 0，因此方差的最大值为 0.5（与所考察的每个独立随机变量的标量常数$(\sqrt{.5})^2$, 成比例且等价；当对同一变量计算两次时，该值几乎必然不大于 1）。方差几乎必然收敛且有限，均值亦然，这由引理 1.1保证；这是 Kolmogorov 强序性质的推论。由于同一映射的最小值为 0，且在 $\mathcal{M}$上取不到，因此，方程 2 的支撑集为 $\sigma_{\kappa}^{2} \in (0,1],$，适用于扩展实数轴上任意有限长度的所有可能非恒定向量 n。请注意，考虑子集域 S_n时，σ_{κ}^{2} 对于所有元素向量均为常数 1。因此，排列空间中的并列值会降低样本估计的变异性，使得 $\sigma_{\kappa}^{2}(\mathcal{M} \setminus S_n) \in (0,\ 1)$。

由于每个向量 x 均以 κ 矩阵的像为中心，因此在估计方差时会损失一个自由度（中位数）。这就需要引入 Bartlett 校正，以确定在存在并列值时，Kemeny 度量空间上次序统计量的方差。假设采用独立同分布抽样，可得以下表达式

$$\dot{\sigma}_{\kappa}^{2}(\mathcal{M}) = \frac{(n-1)^2(n+4)(2n-1)}{18n}, \tag{3}$$

表示固定长度的所有 Kemeny 距离的总体方差 n，这些距离取自双变量独立随机变量，与表 1中的实证结果一致。请注意，该定义域明确假定每个排列仅抽样一次；不过，这一点将在第 2.4 小节中讨论。

对连续随机变量进行这些修正，自然会产生$n!$此类情况确实可以观察到。我们还注意到，在这些条件下，诸如二项随机变量这样的极限情形，在基于任一度量函数空间构造时，所得方差相同：这使得有限数据的正交归一刻画具有一个基本恒等式

得以实现。不过，目前请考虑，如果不允许存在并列测度，Kemeny 方差函数及其与欧几里得函数空间的相等关系将无法定义。因此，作为统计学习的有效工具，如果不允许并列值，多种常见情形只能渐近无偏（因而得以识别）。

接下来，我们证明基于距离函数构造的估计量是无偏的：

引理 1.2。 Kemeny 相关（方程 *4*）是无偏估计量。

证明。对于 Kemeny 相关所对应的空间，注意，在可数有限总体 M 上，偶函数的内积期望（由任何度量空间关于 U(M) 的对称性所保证）对应的距离为 0，等价地说，相关为 0。对于 $n = 2$，M 中具有对称邻域的 $= 2^2 - 2 = 2$ 排列（由方程 1a 的偶函数性质可知），其距离位于 $[-(\sqrt{0.5})^2, +(\sqrt{0.5})^2]$。对 M 中所有元素 m 求和时，总体期望 $\lim_{m\to\infty^+} E(U(\mathrm{M}) = 0$。因此，Kemeny 估计量在期望意义下渐近无偏。由归纳法，考察有限望远镜序列 $m_i \in \mathrm{M}\ i = 1, \ldots, n^n - n$，由此可得有限期望为 0（这是因为方程 1b 的斜对称矩阵在 Kemeny 度量空间中对加法封闭），且望远镜序列中的正距离与负距离具有对称性；这些性质源于以任意原点为中心的 κ函数为偶函数（见引理 A.8）。由于该空间紧且全有界，所有绝对距离均不大于 $\frac{n^2-n}{2}$，无论扩展实数轴上出现何种取值。对于任意有限 n，总体 M 上由 $m \mid n$ 索引、长度为 n 的两个随机变量 x, y 的和与内积为

$$\mu_1 = \lim_{m\to\mathcal{M}} \frac{n^2-n}{2} + E_m(\tau_{x,y}) = \frac{n^2-n}{2} + \sum_{k,l=1}^{n} \kappa_{kl}(x_m) \odot \kappa_{kl}^{\intercal}(y_m) = \frac{n^2-n}{2} + 0,$$

由此完成证明：Kemeny 相关是中位数双变量距离的无偏估计量。■

通过减去首要中位距离，可将距离从希尔伯特空间的完备度量性质变换到有符号距离空间，从而得到有限子集 m，该子集在 $U(\mathcal{M})$ 的邻域内整体平均为 0，并由 Glivenko-Cantelli定理（定理 1.8) 对该线性函数空间进行一致处理。进一步考察将表明，逼近误差本身关于期望对称分布（因为 Kemeny 度量是偶函数；引理 A.8)，并且根据次可加性（定理A.1)，通过减去无偏期望进行的仿射线性变换不会使估计量产生偏差；对于长度为 n 的随机变量之间的所有有限双变量距离，结论均成立。

1.1.1 Kemeny 相关 τ_κ

将距离函数（方程 1a）转化为内积相关系数，可通过方程 4 实现，

$$\tau_\kappa(x,y) = -\frac{2}{n^2-n}\sum_{k=1}^{n}\sum_{l=1}^{n}\kappa(x)_{kl}\odot\kappa^{\intercal}(y)_{kl} = -\frac{2}{n^2-n}\left(\rho_\kappa(x,y)-\frac{n^2-n}{2}\right), \{x,y\}\in\overline{\mathbb{R}}^{n\times 1}. \tag{4}$$

通过取以 0 为中心的有限、紧致且全有界邻域，将距离 $\rho_\kappa \in [0, n^2-n]$（对于任意 n）及其在长度为 n 的两条扩展实数轴所构成的扩展实数域上的支撑集，转换为一个以 0 为中心的闭且紧致邻域，$U(\mathbf{M}) = [-\frac{n^2-n}{2},\ldots,0,\ldots,\frac{n^2-n}{2}]$。乘以 2 将极值重新缩放，使其覆盖 $\pm(n^2-n)$，同时不改变期望；取负则确保实现的距离越远离原点 0（该原点始终为有限值），随机变量就越接近逆序排列。取负后的极端距离 $\frac{n^2-n}{n^2-n} = 1$ 变为 -1，这支持了这样一种解释：相对于原点，最远距离对应一个所有元素都完全逆序的向量（称为逆序恒等排列向量，并注意原点即为恒等排列向量）。

在 $-(n^2-n)$ 邻域上的尺度化距离，在减去中位距离后，通过比较原点与其自身得到，可重新缩放为 -1；取反后则为 1（由此表示与原点具有相同排序的点，其距离为 0，且余弦相似度的度量值为 $\cos(0) = 1$），这符合相关度量的要求。最后，所有等于中位距离的中点距离均等于 0，而按照规定常数对它们进行的任何仿射线性变换仍为 0；这表示一种赋范相关或距离度量，它在有限域 $\{-1,\ 0,\ 1\}$ 的各点处固定且定义唯一，适用于原点邻域上的所有有限度量。由此便可应用布劳威尔不动点定理，因为等距欧几里得距离及其余弦在其他所有点处均为非线性。

1.1.2 估计量的理论性质

现在我们证明，Kemeny 相关估计量也满足高斯-马尔可夫定理，在以下定义下，它是满足 Lehmann-Scheff´e 定理的最佳线性无偏估计量：

定义 1.1。满足高斯-马尔可夫定理的必要条件，能够保证距离最小化函数在样本上提供可能的最佳线性无偏估计（BLUE）点估计。这五个高斯-马尔可夫条件如下：

1. 线性：估计参数必须是线性的。4. 外生性：随机变量在条件上正交归一。2. 变量通过从共同总体中进行随机抽样，以独立同分布的方式产生。5. 同方差性：给定总体中误差的方差保持不变。3. 不存在完全相关的变量。

定理 1.3. 对于从共同总体中独立抽样得到的任意长度为 n 的双变量向量对，Kemeny 相关是一个高斯–马尔可夫估计量。

证明。 根据定义，对于任意希尔伯特范数空间，Kemeny 函数空间都具有线性。Kemeny 距离及其相关函数（例如方程 4中所示的函数）在希尔伯特空间中已得到证明（另见定理 A.1），其依据是有效内积的存在。无偏性由引理 1.2得出，而格拉姆矩阵正定性则可由满足 Mercer 条件得出，或由 Kemeny 度量空间中全有界正方差的平方和得出（引理 1.1 作用于 $\mathcal{M}$；Schoenberg，1938）；这两个条件等价，因此对 Kemeny 测度空间均成立。利用 Bochner 定理来保证格拉姆矩阵的性质同样有效，因为希尔伯特空间是局部紧阿贝尔群上的连续正定函数。等价地，两个向量之间的相关必须仅取决于它们之间的距离，同时对于假设所定义的任意定义域，几乎处处保持正定（即为有效的距离函数）。Kemeny 距离满足这一性质（方程 4），而移除两个共线点则保证对所有 $n < \infty^+ \in \mathbb{N}^+$总能观测到正定且有限的距离。因此，对于任何非共线且非退化的总体，相关矩阵（因而协方差矩阵）始终正定。

条件 3 根据公理假设得到满足，其中唯一点I，$I' \setminus \mathcal{M} \equiv Q$（希尔伯特空间上的 Riesz 表示定理保证了唯一性，该空间建立在具有单调仿射线性不变性的拓扑函数空间之上）。然而，对于 Kemeny 邻域上的任意有限 n 随机抽样，仅存在 2 κ 映射$\mathcal{M} \setminus Q$，它们对于任意一对独立随机变量都是共线的。观测概率（另见定理 1.7）随后在关于 n的极限下迅速趋于 0：

$$\Pr(\mathcal{M} \cap Q) = \lim_{n \to \infty^+} \frac{|Q| = 2}{n^n - n} = 0.$$

外生性进而可由任意希尔伯特空间的 Riesz 表示定理推出，因为在该定理定义域的公理假设下，独立随机抽样随之成立，使得 $\lim_{n \to \infty^+} \frac{2}{n^n - n}$。同方差性断言根据 Kemeny 方差的定义得出 $0 < \sigma_\kappa^2 < \infty^+$，因此只有线性可比较的得分才能进行排序，因为对于任一共同总体函数（累积分布函数），只能存在一组排列。因此，这排除了在周期函数上进行唯一测量的可能性；根据公理假设，这是成立的，否则必定存在

扩展实数轴上的实现值可能无法通过方程有效评估 1b。由此证明 Kemeny 相关满足高斯–马尔可夫估计量的所有必要条件。 ■

每个变量的非负定性由斜对称矩阵的平方给出（其转置取负后再平方），只有在 n 退化样本分布时才会得到 0，因此在其他情况下均不考虑 M 上的情形。任何此类双变量数列中 $n^2 - n$ 这些数的总和，是斜对称矩阵所导致的结果；该矩阵的对角线为 n 0，乘以 $\sqrt{0.5}^2$后几乎必然落在区间 $(0, \ldots, n^2 - n]$ 内。根据度量空间的性质，该值几乎必然为正；对于所有有限的 n，也显然可以确认其有限。

1.2 Kemeny 距离及所有仿射线性变换的概率正则性

不失一般性，假设 p 的扩展实向量空间可由正定方差-协方差矩阵 $\Xi_{p\times p}$表示。根据对任意非常值随机数列（即变量）观测到正有限方差的定义，并结合 $U(\mathcal{M})$ (方程 2) 上有符号距离测度在 0 附近的紧支撑，距离在 $U(\mathcal{M})$ 上的分布满足成为次高斯分布的必要条件（Buldygin & Kozachenko，2000，第 1 章）。

现在讨论最小必要矩序列的必要性扩展（即， n不是充分的总体参数）。当且仅当满足以下条件时，次高斯变量 ξ 称为严格次高斯变量

定义 1.2。

$$E(e^{\lambda\xi}) \leq e^{\frac{\lambda^2\sigma_\xi^2}{2}}, \forall\ \lambda \in \overline{\mathbb{R}} \equiv \|\xi\|_\kappa \leq \sigma_\xi = \|\xi\|\ell_2(\Omega). \tag{5}$$

定理 1.4。 对于任意有限样本*1.2*，Kemeny 距离的分布（定义 $n < \infty^+$）都是严格次高斯的，因此具有稳定性。

证明。 对于所有有限的 n，Kemeny 距离的分布定义在邻域 $U(\mathcal{M})$ 上，因此几乎必然有限且可测（完美正规希尔伯特空间的存在确保了 Borel σ-测度的存在）。所有可能观测到的方差均已证明具有有限确定性（引理 1.1），其定义域为 $\mathcal{M}$。次高斯变量定义为有限方差与紧致全有界支撑的组合，因此对于有限的 n (Buldygin& Kozachenko,2000，第 1 章），Kemeny 距离是次高斯的。因此，无论是双变量还是单变量情形，我们都观察到次高斯分布：在单变量情形中，距离度量的谱为 $(0,\ b],\ b \in (0,\ \ldots\ldots,\ \mathbb{N}^+,\ \infty^+)$，并满足平凡限制

使 $n > 0$ 能够允许 $b \neq 0$，且在仿射变换 0 下的期望满足条件 5。对于双变量情形， κ 反对称矩阵 $\{\kappa_x, \kappa_y\}$ 的向量内积之谱也是有限且全有界的，因此对于任意有限的 n，它也严格次高斯，因为其距离几乎必然小于或等于欧几里得距离 ℓ_2范数，且该关系在整个 $U(\mathcal{M})$ 上一致成立。因此，对于所有有限的 n，边缘分布和双变量分布都严格次高斯。

对于逆否条件，考虑当有限的 n 时，总体的 Kemeny 距离分布为高斯分布的情形。若确实如此，则将存在一些可测事件：它们在高斯概率密度函数上以非零密度发生，但对于任何有限的 n 都无法观测到：

$$\lim_{n\to\infty^+} \Pr(x \backslash U(\mathcal{M}) \in \pm\tfrac{n^2-n}{2}) = \int_{\infty^-}^{-\frac{n^2-n}{2}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x - \int_{-\frac{n^2-n}{2}}^{\infty^+} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x \geq 0,\ x \in \mathbb{R}. \tag{6}$$

因此，对于所有排列总体，必然存在在高斯概率分布上可观测、但不出现在 Kemeny 距离支撑集上的距离；所以，只有在极限情况下才能取得相等，此时等式成立且值为 0。因此，该分布在定义域 $x^{n\times 1}$、$y^{n\times 1} \in \overline{\mathbb{R}}$、$n < \infty^+$ 上并非高斯分布。■

推论1.4.1。如果 Kemeny 距离几乎必然严格为次高斯，则我们观察到该分布以 0 为中心，并具有对称的密度尾部，其密度小于或等于标准正态分布的密度。因此可以立即得出，对于大小为 M 的任意总体，四阶矩足以刻画该概率分布。

证明。 显然，0 的期望的任意次幂也为 0，因此，对于 Kemeny 度量，所有高阶奇数阶矩均为 0（从而也确立了其在所有m上的分布具有偶对称函数的性质）。二阶中心矩在方程 2 中定义，因此是关于期望 0 的正交归一量；根据距离的对称性（引理 A.8），偏度为 0，因此最后需要考察的自由矩是 μ_4 有限样本上的超额峰度。然而，Kemeny 分布严格次高斯，因此对于任何有限 n，其超额峰度必为负值，因而不可能服从正态分布。因此，这种独特的概率分布具有对称性且无偏，其所有得分的离散程度和度量几乎必然为正且有限；它不具有偏度（由偶函数性质可知），并且具有有限的负超额峰度，该峰度从下方渐近趋于 0（Buldygin & Kozachenko，2000，第 1 章）。这与外推得出的断言结论相矛盾：渐近性能适用于有限样本（即具有稳定性），因为在有限样本中存在并列值时，有限样本距离的分布（严格次高斯）不可能服从正态分布。■

文中指出一个有趣的推论：Kemeny 距离和相关函数与 Kendall τ 距离及仿射

线性函数（包括相关性）有关。因此，Kendall τ 函数空间是更大的完备 Kemeny 空间的真子集。

推论 1.4.2. 假设对于 $n > 0$，相应度量下排列空间的密度分别为：Kemeny 度量空间为 M $= n^n - n$，Kendall 度量空间为 M′ $= n!$，其中 M′ $\subset$ M。由此可知，数值 M 必须能由阶乘的斯特林近似充分表示

$$\sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}},$$

否则 M′ ⊂M。

证明. 假设 M $\subseteq$ M′，这表示 τ-距离空间的密度大于或等于 Kemeny ρ_κ-空间的密度。若确实如此，则可推出 $2n \leq \log(2\pi n)$（完整展开见引理 B），但在所有条件下，这一结论均因矛盾而严格不成立，因为不存在满足这一严格不等式的 $n \in \mathbb{N}^+$。由此可知 M′ $\subset$ M，因此关于 Kendall 距离的所有度量结果都是 Kemeny 距离的严格子集，$\tau(x, y) \subset \rho_\kappa(x, y)$。由于含并列值的排列必然严格包含所有不含并列的排列，因此集合差必然为正 $\lim_{n\to\infty^+} n^n - n - n! > 0$; 这一关系对所有$n > 2$均一致成立，从而确保含并列值的所有排列所构成的集合空间密度几乎必然更高。■

推论 1.4.3. 两个距离函数的线性比例关系在子集 x 上成立，其中 $\sigma_\kappa^2 = \frac{n^2-n}{2}$ 是 $n!$ 个无并列值的此类置换的集合。对于该集合，注意到 $\tau(x,\ y) \in [0, \frac{n^2-n}{4}]$，而 $\rho_\kappa(x,\ y) \in [0, \frac{n^2-n}{2}]$；并且对于每个置换，$\tau(x,\ y) = \frac{1}{2}\rho_\kappa(x,\ y)$。

证明。由于对于 $a \in \overline{\mathbb{R}}^+$[1]，$\tau$函数的单位距离与其成正比，因此可直接得到：当 $x = \{a, b\} \xrightarrow{\tau} y = \{b, a\} = \tau(x, y) = 1a.$ 时。然而，对于 $x = \{a, b\} \xrightarrow{\tau} y = \{b, a\} = \rho_\kappa(x, y) = 2a$，。这表明，Kendall 度量上的距离 1 是 Kemeny 度量上的距离的一半；按比例重新缩放后，由 Kemeny 度量衡量的 S_n 的最小非零距离为 1，反映了并列值出现时的测量情况，而这种情况同样无法观测到。由于在 τ 距离定义域（S_n）上不存在并列值，相对于全序而言，并列值对应的最小距离不可测，这与推论 1.4.2 一致。因此，存在仿射线性双射 $\tau(x, y) = 1a \propto \rho_\kappa = 2a \ \therefore 2 \times \tau(x, y) = \rho_\kappa(x, y)\ \forall (x, y) \in U$(M)，假设两个随机变量均为非退化且具有相同的有限长度 n。因此，每个 τ 距离都能在 ρ_κ 距离上得到有效且唯一的定义，反之则不然，因为对于并列值可能出现在 Kendall 的 τ 上的测度空间，不存在 Borel σ 映射。因此，Kendall 的 τ 总体是 Kemeny U(M) 的真子集。■

[1]注意，无需明确指定标量常数 $a = |\sqrt{.5}|$。

注释 1。请注意，根据 $\tau = \frac{n^2-n}{2}$ 和 $\rho_{\kappa} = n^2 - n$ 的定义支撑集，这一点更容易等价地得到确立：它们具有共同的极值（及原点）；对于相同的随机变量对，距离唯一地定义为 0，而双变量排列在其他所有距离上都按标量 2 成比例。由 $n = 2$ 归纳；在 n 上应用 $\mathbf{M} \cap S_n$，即可立即确立共同距离的比例关系。

对于这个线性函数空间，我们由此建立距离估计量和相关性估计量的渐近正态性：

引理1.5。双变量严格次高斯 Kemeny 度量空间渐近地 $wrtn$ 服从正态分布。

证明。考察贝塔-二项分布下 Kemeny 距离函数的概率分布的峰度。作为 n 的函数，表示双变量独立随机变量之间 Kemeny 距离的随机变量函数显然是单调递增函数。根据严格次高斯性的定义，对于所有具有渐近正态性的有限 n，超额峰度始终为负（因此在 n 的极限下超额峰度为 0）。由于所有奇数阶矩均为 0 且方差为正，单调收敛的峰度在 n 的极限下从下方趋近于 3。这使其在所有有限样本量下保持严格次高斯性；在 $n \to \infty^{+}$ 的渐近极限下则退化，并且对于 M 的总体中的所有 n 均无偏。

反证法确保集合 $U(\mathbf{M})$ 的渐近正态性。在总体极限 $n \to \infty^{+}$ 下，该集合的基数偏离实数直线，且不存在有限距离或基数的集合，如 $U^{*}(\mathbf{M}) \supset U(\mathbf{M})\ \forall n$。将上述元素比率映射到邻域的有限支撑集上，满足 Kolmogorov 强序性质的上极限，但也会在无限集合上产生退化界。在该集合中，无限元素集合上的距离不再能够相对于唯一原点 I_n 按有限距离进行唯一排序，而该原点被视为等价于大小为 $0 = (U^{*}(\mathbf{M}))^{-1}$ 的概率测度，因此无法稳定地满足严格次高斯性质。因此，在秩与分数测度距离之间的严格双射关系下，高斯分布足以且是唯一能够表示该线性函数空间的分布。在总体中，根据大数弱律，可以观察到一个完全高斯的随机变量，并且秩与得分的线性双射几乎必然成立。因此，发散的 Kemeny 距离与高斯概率密度函数及其积分呈双射共线关系；一旦经过中心化和尺度变换，共同分布上的两个测度空间便可由欧几里得距离唯一确定。■

虽然 Kemeny 度量空间的邻域严格来说是次高斯的，且仅渐近正态分布，但基于平方向量矩阵刻画（即映射 κ 的刻画）的运算函数空间在有限样本中存在实现困难（尽管我们通常可以作此假设，但

<escape>（很少会验证渐近正态分布在有限样本下是否也呈正态分布）。这种关系将在后文第 2.4 节中进一步阐述。</escape>

引理 1.6。对于任意一对独立同分布的独立随机变量，Kemeny 度量作为线性距离函数满足大数强律。

证明。根据马尔可夫不等式，有限期望的存在意味着满足大数强律；对于 Kemeny 度量下的任意有限 n，这一点都有保证，因为 Kemeny 距离函数具有全有界性，由此可得 $E(I\{\kappa(x_i) \leq \kappa(x)\}) < \infty$，因而对所有有限 n 均成立。该条件等价于证明 $\sup \sqrt{(\pm\frac{n^2-n}{2})^2} \equiv \frac{n^2-n}{2} \equiv \sup \rho_\kappa(x,y) \ \forall \ n \in \mathbb{N}^+$，这一点很容易验证。第二个条件显然成立：对于任意有限向量序列 $\{x_i\}_{i=1}^n$，都可以得到一个线性序，例如使用从扩展实数轴上独立同分布抽样的任意随机变量的 κ 映射的像。因此，对于任何能够对 U(M) 上的有限样本有效测量 $\kappa(\cdot)$ 的情况，大数强律均成立。■

因此，根据大数强律，我们可以确保线性函数渐近收敛于真实分布，并且对所有有限的 n，由于其分布具有严格的次高斯性质，该收敛稳定成立（定理 1.4）。此外，对于每个 n，我们可以将独立随机向量 x 在 $\mathcal{M}$ 中的观测划分开来，这是由在孤立状态下条件观测到的正交关系所致（即假设独立同分布）。因此，得到 F_κ 的累积分布函数的一个无偏估计量；引理 A.13 和引理 A.14 表明，无论非线性评分关系如何，通过 Kemeny 度量空间的 σ_κ-代数上的全变差距离，该估计量都依大数强律收敛于真实分布。然而，该分布并非显而易见；不过，通过经验矩匹配，我们得出，对于所有有限的 n，Kemeny 距离的分布均为贝塔-二项分布。该概率分布的复合性质是一个常见问题；不过，由于其支撑集几乎必然是紧且全有界的，且概率正则，因此，对于任意样本，都可以通过乘以 $U(\mathcal{M}) \cdot$ 上确界 $U(\mathcal{M})$来消除分布的二项分布性质，从而将支撑集整体归一化至 0 到 1。

1.3 Kemeny 参数空间的充分性

给定一组在未知参数向量 θ 条件下独立分布的数据 x，充分统计量是一个函数 $T(x)$，其值包含计算参数任何估计值所需的全部信息。根据因子分解定理，对于充分统计量 $T(x)$，概率密度随后可表示为 $f_x = h(x)g(\theta,\ T(x))$；未知参数向量 θ 由原始数据空间上的函数组成，这些函数是充分的，在

对于给定统计量$t = T(x)$的条件概率分布，数据 x满足$(x \perp \theta \mid t)$。如果一个统计量与 θ 之间的互信息严格等于 x 与 θ 之间的信息（根据平方函数的单调性，所有不相等的元素都严格小于），那么该统计量就是充分统计量。因此，将证明充分统计量$\mu_\kappa = \sum_{j=1}^{n} \sum_{i=1}^{n} \kappa_{ij}(x)$对于任何在实数集上独立分布的分布的集中趋势都是充分的。

此外，满足 $\mu_\kappa^1(x) = 0$ 对于任意 $x \in \overline{\mathbb{R}}^{n\times 1}$，可得到一个充分且无偏的位置参数估计量，使得 $E(\kappa_i(x)) = E(\kappa(x)) = 0 \,\forall\, \theta \implies \kappa(x) = 0 \,\forall\, \theta,$ 几乎处处成立（引理 1.5）。因此，Kemeny 度量满足 Lehmann-Scheff´e 定理关于完备性的必要条件，因此它也是最佳无偏估计函数，这一点已由定理 1.3证明。充分性则由 Kemeny 测度空间上的渐近正态性推出；对于所有 n，则由严格次高斯性推出。

定理 1.7。 设 $\{x_1, \ldots, x_n\}$ 是一个大小为 n 的随机样本，其分布的期望值为 μ，方差为 σ^2。根据 Kemeny 度量函数空间上的大数定律（引理 *1.6*），样本几乎必然收敛到期望值 μ，相对于 n的极限如 Lindeberg-L´evy 中心极限定理所示：

假设 $\{x_1, \ldots, x_n\}$ 是一列独立同分布的随机变量，且 $E[x_i] = \mu$ 和 $\mathrm{Var}[x_i] = \sigma^2 < \infty$。那么，当 n 趋于无穷大时，随机变量$\sqrt{n}(\bar{x}_n - \mu)$ 在分布上收敛于正态分布 $\mathrm{N}(0, \sigma^2)$:

$$\sqrt{n}\left(\bar{x}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

证明。要将（广义）中心极限定理有效地应用于严格次高斯 Kemeny 度量，必须满足三个条件。首先，对于任意有限的 n，以 $\lim_{n\to\infty^+} \frac{1}{n} \equiv |\mathbf{M}|^{-1} = 0,$ 替代，这是无原子概率测度所必需的转换。对于有限的 n，在 n 个独立抽样的观测值下，所有允许重复的排列向量构成的总体并非无原子总体，而是包含 $|\mathbf{M}| = n^n - n$ 个元素；具体而言，排除所有长度为 n 的退化但可观测的常向量。这将得到一个概率为$\frac{1}{|\mathcal{M}|}$的完备测度空间，其中不可观测的概率趋于 0，M 上不可测点被观测到的概率也几乎必然趋于 0，因此整个非退化扩展实数域均为线性可测：

$$1 - \lim_{m\to\infty^+} \Pr(\mathcal{M} \setminus \kappa(x)) = 1.$$

马尔可夫不等式对此进行了确认，该不等式表明，在映射函数 κ下不存在任何不可测有限向量；忽略退化分布时，对于任意长度递增的序列，该结果趋于 0 n。其次，令 μ_κ^1 存在，这一点已在 Kemeny 度量的引理1.2 和引理A.6 中得到证明。最后，令任意总体的方差均为有限值，则

这在 Kemeny 度量空间上同样得到证明（引理1.1 和A.7）。因此，根据这三个条件可以得出，中心极限定理是 Kemeny 度量空间上的一个有效构造，从而确保任意 p 未知但已确定的线性估计目标均几乎处处分布中的线性收敛；对于有限的 n，它们是严格服从次高斯分布的变量。■

在 Kemeny 度量空间上，利用 Lyapunov条件的中心极限定理（推论A.14.1）和 Kolmogorov 定律（推论 A.14.2）的等价结论见附录。

定理 1.8. M 上的 Kemeny度量函数满足 Glivenko-Cantelli 定理：设 $\{x_i\}_{i=1}^{\mathcal{M}}$ 是一列独立分布的均匀随机变量，其分布函数为 $F \in \overline{\mathbb{R}}$。那么

$$\sup_{x\in\overline{\mathbb{R}}} |\hat{F}_m(x) - F(x)| \to 0, a.s.$$

证明。对于任意 $\epsilon > 0$，有

$$\lim_{m\to\infty} \sup_{x\in\overline{\mathbb{R}}} |\hat{F}_m(x) - F(x)| \leq \epsilon, a.s.$$

根据引理 A.11，存在一个划分指标 j，使得 $t_j \leq x < t_{j+1}$，满足：

$$\hat{F}_m(t_j) \leq \hat{F}_m(x) \leq \hat{F}_m(t_{j+1}^-) \wedge F(t_j) \leq F(x) \leq F(t_{j+1}^-),$$
$$\implies \hat{F}_m(t_j) - F(t_{j+1}^-) \leq \hat{F}_m(x) - F(x) \leq \hat{F}_m(t_{j+1}^-) - F(t_j) \equiv$$
$$\hat{F}_m(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-) \leq \hat{F}_m(x) - F(x) \leq \hat{F}(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j)$$
$$\therefore \hat{F}_m(t_j) - F(t_j) - \frac{\epsilon}{2} \leq \hat{F}_m(x) - F(x) \leq \hat{F}_m(t_{j+1}^-) - F(t_{j+1}^-) + \frac{\epsilon}{2},$$

根据大数强律，该值趋于 0 处的等式。因此，对于任何有限样本（也就包括可数样本），任意扩展实数分布的秩序在 Kemeny 度量下都满足 Glivenko-Cantelli定理。■

1.4 哈尔测度

哈尔测度的存在使我们能够定义可容许程序，从而建立最优不变决策准则。更实际地说，它们使我们能够利用巴拿赫赋范空间存在所保证的概率结构，例如对频率学派推断和贝叶斯推断均可容许的优化泛函不变变换。尤其是，哈尔测度使统计学中先验概率的构造和条件推断成为可能（进而能够扩展到包括多元回归在内的函数空间）。因此，我们在此基于 Kemeny 度量空间确立若干数学基础，包括拉东导数的存在与有效性，以及在样本而非总体的不变体积上进行积分。这两项性质是统计分析的基础，如今也证明它们能够有效地应用于 Kemeny 函数空间。此外，我们注意到，根据哈尔测度的性质，由于 M

及其函数均为离散的，依据选择公理，可将可数可加的左不变测度完整定义在定义域的所有子集上。

作为哈尔测度的函数定义为唯一的可数可加、非平凡测度 μ，定义在 G 的博雷尔子集上，并满足以下性质：

定义 1.3.

1. 测度 μ 在左平移下保持不变：对每个 $g \in G$ 和所有博雷尔集 $S \subseteq G$，$\mu(gS) = \mu(S)$。

2. 测度 μ 在每个紧集上均为有限：$\mu(K) < \infty^+$ 对所有紧 $K \subseteq G$

3. 测度 μ 在博雷尔集 $S \subseteq G$上是外正则的：

$$\mu(S) = \inf\{\mu(U) : S \subseteq U\}.$$

4. 对于紧致的 K $U \subseteq G$，测度 μ 在开集上内正则：

$$\mu(U) = \sup\{\mu(K) : K \subseteq U\}.$$

G 上满足这些条件的测度称为左哈尔测度；这是右哈尔测度存在且成比例（因而等价）的充要条件。

引理1.9. Kemeny 度量空间满足*1.4* 中对拉东测度空间的定义，从而证明拉东导数的存在性。

定义 1.4. 如果 x 是一个豪斯多夫拓扑空间，那么 x 上的拉东测度是 x 上的一个博雷尔测度 m，满足 m 局部有限，且在所有博雷尔子集上内正则。

证明。 Kemeny 度量是一个希尔伯特度量空间（由方程1a可知），因此是一个 T_6，或完全正规空间、豪斯多夫拓扑向量空间。因此， $\kappa(x)$ $x \in \overline{\mathbb{R}}^{n\times 1}$ 必定既局部有限又内正则。根据 Riesz 表示定理，所有度量空间在开集上都是内正则的 K。根据引理1.1，度量空间总是局部有限的。因此，Kemeny 度量上的有限博雷尔测度是紧的（按 Bogachev 的含义，2007 定理 7.1.7），且在 Kemeny 测度上存在一个拉东导数。■

引理1.10。 Kemeny 度量空间满足定义 *1.3*的所有性质，因此是一个建立在所有从扩展实数轴独立同分布抽取的随机独立变量之上的哈尔测度空间。

证明。 Kemeny 度量及其 Borel σ-代数在加法和乘法下封闭，因此具有左平移不变性。引理 1.1 中的全有界性保证测度 $\kappa(K)$ 对所有 $K \subseteq \overline{\mathbb{R}}$都有限。引理 1.9 证明了内正则性；外正则性和完备性则由希尔伯特空间的性质推出。■

表 1：有限 n 时 Kemeny距离函数的分布特征：对 $n \leq 8$ 穷举排列，并对所有更大的 n 抽取 3,294,172 个样本。

n	μ_1	σ_κ	超额 μ_4	n	μ_1	σ_κ	超额 μ_4
2	0.000	0.707	-1.875	25	0.002	42.647	-0.091
3	0.000	1.610	-1.171	26	0.019	45.183	-0.083
4	0.000	2.646	-0.747	27	-0.066	50.477	-0.080
5	0.000	3.795	-0.548	28	0.040	53.177	-0.076
6	0.000	5.046	-0.432	29	0.040	55.900	-0.075
7	0.000	6.392	-0.356	30	0.046	58.674	-0.072
8	0.000	7.826	-0.302	31	0.011	61.506	-0.069
9	0.000	9.345	-0.259	32	0.043	64.418	-0.066
10	0.006	10.939	-0.230	33	0.014	67.287	-0.061
11	0.009	12.622	-0.212	34	0.082	70.272	-0.062
12	-0.007	14.352	-0.191	35	-0.016	73.262	-0.065
13	-0.017	16.168	-0.173	36	-0.036	73.262	-0.057
14	0.006	18.064	-0.161	37	-0.005	76.255	-0.063
15	-0.010	19.996	-0.148	38	0.010	79.419	-0.060
16	-0.025	22.005	-0.141	40	-0.035	85.764	-0.058
17	0.001	24.066	-0.131	45	-0.052	102.107	-0.052
18	0.010	26.216	-0.122	50	0.035	119.342	-0.043
19	-0.021	28.386	-0.117	55	0.075	137.645	-0.039
20	0.025	30.645	-0.105	60	0.057	156.656	-0.036
21	0.006	32.942	-0.105	62	-0.100	164.530	-0.039
22	0.008	35.272	-0.096	64	-0.020	172.447	-0.039
23	0.031	37.694	-0.096	68	-0.032	188.808	-0.030
24	-0.012	40.155	-0.095	75	0.042	218.527	-0.026

已知 Kemeny 距离关于 $U(\mathcal{M})$ 的分布函数服从贝塔-二项分布，因此超额峰度是 n的显式函数，无需单独估计。相反，对于关于 n的渐近极限，该分布的非正态性通过对称分布收敛于标准正态分布而得到解决，这是由贝塔分布的指数族特征所致（另见引理 1.5）。另一种证明方法是指出该概率分布服从贝塔分布，而贝塔分布属于指数族，因此在独立同分布抽样条件下必然几乎必然收敛于正态分布。只关注贝塔分布（并可放心忽略二项分布分量，因为它仅用于离散化连续概率流形，而这是希尔伯特空间的必要特征）的合理性将在第 2.2 小节中讨论，以构造最大似然估计。

1.4.1 任意均匀抽样变量下的 Cochran定理

如果我们接受贝塔-二项分布对于将支撑集限制在 κ 函数的区间内、进而限制 Kemeny 度量函数（引理 1.1）是必要的，那么方差的单变量分布也必须是中心 $\chi^2_{\nu=1}$ 分布，并使用第一类修正贝塞尔函数 I_v，其形式为：

$$f(x \mid \sigma_\kappa, n) = \frac{1}{2}\left(\frac{x}{\sigma_\kappa^2}\right)^{\frac{M-1}{2}} \exp\left(-\frac{x+\sigma_\kappa^2}{2}\right) I_v(\sigma_\kappa\sqrt{x}),\ x = \kappa(x) \qquad (7)$$

引理1.11. Kemeny 估计量在均匀抽样下满足 Cochran定理。

证明。 为证明 Kemeny 估计量在均匀抽样下满足 Cochran 定理，考虑任意定义域空间 κ^2 中的取值$\mathcal{M}$，其中方差严格大于零且小于一，并且对所有 m ，一阶期望始终为 0（由引理 1.2）。据 Chernoff 界，可在截断点 $a = \frac{n-n^2}{2}$，$b = \frac{n^2-n}{2}$处得到累积分布函数的已知尾部，这些点定义了 $U(\mathcal{M})$ 对所有有限n的支撑集。由于 κ 函数是完备度量希尔伯特空间的线性基，它在加法和乘法下均封闭，因此其和也允许 χ^2 变量具有同样的分布。之所以接受这一点，是因为定义域 x上的线性加法 y 必然具有由方程1b给出的相应余像映射，该映射在线性和平移单调变换下保持不变；Semrl（1996）已证明了这种扩展所需的相应条件。因此我们得出结论：对于 Kemeny 测度空间上长度为 n 的随机抽样变量，只要断言所有得分均可线性排序（例如来自指数族分布，或其任意离散化），Cochran 定理便成立。由近似线性的投影函数希尔伯特空间的正态性，我们可直接推出该 χ_1^2 分布成立；这使我们能够毫无障碍地对离散均匀分布采用 Chernoff 界，完全基于此前已证明的严格次高斯随机变量的广义中心极限定理，并且只需假设对观测到的可线性排序分布进行均匀抽样时相互独立。■

由此可知，对于可在共同总体上排序的独立随机变量，Kemeny 度量的有效应用导出引理 1.11。根据这一结果，我们得到以下带符号的 Wald 检验统计量，其对应的标准化贝塔-二项分布在渐近意义下为正态分布（当 $n \to \infty^+$ 时成立），且对所有有限的 $\mathcal{M}$ 均无偏（分别依据贝塔分布的指数性质和引理 1.5所述的渐近正态性）：

$$z_\phi(x, y) = \frac{-\rho_\kappa(x, y)}{\sqrt{\dot{\sigma}_\kappa^2}}, \tag{8}$$

表示中心化 Kemeny 距离的比值 ρ_κ，$E(\rho_\kappa) = 0$，该比值以$\mathcal{M} \mid n$上的标准差进行标准化。显然$(z_\phi)^2 \sim \chi^2_{df=1}$。由于样本量这一常数已知函数只有一个自由度（由方程3给出的方差），因此该分布是标准正态 z 统计量，用于估计从共同总体中抽取的有限样本中两个向量之间的 Kemeny 距离（注意，这里使用的是标准差而非方差，因为分子是有符号度量，而不是各距离相对于唯一原点的方差）。更一般地说，尽管本文未作探讨，Cochran定理（引理1.11）允许应用一般线性模型理论（例如，对期望上无偏的有限样本分布无关估计量使用多元回归），其方法是确立条件多元分布的唯一识别，并通过独特的贝叶斯方法实现有效近似。

在扩展域$\mathcal{M} \setminus\ S_n > 0$上，Kendall 的 τ_b 在所有有限样本中都有偏，因此不是最小方差估计量。表3、4和5中的实证模拟证实了这一点；此外，理论证明表明，当 $x^{n\times 1}$，$y^{n\times 1}$ $\not\in S_n$时会出现不可测事件。这些表展示了 Kemeny 相关（及其产生的 Wald 统计量）在离散有序数据上的预期理论最小方差性质，表明既有的替代估计量在 Kendall τ 估计量下会产生过度分散（且正偏）的实证结果。因此，对于给定的 α阈值，将其应用于非连续数据（例如二元有序有限响应集或计数）时，I 类错误必然会膨胀。常见示例既包括行为评估中常见的序数量表项目，也包括分级反应模型，例如项目评审；这些示例表明，线性距离相对于弗罗贝尼乌斯范数或 Kendall 范数并非常数，而相对于 Kemeny 范数则为常数。因此，在所有情形下，实证上均更倾向于使用 Kemeny 相关估计量而非 Kendall 相关估计量：在有限的连续情形和渐近情形下（此时不会出现并列值），其分布与 Kendall 的 τ 相同；但在其他情况下，它更准确且效率更高。文中所引用的所有实验模拟，均使用 Kendall τ_b 相关估计及检验统计量的方差渐近近似 z_b。

2 非参数肯德尔相关的估计方程

前述推导给出了 Kemenyρ_κ 距离泛函拓扑框架的解析解和证明，该框架进而提供了赋范相关。与此同时，也给出了相应的 Wald 检验统计量，并证明了在若干标准二元分析情形下，Kemeny 估计具有无偏性和最小方差特性（即任何常见的二元总体，只要其元素可依据累积分布函数在扩展实数轴上排序）。用于构造 Wald 统计量的随机变量两两距离的贝塔-二项分布，也适用于最大似然估计。这使我们能够确认：原则上，与高斯分布等价，Kemeny 度量是非参数二元相关的线性最大似然估计量，同时对于范围更广、经独立抽样的累积分布函数集合，它也是无偏最小方差估计量。这里还指出了 Kemeny 相关与 Spearman 相关之间对偶性这一重要特性；在适当的参数限制下，这种对偶性还可进一步推广至 Pearson$r$ 。最后，我们提出了一个对 Kemeny 相关空间进行 Student 化的框架，从而能够构造通用的 Student t 检验估计量、Welch t 检验估计量和配对样本 t 检验。

2.1 矩估计量

已知总体特性后，可通过矩估计法估计贝塔分布的形状参数 α_1 和 α_2。该分布由距离的贝塔-二项总体除以 $N = n^2 - n$ 得到，从而进行归一化；对于任意有限大小的样本，距离都成为实区间 $\frac{\rho_\kappa}{N} \in [0, 1]$ 上紧且全有界的元素。

备注 2。注意，在考察 B 分布（即 Beta 分布）时，经验观察表明，中心极限定理（以强大数定律而非弱大数定律的形式）仅对 $n \geq 9$ 有效，正如前文所述。对于每个 $n \in 1, \ldots\ldots, \mathbb{N}$，通过将对称 B 分布的方差与支撑集之比进行比较，我们注意到，当且仅当 $n \geq 9$ 时，方差才大于 1：

$$\begin{aligned} E[x]_\kappa &= \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{n}{2n} = \frac{1}{2} \mid \alpha_1 = \alpha_2 = n \\ \operatorname{var}[x]_\kappa &= \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} = \frac{1}{8n + 1} \end{aligned} \tag{9}$$

使得 $(\frac{n^2 - n}{8n+1}) > 1 \iff n > 9$. 这在限制较少的条件下唯一地定义了标准伪中位数构造（Hollander 等人，2014，第 39 页），从而允许在扩展实数域上对所述位置参数构造最小方差无偏估计量。这进一步表明，为使中心极限定理适用于单峰分布，方差必须小于测度空间的支撑集，从而也使支撑集的最稠密划分能够在单调收敛空间上得到唯一确定。这一有偏或约束优化问题（通过 KKT 估计唯一求解）的影响将在后续工作中加以讨论和发展。

另请注意，对于 $\lim_{n \to \infty^+}$ （定理1.4），Kemeny 估计量的贝塔分布是渐近正态的；而在其他情况下，对于所有有限样本，它仍构成一个闭且全有界的偶函数空间，这支持了这种概率分布的选择。贝塔分布的指数特性也证实了这一点，由此可推出估计方程的渐近正态性。我们还注意到，考察比率 $\mathcal{M}\ n!$ 时，可以观察到变化率在约 $n \geq 15$. 处明显呈非指数形式。具体而言，请注意,—在该比率下，标准中心极限定理阈值 $n \geq 20$ 处的密度，在 log(log(n)) 尺度上近似呈线性。

在贝塔分布的估计方程中求解自由参数 α_1， α_2；其中，随机数据向量（变量）(x， y)的 n， ρ_κ、样本量以及双变量 Kemeny 距离均为已知量，且满足识别约束 α_1， α_2， n， $\rho_\kappa > 0$。这些条件与既有假设一致，且不会在证明高斯-马尔可夫条件所施加的假设之外进一步限制解空间。

利用贝塔-二项分布中的贝塔分量（它控制离散支撑集 $[0, n^2 -$ 所依据的概率分布的形状）

$n]$ 由此产生，我们观察到误差的总体概率分布如下：

$$\begin{aligned} f(x \mid \alpha_1 = \alpha_2 = n) &= \text{constant} \cdot x^{n-1}(1-x)^{n-1} \\ &= \frac{x^{n-1}(1-x)^{n-1}}{\int_0^1 u^{n-1}(1-u)^{n-1}\ \mathrm{d}u} \\ &= \frac{\Gamma(n+n)}{\Gamma(n)\Gamma(n)}\, x^{n-1}(1-x)^{n-1} \\ &= \frac{1}{\mathrm{B}_I(n,n)} x^{n-1}(1-x)^{n-1}, \end{aligned} \tag{10}$$

其中 Γ 是伽马函数，而不完全贝塔函数 B_I 对该函数进行归一化，使其在有限总体 M 上满足 Haar 测度所要求的和为 1。

接受$\mathrm{B}(\hat{}\alpha_1, \hat{}\alpha_2 \mid x, y)$ 作为样本 Kemeny 距离概率分布的似然函数；该分布的支撑集由大小为 N 的二项分布有效离散化。由此可知，最大似然估计等价于方程 1a 中给出的解析解，当且仅当在 $m \to \mathcal{M} = n^n - n$ 的极限下，我们观察到凹概率分布（我们已指出该分布为指数型，因此既可数值求解，也可解析求解）必然在一个共同的点估计处达到最大值，该估计表示为 $\hat{}\alpha_1$、$\hat{}\alpha_2$ 的线性函数；由于总体定义域上的所有度量空间均具有等距等价性，弗罗贝尼乌斯范数形式的概率测度也就直接成立。如果我们只接受这样的概率结构：矩的分布为贝塔-二项分布，且它表示两个服从贝塔分布的随机变量向量之间的距离，那么

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{\rho_\kappa}{(n^2 - n)} \tag{11a}$$

$$\frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} = \frac{(n-1)^2(n+4)(2n-1)}{18n \cdot (n^2 - n)} \tag{11b}$$

$$\hat{\alpha}_1 = \frac{\rho_\kappa(18\rho_\kappa^2 + 2n^5 + n^4 - 19n^3 - 18n^2\rho_\kappa + 31n^2 + 18n\rho_\kappa - 19n + 4)}{n(n-1)^4(2n^2 + 7n - 4)} \tag{11c}$$

$$\hat{\alpha}_2 = \frac{(n^2 - n - \rho_\kappa)(18\rho_\kappa^2 + 2n^5 + n^4 - 19n^3 - 18n^2\rho_\kappa + 31n^2 + 18n\rho_\kappa - 19n + 4)}{n(n-1)^4(2n^2 + 7n - 4)}, \tag{11d}$$

据此估计 $(\hat{}\alpha_1, \hat{}\alpha_2)$ 余积的联合分布$Z = xy$，并且对任意 $n, \rho_\kappa\ x, y \in \overline{\mathbb{R}}^+$ 都有确定定义。然后，使用样本估计的形状参数（方程 11c 和 11d）求解距离，进而（通过方程 4）得到相关性估计。由此，我们观察到一个函数型且收敛的估计量，用于估计给定有限样本量 $N = n^2 - n$下、以观测数据为条件的两个总体向量之间 Kemeny 距离的概率分布。

根据样本依赖的分布参数$\hat{}\alpha_1$ 和$\hat{}\alpha_2$，我们可以看出，贝塔分布的方差也有一个确定性表示，如方程11b左侧所示。

当然，直接根据紧且全有界支撑集上的 Kemeny 距离空间的唯一性可知，任一极值的方差也同样是唯一确定的，因此是唯一退化的，因为仅存在 $|m| \in \mathcal{M} = 2 \mid \rho_\kappa \in \{0, n^2 - n\}$。因此，该估计量不确定性的经验估计在总体中达到最小值，并且对于有限样本也可以进行适当的学生化处理（详情见第2.4 节），同时提供了用于处理极限情形的上界。

2.1.1 边缘矩估计量

对于从共同总体中均匀抽取的两个长度为 n 的向量之间的 Kemeny 距离的二元联合分布，我们接下来同样使用矩估计法来刻画边缘分布。我们认为，对于任意希尔伯特空间（包括 Kemeny 距离函数空间），给定向量与其自身之间的距离始终唯一确定为0 。因此，继续使用贝塔分布来描述 Kemeny 距离函数的概率，我们估计 $\alpha_\bullet$，即边缘分布 X的形状参数。根据均匀分布的对称性，只有一个未知量，它可表示为方差的函数。对于从共同总体中独立抽取的样本，其方程为2 乘以 $n^2 - n$：

$$\frac{\alpha_1^2(n^2-n)}{4\alpha_1^2(2\alpha_1+1)} = \dot{\sigma}_\kappa^2$$

$$\alpha_1 = \frac{(n^2 - n - 4\sigma_\kappa^2)}{8\dot{\sigma}_\kappa^2}, \tag{12a}$$

对于贝塔分布，将其按 $2^t(n^2 - n)$ 缩放后，可获得贝塔-二项分布所需的特征，而该分布可作为 Kemeny 距离观测到（特别是，其在扩展实数域上具有紧且全有界的有限支撑）。利用序中心统计量的位置参数（对应于点 $X = \{x_{(1)}, \ldots, x_{(i)}\}$，有序对称元素 κ_{kl} 的和满足 $\sum_{k=1}^{n} \kappa_{kl}(x) = 0$），可得到若干易于识别的特征。

首先，顺序统计量向量具有唯一确定的比例形式，其本质上是二次的，因此在使 κ^2距离达到最小值的相应点处也可得到该向量，即在期望处。为便于说明，我们注意到，该得分方程（通过对所有元素的 2^t 统计功效求导而解得，$\kappa^2(\cdot)$）足以确定从非退化总体中均匀抽取的任意边际随机变量的一阶矩。进一步将 κ 函数扩展为由 Kemeny 距离确定的函数空间的基线性算子，则随机变量 X 的 t^{th} 阶矩（$t = 1, \ldots, \mathbb{N}^+,$）等于 $\sqrt[t]{2^t \quad (n^2-n)} \sum_{kl}^{n} \kappa_{kl}^t(x)$，是 $\alpha_1 \in [0, \infty^+)$.的函数。

因此， X 的边缘分布仅由$\hat{}\alpha_1$ 和观测向量长度（即样本量）唯一确定，与任何仿射线性变换无关，这使得方差可以表示出来。

2.1.2 数值示例

以 Fisher 鸢尾花数据为例，可得到一个简单的数值示例，其中$n = 150$。两两距离和通过解析方法求解的参数列于表2中，可以立即观察到，支撑尺度与 Beta 函数的期望之积就是 Kemeny 距离：

$$\mu(n^2 - n) = \frac{\alpha_1}{\alpha_1+\alpha_2}(n^2 - n) = \rho_\kappa \mid \rho_\kappa = 13145, \alpha_1 = 0.5797333, \alpha_2 = 0.4059677$$
$$\frac{0.5797333}{0.5797333+0.4059677} \cdot (150^2 - 150) = 0.5881432(22350) = 13145$$

表 2：Fisher 鸢尾花数据集，$n = 150$，这使得 α_1，α_2 可以表示为已知距离和样本量的函数。

	萼片长度	萼片宽度	花瓣长度	花瓣宽度
萼片长度	0	$\alpha_1 = 0.5285905, \alpha_2 = 0.4567304$	$\alpha_1 = 0.151408205, \alpha_2 = 0.8409595$	$\alpha_1 = 0.1881198, \alpha_2 = 0.8028011$
萼片宽度	11990	0	$\alpha_1 = 0.5797333, \alpha_2 = 0.4059677$	$\alpha_1 = 0.5646111, \alpha_2 = 0.4209448$
花瓣长度	3410	13145	0	$\alpha_1 = 0.1171291, \alpha_2 = 0.8767339$
花瓣宽度	4243	12804	2634	0

2.2 最大似然估计量

矩估计量没有明确利用由概率定义的似然函数，因此会失去某些优势。然而，历史上观察到，与最大似然估计相比，该方法在贝塔分布的独立随机变量上表现更优。虽然次序统计量的构造历来基于连续随机变量，且其观测方差为常数，但我们借助 Kemeny 度量提出了一个框架，可通过最大似然估计量获得边际随机变量分布 x 和 y 的形状参数。基于这些参数，我们还考察了两个独立贝塔随机变量的乘积所产生的 Kemeny 相关统计量的最大似然估计，并据此构造出最大似然估计量。值得特别指出的是，由于不再假设不存在并列值（该方法适用于离散和连续随机变量），方差不再是先验明确已知的量，因此我们利用得到的形状参数直接根据数据对其进行估计，从而也得到方差的最大似然估计量。这使我们能够证明，两个随机变量乘积的似然函数等价于 Kemeny 度量的贝塔分布，因此其本身也是一个最大似然估计量。我们提出了一个构造长度为 n的贝塔分布（归一化）扩展实随机变量 X_j 的框架，并证明这些变量的均值和方差可通过利用估计得到的形状参数，直接由已知的贝塔分布矩构造方法求得。

将随机变量 $X_{i,\ j}$ 映射到次序统计量空间 $(X_{(i,j)})$，对于 $i = 1$、.....，n 以及 $j = 1$、.....，p，很容易验证其

对于支撑集 $[-\frac{N}{2},\ldots,0,\ldots,\frac{N}{2}]$ 中的每个独立分布随机变量，其分布为贝塔分布；通过对 $\kappa(X_j)$上的向量化结果取极值，可得到得分。将每个斜对称矩阵 κ 矩阵按 n 行求和，可得到一个 $\neg 1\times n$向量，其中包含随机变量 j 在先前定义的支撑集上的有序取值集合 $X_{(j)}$。在不存在并列值的情况下（如连续随机变量所致），极差为 $[(1-n)\ n-1]=2(n-1)$；然而，由于方差是自由估计的（依赖于数据），可能出现并列值，因此对 X_j的每个次序统计量分布进行识别时，必须使用两个参数。此外，假设贝塔分布随机变量非负 $x\in[0,1)$，因此还需要从每个向量 $\vec{X}_{(j)}$ 中减去最小次序统计量（根据对称偶 κ 函数，这也使观测到的最大样本值能够确定贝塔分布）。然而，在变换之前 $E(\vec{X}_{(j)})=0$，因此可见其无偏，正如底层的 Kemeny 度量本身也是无偏的一样（引理 1.2）。进一步考察确认，$\vec{X}$ 方差的解析解与 $\sigma^2_\kappa(X)$等价，从而证实了该函数表示的对偶性：

由此得到一组 j 贝塔分布随机变量，它们相互独立且非退化。基于这些变量，可通过矩估计法或最大似然法估计一组 j 参数对 $(\alpha_{(1,j)},\ \alpha_{(2,j)})$，从而得到一阶矩和二阶矩的唯一估计（唯一性由贝塔分布变量的紧致且全有界性质定义）。这也为变量 $Z_{(j)}=\frac{X_{(j)}}{\sqrt{\bar{\sigma}^2_j}}$的标准化样本特征提供了自然构造。我们指出，这为构造 Spearman 的 ρ奠定了基础，并将其推广为在存在并列值时也能识别（即可测）的形式。与现有应用于非参数数据分析的 Wald 统计量标准构造不同，它还是一种有符号测度。

这两个贝塔随机变量的内积$\binom{p}{2}=\{k,\ k'\}=K$由方程14a 得出（即方差—协方差矩阵和相关矩阵的下三角元素）。肯梅尼张成基及其等距欧几里得映射自然对应于相关系数和偏沃尔德检验统计量；基于样本进行估计时，其在总体中的方差为 1。但请注意，$\bar{\sigma}^2_j$ 既是所有 n几乎必然有限方差的经验表示，也由方程11a 和11b得出。因此，我们得到了恰好识别的最大似然估计量，其内积即为 Spearman 的 ρ 相关，因此也是无偏的最大似然估计量，其中心为期望值 0，概率分布定义为两个贝塔分布的乘积$\langle x,\ y\rangle$。

定理 2.1。Kemeny ρ 相关系数（在极限情况下，当存在连续独立随机变量时，它等价于 Spearman ρ）是 Kemeny 度量拓扑函数空间上的最大似然估计量。

证明。 首先对每个随机变量的边缘分布进行评估 x_k， $y = x_{k'}$，表示两个 κ 从定义在扩展实数集上的独立同分布有限分布中得到的映射，采用标准最大似然估计进行估计，其高斯误差在渐近意义下是正确的。令 $\vec{Z} = \langle xy \rangle$ 为边缘参数 $\alpha_{(1,x)}$， $\alpha_{(1,y)}$， $\alpha_{(2,x)}$， $\alpha_{(2,y)}$的函数。这些参数源于随机变量通过公式 κ 14a中的映射进行变换，从而在单位区间内产生严格非负值。根据以下估计方程，利用每个随机变量的两个参数构造 Z， $\mu^1_{\vec{Z}}$， $\mu^2_{\vec{Z}}$的一阶矩和二阶矩：

$$\mu^1_{\vec{x}} = \frac{\alpha_{1,\vec{x}}}{\alpha_{1,\vec{x}} + \alpha_{2,\vec{x}}} \tag{13a}$$

$$\mu^1_{\vec{y}} = \frac{\alpha_{1,\vec{y}}}{\alpha_{1,\vec{y}} + \alpha_{2,\vec{y}}} \tag{13b}$$

$$\mu^2_{\vec{x}} = \frac{\alpha_{1,\vec{x}}\alpha_{2,\vec{x}}}{(\alpha_{1,\vec{x}} + \alpha_{2,\vec{x}})^2 \cdot (\alpha_{1,\vec{x}} + \alpha_{2,\vec{x}} + 1)} \tag{13c}$$

$$\mu^2_{\vec{y}} = \frac{\alpha_{1,\vec{y}}\alpha_{2,\vec{y}}}{(\alpha_{1,\vec{y}} + \alpha_{2,\vec{y}})^2 \cdot (\alpha_{1,\vec{y}} + \alpha_{2,\vec{y}} + 1)} \tag{13d}$$

$$\mu^1_{\vec{Z}} = \left(\frac{\hat{\alpha}_{1,x}}{\hat{\alpha}_{1,x}+\hat{\alpha}_{2,x}}\right) \cdot \left(\frac{\hat{\alpha}_{1,y}}{\hat{\alpha}_{1,y}+\hat{\alpha}_{2,\vec{y}}}\right) = \mu^1_{\vec{x}} \cdot \mu^1_{\vec{y}} \tag{13e}$$

$$\mu^2_{\vec{Z}} = \left(\frac{\alpha_{1,\vec{x}}\alpha_{2,\vec{x}}}{(\alpha_{1,x}+\alpha_{2,\vec{x}})^2 \cdot (\alpha_{1,\vec{x}}+\alpha_{2,\vec{x}}+1)}\right) \cdot \left(\frac{\alpha_{1,\vec{y}}\alpha_{2,\vec{y}}}{(\alpha_{1,\vec{y}}+\alpha_{2,\vec{y}})^2 \cdot (\alpha_{1,\vec{y}}+\alpha_{2,\vec{y}}+1)}\right) = \mu^2_{\vec{x}} \cdot \mu^2_{\vec{y}} + (\mu^1_{\vec{x}})^2 \cdot (\mu^1_{\vec{y}})^2 \tag{13f}$$

因此，构造性地，取两个独立同分布抽样的扩展实向量 X， Y。由此可得以下等价关系

$$\vec{x} = \sum_{k=1}^{n} \kappa_{kl}(x) - \min \sum_{k=1}^{n} \kappa_{kl}(x) \tag{14a}$$

$$\sigma^2_{(x)} = \frac{1}{2(n-1)} \sum_{k=1}^{n} \sum_{l=1}^{n} \kappa^2_{kl}(x)$$

$$\vec{y} = \sum_{k=1}^{n} \kappa_{kl}(y) - \min \sum_{k=1}^{n} \kappa_{kl}(y) \tag{14b}$$

$$\sigma^2_{(y)} = \frac{1}{2(n-1)} \sum_{k=1}^{n} \sum_{l=1}^{n} \kappa^2_{kl}(y)$$

$$\rho_{(x,y)} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{\vec{x}_i \cdot \vec{y}_i}{\sqrt{\sigma^2_{(x)}}\sqrt{\sigma^2_{(y)}}} = \frac{n}{n-1} \frac{\mu^1_{\vec{Z}} - \mu^1_{\vec{x}} \cdot \mu^1_{\vec{y}}}{\sqrt{\mu^2_{\vec{y}} \cdot \mu^2_{\vec{y}}}} \tag{14c}$$

注意，经验最大似然估计得出的边际变量参数用于计算内积的均值和方差，并进行必要的贝塞尔校正。在极限$\frac{n}{n-1} \to 1$ 下，因此方程14c与方程13f 在总体上等价。由此证明， $Z = xy$ 的概率分布，在将原始次序统计量视为归一化贝塔分布时，无论是否存在并列值，均会产生 Kemeny 的 ρ。该估计量是该余积的最大似然估计量和最小方差估计量，因为它是最大似然估计参数的线性函数，至此证明完成。由于 κ 函数本身是无偏估计量，其线性运算直接保证了无偏性。

■

备注 3。基于 Kemeny ρ和 Kemeny τ_κ 函数测得的随机变量方差相同，因此可在所有有限 n上进行识别。此外，根据最大似然估计得出的 Kemeny τ_κ相关系数，也可以通过最大似然估计 Kemeny ρ 参数的单调正交归一变换（使用方程 18得到）视为最大似然估计量。

注 3。根据函数等价性，最大似然估计量的任何函数本身都是最大似然估计量。互积和方差本身就是最大似然估计量（如方程13所示），而 Kemeny ρ作为这些参数的仿射线性变换，本身也是无偏最大似然估计量。但请注意，该最大似然估计量是基于 n 双变量 $i.i.d.$ 样本的定义域构造的，而非$\mathcal{M}^*$。根据该公式的双射性质，经方程18 变换的最大似然估计量（对其求导 τ_κ ）按定义在同一个稳定点处同时达到最大值，符合投影对偶性，因此确保最大似然解在两个度量定义域上均取得共同解，无论是在极限情况下还是有限样本下皆如此。由于存在两个正交归一线性方程判据，至少其中一个在所有有限样本下几乎必然是高斯–马尔可夫估计量，因而还能观察到更紧的界。■

2.3 点二列相关

标准非参数估计量也用于来自共同总体的组间双变量相关，最常见的名称是 Wilcoxon-Mann-Whitney 检验。然而，由于缺乏既定的 σ-代数，在结果变量存在并列值时进行估计一直较为复杂；如今，这种代数已得到定义。这与非参数检验的零假设显著性检验方法中，Wald 框架和一般概率框架各自存在不同估计程序相对应，相关研究中也采用了这些方法。根据本文所提出的估计量共同定义域的严格子空间刻画，可以立即得出：不存在并列值时，两者的性能相同。不过，存在并列值时，我们提出的估计量所具有的无偏性和最小方差性质，同样适用于位置差异偏移问题。例如，点二列距离偏离原假设（即 Kemeny 距离等于中位数，也就是组别与结果相互独立）的概率为 $\Pr^{z_\phi}(-\frac{-49}{30.63658}) = 0.1097329$，这与 Pearson 相关、Spearman 相关、Kendall 的 τ_b相关，以及采用连续性校正的 Wilcoxon秩和检验结果一致（W =25.5，p = .06933）。采用常规的 α-决策阈值时，原假设仍无法被拒绝；然而，正如下一节的结果所示，并列值的存在会导致非参数空间中的估计量统计功效较低（分布更分散）。

2.3.1 数值演示

通过自助法重复抽样，我们证明，对于按标准非参数双样本位置差异检验构造的估计量而言，我们提出的估计量在所有竞争方法中统计功效最高，并呈现出预期的经验最小方差特征。数据集取自 R 中经典的“sleep”数据集，通过重复抽样生成 25,000 个、每个包含 150 个观测值的数据集；检验统计量结果见表 20×3 3，各种基于相关系数得出的非参数效应量见表 3 和表 4。从这些结果可以看出，（已知有偏的）非参数估计量的一阶近似（均值）和二阶近似（标准差）都大于准确值，这与任何 James-Stein 风格估计量（即有偏或非最小方差估计量）的预期性能完全一致。正如假设的那样，竞争估计量的极差和中位数绝对偏差也都更大。列出这些结果是为了确认，部分 Wald 检验的最小方差性质按定义同样适用于这些估计量，也为了确认标准相关效应量与 Kemeny 度量之间存在直接关系；根据几乎所有估计（统计学习）问题的射影几何对偶性，这种关系是必然的。

对数值结果的考察直接证实，基于 Kemeny 的估计量类别使所有离散程度度量（如极差、平均绝对偏差和标准差）达到最小值。此外，我们发现，对于 Kemeny 参数，估计量类别的检验统计量的非正态性得到最有效的最小化，从而展现出最正态的无偏性和最小方差性能特征，并满足广义中心极限定理所预期的高斯–马尔可夫正态性，而无需关注得分本身的目标线性分解（即 Nelder 和 Wedderburn（1972）的单调函数空间）。具体而言，我们指出，这使得对评分尺度采用更为便捷的序数线性框架进行分析成为可能

表 3：比较双样本位置估计量之差的检验统计量，以评估其在不同样本量下对 Sleep 数据集的统计功效和性能

样本量		mean	sd	中位数	mad	min	max	极差	skew	峰度
20	Kemeny	1.51678	0.71893	1.56675	0.72590	-1.99108	3.26407	5.25516	-0.43512	0.12202
	Wilcox	24.26342	11.00131	23.5	11.1195	0	80	80	0.47789	0.16407
	Kendall	1.81228	0.85218	1.87730	0.84998	-2.34019	3.85763	6.19782	-0.47474	0.16679
	Pearson	1.94842	1.09417	1.89843	1.04310	-2.58812	8.59658	11.18469	0.36155	0.74704
150	Kemeny	4.446	0.726	4.47	0.72	0.90	6.92	6.01802	-0.18223	-0.01279
	Wilcox	1425.745	223.317	1417	222.39	664.50	2516.50	1852	0.18139	-0.01912
	Kendall	5.174	0.843	5.21	0.84	1.05	8.04	6.98899	-0.18372	-0.01263
	Pearson	5.363	1.03554	5.35	1.02	1.24	9.92	8.67795	0.09475	0.02065
750	Kemeny	10.03460	0.73047	10.04483	0.73596	6.42783	12.68464	6.25682	-0.08382	0.02329
	Wilcox	35833.29776	2503.45323	35803.00000	2521.90260	26834.00000	48158.50000	21324.50000	0.08592	0.02627
	Kendall	11.62749	0.84565	11.63919	0.85096	7.45119	14.68909	7.23790	-0.08438	0.02387
	Pearson	12.00136	1.02885	11.99176	1.03368	7.33255	16.16491	8.83236	0.06177	0.02354

这些结果中有若干值得注意的特征，值得重点指出。

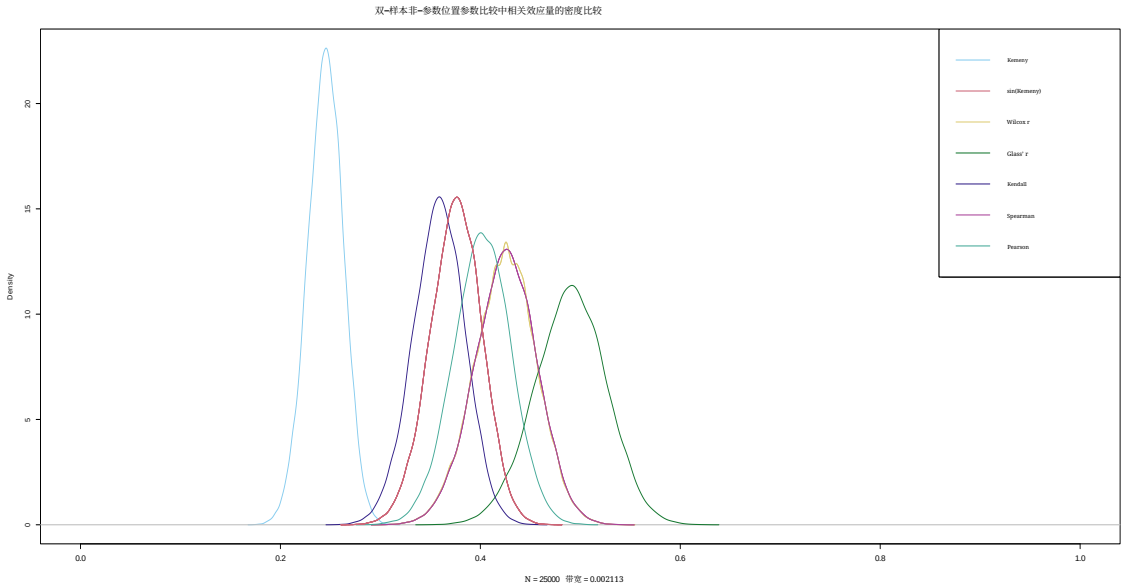
表 4：比较双样本位置估计量之差的效应量，以评估不同样本量下 Sleep 数据集的统计功效和性能。

样本量		mean	sd	中位数	mad	min	max	极差	skew	峰度
20	Kemeny	0.25789	0	0.25789	0	0.25789	0.25789	0	—	—
	sin(Kem τ_κ)	0.39411	0	0.39411	0	0.39411	0.39411	0	—	—
	Wilcox r	-0.41689	0.19097	-0.44700	0.17717	-0.78900	0.26200	1.05100	0.72752	0.22675
	Glass' r	-0.50094	0.23034	-0.53100	0.20534	-0.94800	0.33300	1.28100	0.72624	0.26976
	Kendall	0.36761	0.16849	0.39367	0.15338	-0.23295	0.69614	0.92908	-0.72588	0.23961
	Spearman	0.42775	0.19591	0.45893	0.18081	-0.26784	0.81054	1.07838	-0.72792	0.22655
	Pearson	0.40385	0.18343	0.42146	0.17135	-0.22409	0.74867	0.97276	-0.73520	0.61920
150	Kemeny	0.24500	0.03954	0.24635	0.03940	0.08116	0.38917	0.30801	-0.19257	0.03618
	sin(Kem τ_κ)	0.37469	0.05760	0.37739	0.05738	0.12715	0.57394	0.44680	-0.26570	0.08003
	Wilcox r	-0.42277	0.06812	-0.42500	0.06820	-0.67000	-0.14000	0.53000	0.19522	0.03770
	Glass' r	-0.48998	0.07898	-0.49300	0.07858	-0.77300	-0.16100	0.61200	0.19604	0.03656
	Kendall	0.35828	0.05761	0.36033	0.05751	0.11875	0.56640	0.44765	-0.19380	0.04031
	Spearman	0.42419	0.06834	0.42672	0.06824	0.14030	0.67137	0.53106	-0.19526	0.03775
	Pearson	0.40066	0.06396	0.40293	0.06354	0.10657	0.64143	0.53486	-0.17317	0.03530
750	Kemeny	0.24501	0.01779	0.24532	0.01776	0.17397	0.31319	0.13922	-0.08538	0.00562
	sin(Kem τ_κ)	0.37529	0.02591	0.37588	0.02584	0.26988	0.47235	0.20247	-0.11913	0.01306
	Wilcox r	-0.42487	0.03085	-0.42700	0.03262	-0.54400	-0.30200	0.24200	0.08827	0.01076
	Glass' r	-0.49003	0.03558	-0.49100	0.03558	-0.62600	-0.34800	0.27800	0.08577	0.00614
	Kendall	0.35834	0.02593	0.35877	0.02591	0.25479	0.45754	0.20275	-0.08526	0.00694
	Spearman	0.42515	0.03085	0.42569	0.03084	0.30184	0.54302	0.24119	-0.08581	0.00609
	Pearson	0.40153	0.02876	0.40192	0.02873	0.28588	0.50723	0.22135	-0.08318	0.03151

具体而言，表 4提供了可比较的度量（即效应量），揭示了现有 Kendall τ_b估计量的若干重要缺陷：其平均有偏相关系数更接近 Pearson 相关，而非 Kemeny 相关。然而，Kemeny 相关 (τ_κ) 及其正弦变换（与 Spearman 的 ρ 类似）的方差（以及极差和平均绝对偏差；表 4）均小于所有其他估计量。值得注意的是，这些统计量所对应的估计量均旨在实现非参数估计，其结果如图 1所示。

尽管最大的效应量估计量确实可行，甚至是理想的候选方案，但通过实证和理论推导可以明确得出，即使采用完全非参数的自助重抽样程序，针对任意分布样本，使用本文提出的参数框架也能获得最紧的界。我们还要指出，Kemeny τ_κ 相关的正弦变换具有一个中心估计量，从定性上看，它是现有替代估计量的平均值，其中也包括使用 Pearson 相关时出现的模型设定错误。这充分展示了对偶原理，同时进一步表明，并列值的存在会显著影响非参数、基于秩的估计量的估计特性。Kemeny 拓扑的规律性在函数层面上更加明确，因此能给出极受期待的最紧界；否则，这些界只能通过实证估计，并且只能渐近地达到有限样本的 Cramér–Rao 下界。而我们的估计量已被证明具备这一性质，并且通过对预期偏度和超额峰度采用最“正态”的近似即可达到该下界

图 1：在 25,000 个由 $n = 750$ 个元素组成的重采样集中，各种点二列相关系数的分布可视化。



值（对于任意线性函数空间，它们渐近地都为 0；引理 A.12）。

2.4 在凯梅尼线性模型中引入更强的样本依赖性（Student 化）

现在自然会产生一个问题：当统计估计量服从参数分布，且对任意一对独立随机变量都满足高斯-马尔可夫定理时，若 M 不一定是无放回抽样得到的，那么在有限样本中应如何构造检验统计量，以检验其是否与 0（原假设）存在显著差异？此前推导出的表示形式使我们能够避免在存在并列值与不存在并列值的样本之间对方差近似作两分处理，并允许利用距离的分布（对于相关性，则是相应距离的仿射线性变换）除以这些分布的标准差，为所有有限样本唯一确定精确的 p值。

然而，这种表示并不全面，因为它排除了某些情况的存在，例如有限响应集，其中秩序（因此标准差）小于极值，且其分布也并非总是均匀的。这使我们能够在包含更多自由度的定义域上构建估计量：从本质上讲，这使我们能够将 t 分布的特征描述扩展到非参数估计情境中。从实际应用的角度来看，这为 Kemeny 估计量提供了一种估计方法，而该估计量的……

Student 的 t 检验和 Welch t 检验可能会出现。我们通过缩放干扰参数对所得 t-统计量进行经验近似，以保持 Kemeny-Wald 统计量的紧致且全有界性质。不过，鉴于 Kemeny 的方差 ρ 和方程 2 已被指出是等价的（作为同一线性 κ 函数映射的方差），这既允许在总体方差已知（且实现距离任意）的情况下进行极限单样本检验程序，也允许对估计的共同样本总体进行有限样本校正。

2.4.1 单样本 Student t 检验的非参数替代方法

人们自然会想知道，与方程中定义的 Kemeny 距离充分统计量相比，构造 t 分布所需的自由度从何而来！8。在处理前文所述的相关性估计量时（见方程 4），交叉乘积已经经过归一化，只有 1 个自由度，对应于两个向量相对于唯一确定的原点 0 之间的距离： $\frac{n^2-n}{2} \pm n^2 - n$，因此 $df = n$。但特别需要指出的是，这些事件出现的情形，即在 $\mathcal{M}$ 上进行非均匀抽样，这自然会增大所有 Kemeny 距离分布的期望方差 ρ_κ。因此，我们希望采用一种依赖于样本的方法来增大方差，以反映某些（奇数）距离被抽取的概率高于通常情况，也就是这些距离更容易被抽中。这就构成了应用斯卢茨基定理的典型情形。现在我们证明，这种扩张是单侧的（即， ρ_κ 函数空间的方差下界为 $\dot{\sigma}^2_\kappa(n)$）：

引理 2.2。 所有 Kemeny 距离总体的依赖样本的方差 $\mathcal{M}$ 几乎必然大于或等于 $\dot{\sigma}^2_\kappa(n)$ (公式 3)。

证明。 对于 ρ_κ ，有符号距离函数空间的支撑集由紧致且全有界的区间 $U(\mathcal{M}) \in \mathbb{N}$给出，其期望为 0，且方差几乎必然有限（使用方程 3定义）。由于所有距离的方差均可根据总体 $\mathcal{M}$进行测量，因此，无偏估计量的方差若要降低，就必须对应于对 Kemeny 距离既定支撑集的截断（缩减），或缩小集合 $\mathcal{M}$（该集合涵盖所有无放回排列），否则就会变成有偏估计量。根据 Glivenko-Cantelli定理（定理 1.8），有符号 Kemeny 距离具有一致性，并且鉴于其无偏性（引理 1.2），唯一可行的替代方案是在保持估计量无偏性的同时，以一致的方式缩小测度空间的支撑集。

由于基于长度为 n 的 Kemeny 度量测得的独立随机变量，必须呈现距离不小于 U(M) 的静态邻域，因此可以推知，为了限制可能的可观测空间，缩小方差也必须缩小支撑集。这样的操作自然会产生不可测向量，同时按照公理定义仍保持均匀的独立同分布抽样；而根据定义，这会导致有偏。

样本上。然而，由于 Kemeny 度量是无偏的，这就造成了悖论。因此，我们拒绝“方差必须缩小支撑集”这一假设，并由此通过反证法得出，Kemeny 方差以下式为下界 3 ，该式是观测到的二元样本的函数；否则，在扩展实数轴上可能出现不可测随机变量，从而导致悖论。根据给定样本进行归纳法 n，其中所有 $n < n + k,\ k = 1,\ \ldots\ldots,\ \infty^{+}$，可知，$k$ 中的子样本数据元素就是直接根据这些距离观测到的元素，因此不会引入额外信息，除非相应地增大样本量；而对于任意样本量，样本都必须具有强正则概率结构（引理 1.10），否则无论较大样本还是较小样本都将不可测。■

然而，在处理具有固定距离的分布位置参数时，自然会出现更复杂的估计情形，从而引入以向量特定识别形式呈现的更具体信息。也就是说，两个独立向量 $x,\ y \in \overline{\mathbb{R}}^{n\times 2}$ 分别由 $\kappa(x)$ $\kappa(y)$构造而成，如方程2所示，例如处理非连续随机变量时便属于这种情况。这种情形也涉及有限样本的基数 $m \in \mathcal{M}$问题，因为此时我们可以根据方差，将具有相应频率的特定观测分布识别出来（即重复抽样或非互斥抽样），而不再是从指数族中进行均匀且互斥的选择。

因此，估计每个样本的方差（以及与之成比例的标准差），便可明确刻画生成给定检验的分布；在此基础上，样本独立的测度空间中概率密度函数的集中程度，以及由此而来的似然函数，都必须遵循该分布。随着自由度增加，Student t 分布会出现在 Wald 统计量中；此时方差通常彼此不同（这是不放回抽样的结果），这一点也自然随之体现。根据上述构造，我们得到 Student 单样本 t 分布和 Welch 双样本 t 分布检验统计量的标准情形划分；例如，观察有限响应集而非连续随机变量时，就可能出现重复抽样。

根据方程 2 可观察到，对于典型的二项分布情形（即在两个事件之间观察到带有并列值的完美排序），由于两个希尔伯特空间之间存在等距线性等式，欧几里得表述与 Kemeny 表述下的方差是相同的。将带有并列值的任意置换作为排序的默认原假设 μ_0，相应的样本置换与经验样本之间的距离为 $U(\mathcal{M})$，并以标准差进行标准化。因此，在以下关系下，有符号距离与集中程度之比趋于 0，所得的 Wald 检验收敛于无限统计功效，并能够区分总体中的所有其他距离，而无需假设总体中的数值数量固定 $U(\mathcal{M})$，而是允许在上述邻域内进行重复抽样；这是强大数定律线性收敛的一个简单期望。这里

体现在术语 s_p 中，即两个变量之间的合并标准方差：

$$t_{n-1} = -\frac{\rho_\kappa}{s_p} \tag{15a}$$

$$s_p = \sqrt{\frac{\dot{\sigma}_\kappa^2}{2 \cdot \hat{\sigma}_\kappa^2(X)}} \tag{15b}$$

其中，分母中的各项在区间 (0， $0.5]$ 上紧且全有界，因此其期望之和（即一个随机变量与其自身的组合）在总体二元分布中发生随机并列值的概率不高于随机机会时，总和小于或等于 1。这会得到一个依赖样本的方差，其构造与总体方差相同，且与 Spearman 的 ρ 部分 Wald 检验构造相同；若存在更大的变异性，则会引入必要的 Slutsky 调整。这样便允许总体 Kemeny 距离出现更大的距离，同时仍保持紧致且全有界。经过这一调整，我们对任意可线性排序的二元分布所期望的渐近正态性、一致性和最小方差特征也得以维持。

如前所述，所有样本量下的所有距离均被标准化到单位区间上，因此，单位区间上的距离与有限且全有界的估计方差之比（该方差也紧致且全有界）始终介于 $[0, 1]$, 之间，因此几乎必然一致（忽略退化的经验分布）。若将相对于任意原假设置换的有符号距离加以刻画（标准做法是将距离定义为相对于中位距离期望的 $0n$，这等价于均值差为 0），则可得到几乎必然有限的经验估计，由此允许对可排序的离散和连续独立随机变量进行任意抽样。注意，作为$\hat{s}_p^2$ 的函数， $U(\mathcal{M})$ 的总体分布也必须相应调整：为反映并列值上的更规律抽样（因为非并列值在任何样本量足够大的分布中都只是可忽略的部分），我们必须预期距离的方差会增加，以满足出现超过期望数量的正距离平方的情况。因此，可对总体作简单调整， $U^*(\mathcal{M}) = {}^{U_{(\mathcal{M})}}{}_{s_p^2}$，并按集中率进行膨胀。因此，该新分布下的方差可通过对方程 3 进行标准化，并以 s_p^2调整得到，

$$s_\kappa^2 = U^*(\mathcal{M}) = \frac{\frac{(n-1)^2(n+4)(2n-1)}{18n}}{s_p^2},$$

该分布在总体渐近意义下显然符合引理 2.2所确立的要求（不失一般性），且可立即看出其服从 χ^2 分布。

2.4.2 Welch t 检验

考虑两个向量样本 (*x*， *y*)，它们分别独立地从 M* 中相应的置换映射表示抽取，允许观测值重复（有放回抽样）。注意，在 Kemeny 度量下的方差-协方差矩阵 Ξ 具有三个这样的估计量，由此可得到一个用于 $\overline{\mathbb{R}}^{n\times 2}$ 的恰好识别的方程组，$n < \infty^{+}$. 具体而言，这三个成比例的线性估计量为：(1) $\kappa(x)\kappa^{\intercal}(y) = \rho_{\kappa}$ (*x*， *y*)，(2) $\sigma_{\kappa}(x) = \sqrt{\kappa(x)\kappa^{\intercal}(x)}$，(3) $\sigma_{\kappa}(y) = \sqrt{\kappa(y)\kappa^{\intercal}(y)}$，分别表示 x 和 y 的相关系数与标准差。

通过实证模拟，我们可以观察到，所得到的分布符合自由度为 $n-2$ 的 Student t 分布，因此可以针对非参数相关性构造带符号的检验统计量，使其符合标准 t 分布（斯卢茨基定理）；这一方法利用了非恒定、因而依赖于样本的变量方差度量。实证模拟已证实，二元相关性的 Wald 统计量确实服从 t_{n-2} 分布，而图 2[2] 则提供了 Student t_{1248} 分布的实证比较，其缩放后的非中心参数等于观测分布的均值，从而确保理论分布与实证 Wald 检验统计量的期望在实证上保持一致；图示结果中的实证数值见表5，并表明当随机变量存在并列值时，Kendall 的 τ_b 估计量在有限样本下既非有效，也非无偏。需要指出的是，Kemeny Wald 检验具有最小方差和最小极差（即最紧），且其分布最接近正态分布；这些特性均与高斯-马尔可夫定理所证明的无偏最小方差估计量的性能一致。根据切比雪夫不等式，式ˆ所示的一般 Kemeny 类估计量 θ 及其所有仿射线性和单调变换均具有一致性；由于该概率分布稳定且无偏（定理1.4，引理1.2），这一性质对整个有效排列空间 $m \to \mathcal{M}^{*}$均成立：

$$\lim_{n\to\infty^{+}} \lim_{m\to\mathcal{M}^{*}} \Pr(E_m\Big(\frac{\hat{\rho}_{\kappa}(x,y)}{s_{\kappa}} - \frac{\rho_{\kappa}}{s_{\kappa}}\Big)^2 = 0) = 1, \tag{16a}$$

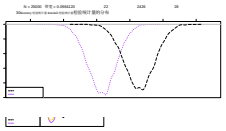
$$\Pr((\hat{\theta}-\theta)^2) \geq \epsilon) = \Pr((\hat{\theta}-\theta)^2 \geq \epsilon^2) \leq \frac{E(\hat{\theta}-\theta)^2}{\epsilon^2}, \tag{16b}$$

$$\lim_{m\to\infty^{+}\equiv\mathcal{M}^{*}} E_m(\hat{\theta}-\theta)^2 = 0,\ \forall\, n < \infty^{+}. \tag{16c}$$

$$t_{n-2} = -\frac{\theta(x,y)}{s_{\kappa}} \tag{16d}$$

$$s_p = \sqrt{\frac{\dot{\sigma}_{\kappa}^2}{\hat{\sigma}_{\kappa}^2(x)+\hat{\sigma}_{\kappa}^2(y)}}, \tag{16e}$$

[2]图例颠倒了 Student 和 z 分布的图标。



$$s_{\kappa} = \sqrt{\frac{\frac{(n-1)^2(n+4)(2n-1)}{18n}}{s_p^2}} = \sqrt{\frac{((n-1)^2(n+4)(2n-1)) \cdot s_p^{-2}}{18n}} \tag{16f}$$

请注意，分母中的两个方差相结合，依据全期望定律，保留了对两个随机变量方差的平均。通过在式$\sqrt{.5}$ 的每个 $\kappa^2(\cdot)$ 矩阵单元中使用缩放常数2，即可实现这种平均。式16f 中的集中度调整项 $\sigma_{\kappa}^2(x)$ $\sigma_{\kappa}^2(y)$ 经线性组合后，得到秩的欠离散率（小于或等于 1）；再用该值去除已知的总体方差，从而得到两个随机向量 x 和 y 在给定 n时排序排列之间差异的估计量。由此对总体 $U^*(\mathcal{M})$ 进行标准化，且所得结果几乎必然有限。

图 2：演示了 Kemeny 部分 Wald 检验的 t_{1248}-分布与理论分布一致，符合必要的高斯-马尔可夫定理，且其高阶矩更接近正态分布。

图 3： n $= 2500$ 时经验分布的可视化结果；我们观察到，即使对于较大的样本，Kemeny τ_{κ} 估计量也更加一致，并且更符合正态分布。

表 5：针对各种估计量，对序数随机变量的双变量相关和 Welch t_{1248}–检验的 25,000 个检验统计量分布进行的经验比较。

样本量		mean	sd	中位数	min	max	极差	skew	峰度
(a) n = 1250	Kemeny t–检验	15.1668	0.8943	15.1709	11.3614	18.4783	7.1169	-0.0392	-0.0008
	Kendall z–检验	16.3493	0.9521	16.3554	12.2476	19.8721	7.6245	-0.0436	0.0056
	学生 t_{1248} $\frac{1}{c^3}$	15.1746	0.9169	15.1709	10.9595	18.9556	7.9961	0.0294	-0.0014
(b) n = 2500	Kemeny t–检验	24.3124	0.8286	24.3218	20.6684	27.5784	6 .9101	-0.0324	-0.0012
	Kendall z–检验	26.5708	0.8893	26.5763	22.7169	30.0486	7.3318	-0.0298	-0.0074

2.4.3 配对样本非参数 t 检验

假设存在一个二元总体，其中独立随机变量 x 服从二项分布，且基态出现的概率为 $1-\pi$；另有随机变量 y，其元素 y_i、$i=1$、.....、n 可以按秩排列。每个随机变量都可以按受试者 i 编号，因此这两个随机变量表示在不同条件下对每个人进行的同一项测量：由此可进行配对 t 检验。非参数对应方法是 Wilcoxon 符号秩检验。再次考虑 Sleep 数据集，其中对接受两种不同药物治疗的同一组个体进行了 $n=10$ 次观测。使用以下线性期望，

$$-\frac{\rho_\kappa(X,Y)}{\frac{\dot{\sigma}_\kappa^2(n)}{\hat{\sigma}_\kappa(X-Y)}} \sim t_{n-1} \tag{17}$$

我们根据方程 3 从方程 2 获得方差调整的标准差，该调整基于第一次和第二次试验排序之间标量差异的差值。当秩的离散程度更大时（例如，并列值更多，因此对于真实排序的不确定性更大），分母会增大。因此，在存在和不存在并列值的情况下，Wilcoxon秩和检验的一致估计量（以及相应的检验统计量）会变小，表明相对于分子而言统计功效较低；分子包含邻域区间内的恒定带符号距离 U^*（$\mathcal{M}$）。

2.5 Kemeny 范数测度空间与 Frobenius 范数测度空间之间的对偶性

射影几何中的对偶性概念是一条自然的研究路径，Kendall 最先将其用于探讨相关性（1948年，第 129 页）。这一概念认为，对于欧几里得空间的平面射影几何，存在一种对偶置换几何；本文在 Kemeny 度量空间与欧几里得度量空间之间观察到了这种几何关系。从形式上说，由于 Kemeny 度量空间具有有限性，我们可以将其视为一个伽罗瓦域，进而视为一个对偶向量空间，满足对偶锥的三个必要性质；这两个空间在各自的度量范数下均为连续空间。

此外，对偶性存在的标准构造，即利用第二个域来区分给定域中的相同元素，在对偶度量空间的刻画中显然是不言而喻的。暂且忽略线性置换域在由线性得分构成的总体上的极限情形（即 Le Cam 所述的标准渐近参数学习问题，2012年）。在这种情形下，可以观察到总体上的完美参数评分拟合，这意味着完美排序等价，从而表明排名与评分之间存在通过累积分布函数（CDF）建立的双射关系。相反，我们将说明，尤其是在针对线性泛函映射进行 Tikhonov 正则化的病态或有偏学习问题中（尽管解不一定唯一），两个度量空间的对偶性能够给出恰好识别的唯一解。后续研究将这一刻画扩展到多种社会科学中常见的有限样本问题，证明 Karush-Kuhn-Tucker 约束优化（即岭回归）存在唯一解。

在 Kendall（1948年，第129页）中，声称 Pearson 的 r 与 Kendall 的 τ_b之间存在以下等价关系：

$$r_{x,y} = \sin\left(\tau_b(x,y) \cdot \frac{\pi}{2}\right). \tag{18}$$

我们将证明这一刻画是无效的，并转而证明，方程18 左侧实际上是 Kemeny 的 ρ，可由κ（x）和 κ（y）构造得到；它仅在总体上渐近收敛于 Pearson 的 r，从而否定了在样本上存在严格双射。我们关注的是确立这样一点：对于存在并列值的定义域，两个相关系数并不相等，因此在没有进一步假设的情况下，Pearson 的 r 不可能从该无效估计量在任何定义域上得到，除非该定义域是 S_n；并且即便如此，也只有在二元分布进一步服从高斯分布且具有良好的线性适定性时才成立。当人们将二元相关扩展到多个预测变量相对于共同目标变量在非欧函数空间上同时求解的情形（即标准多元回归）时，这一认识尤为重要。

首先，考虑 ℓ_2 或弗罗贝尼乌斯范数的定义——插入一个或多个无穷值，明确会导致至少一个随机变量的期望为非有限（因而退化）。因此，由于期望非有限， $n \times 1$ 向量在欧几里得度量空间中既无法中心化，也无法评估相对距离。内积同样未定义，因为内积（即交叉乘积）中非有限值的总和本身也是非有限的，因此皮尔逊相关度量会退化。然而，扩展定义域是必要的，因为 Kemeny 相关可以有效地应用于扩展实数上，并且对于向量的任意有限 n 排序都会产生一个有限的相应度量。这导致了一个悖论：Kemeny 相关存在并取得有限的收敛值，而皮尔逊相关却退化为不可测，这与方程 18相矛盾。因此，对于得分的线性投影为非线性的所有情形，方程 18 中定义的关系均无效

变量之间的关系为非线性时——包括但不限于扩展实数轴上的情形。此外，对于有限样本，即使满足大数弱律，若随机变量服从非高斯分布，皮尔逊相关按定义也存在偏差，因此无法满足等价性。

然而，我们可以从 Kemeny 度量构造出一个皮尔逊相关系数的替代估计量，并且很容易发现它在总体上是等价的。对于每个扩展实随机变量 $\{x \equiv \kappa(x)^{n\times n}, y \equiv \kappa(y)^{n\times n}\}$，我们希望从现有张量 $n \times n \times 2$ 中构造一个阶数为 $n \times 2$ 的双变量向量矩阵，以符合皮尔逊积矩相关系数所必需的内积形式。在斜对称置换域上，对每个斜对称矩阵的所有 k 行求和即可得到这一结果，从而生成两个 $1 \times n$ 向量，它们的期望均为 0，分别记为 $\vec{x}^{\intercal}_{n\times 1}$ 和 $\vec{y}^{\intercal}_{n\times 1}$。具体如下，其中根据需要以 $\vec{x}$ 替换 $\vec{y}$：

$$\vec{x} : \overline{\mathbb{R}}^{n\times 1} \to \kappa(x) \in \mathbb{R}^{n\times n} \to (\mathbb{R}^{1\times n})^{\intercal}$$
$$\vec{x} = \sum_{k=1}^{n} \kappa_k(x) = \left[\sum_{k}^{n} \kappa_{l=1}(x), \sum_{k}^{n} \kappa_{l=2}(x), \cdots, \sum_{k}^{n} \kappa_{l=n}(x) \right]^{\intercal}. \tag{19}$$

我们注意到，期望 $E(\vec{x}) = 0$ 和 $E(\vec{y}) = 0$ 对于所有样本都是固定的，并且对于每个一元随机变量，都可以唯一地估计和确定总体方差以及 Student 化标准差调整集中度量（方程 2）。每个 $\vec{x}$ 随后都是某个变量的秩序向量，即使在 x 或 y 中观测到非有限变值时，其均值和方差仍为有限值，因为线性序仍然是有效的代表性映射：$\mathsf{h}\, \sigma_\kappa(\vec{x})\ \ \sigma_\kappa(\vec{y})\, \mathsf{i}$ 的内积，就是希尔伯特空间中依据平行四边形法则、具有固定单位距离的~~向量上线~~性序的余弦。

不过，我们首先必须证明，这种内积适用于扩展实数向量，这一点不同于皮尔逊相关系数。在 S_n 上，也就是并列值出现概率为 0 的置换空间中，我们发现，有限得分的秩与得分本身的内积是等价的。此前已经证明，Kendall 的 τ 是 Kemeny 的严格子集 τ_κ，二者对应的距离也是如此；因此，在共同子定义域上定义的任何度量，在另一个定义域上也必然且确实会给出等价的度量，这一点也证实了上述结论。然而，这两个空间的内积反映了排名对之间的欧几里得距离；在秩与得分之间存在双射等价的情况下，根据大数弱律，这又意味着 Pearson 的 r 与 Spearman 的 ρ 等价。解决了不同度量空间之间关系上的这一悖论后，我们回到以下问题：在 $\overline{\mathbb{R}}^{n\times p}$ 上，对均匀抽样变量的分布进行双变量秩序排列的效率如何。这使得线性函数能够用于任何可线性排序的二元分布，同时满足 Cr´amer-Rao 下界（引理 A.12），从而评估任何 $n \times p$ 分布的充分统计量。值得注意的是，这最终确定了

每个 p 变量的中位数和方差，以及相关矩阵 $\Xi_{p\times p}$，作为一个 ℓ_2-范数空间，从而确定期望的二次解。

这种对偶性使得即使是经 Tikhinov 正则化的线性方程组，也能够通过收敛评估被唯一识别并求解；根据引理 A.13 和引理 A.14，任何可行解都必定存在于真实解的邻域内，并以概率 1 单调收敛至总体上的期望。因此，即使欧几里得度量给出病态（因而非唯一）的解，与 Kemeny 度量及其射影几何对偶性的联合结合，也必然能够为所有有限样本找到唯一解；该解既最接近观测得分，也最接近真实（无偏）的排序。接受 $\Pr(x \mid \alpha_1, \alpha_2) \propto x$ 和 $\Pr(y \mid \alpha_1, \alpha_2) \propto y$ 后，根据全期望定律，必然有 $E(\Pr(\vec{x}) \cdot \Pr(\vec{y})) = \Pr(\vec{x}\vec{y})$。随后可将似然函数构造为边际形状系数与内积或相关系数的函数，而该函数是恰好识别的。 $L(\rho_\kappa \mid \vec{x}, \vec{y}, \propto \alpha_{x,1}, \alpha_{x,2}, \alpha_{y,1}, \alpha_{y,2}$ 的导数

3 讨论

最后，我们讨论非参数线性模型空间的相对价值。在传统统计学文献中，几乎普遍认为变量取值的效用极其重要。这一点很容易证明：在参数估计问题中，我们观察到秩是得分的双射函数，因此它是识别过程不可或缺的组成部分。然而，在所有有偏或季洪诺夫正则化线性模型空间中，明确定义的逆映射会极为有用。事实上，以 Kemeny 相关矩阵为基来构建 EM 算法，以处理缺失数据、聚类、因子分析和一般多元回归问题，已经取得了极其积极的结果。根本问题始终在于存在并列值时无法识别概率模型；不出所料，线性组合的满射映射几乎必然会产生此类并列值。一个简单的例子是对二分组进行非参数双因素方差分析：其中并列值是由构造方式决定的，因此无法进行估计。这一难题恰好可以解释我们所认为的过去 80 年间相关研究进展不足的原因；否则，我们就必须施加实质性限制，或处理二项式爆炸所带来的混合问题，才能针对每个样本进行唯一估计（即标准次序统计量构造情形）。我们认为，以概率方式为任意可排序分布构造完备的希尔伯特空间，明确解决了当前统计学文献中一个基础且亟待解决的问题。例如，结合基于秩和得分的方法，可以使许多原本无法求解的病态方程组获得唯一解。此外，由这两个空间之间已指出的对偶性，这种结构

与参数情形下原本会得到的唯一解完全相同（这一点已得到实证确认）。

将一般机器学习问题发展为唯一解的可能性（否则这些问题将满足Hadamard不适定问题的定义），并将其作为对偶定义的损失函数，具有极大的研究与发展价值，同时也很符合直觉。在参数情形下，得分本质上表示了累积分布函数上的一种排序。能够正确地排列所有情形，并估计这些数据项的位置参数和尺度参数的估计方程系统，其解就是最小方差无偏估计量。然而，如果处理的是一组病态估计方程，则在欧几里得距离函数空间上，使得得分相对于目标的排序近似程度最接近的解集合，满足拉格朗日乘数（即鞍点）解；在所有候选解中，该解对应于最稳定的解，而其他情况下，这一解由完全任意选择的蒂霍诺夫偏差无限确定。因此，射影几何对偶性能够在所有可用数据上产生唯一且可泛化的解，而无需作出任何额外假设或限制，也不会引入超出欧几里得距离损失函数本已施加的计算负担。

这说明了如何解决似然原理所引出的、原本仍然存在的悖论：根据该原理，所有信息都包含在对正确概率分布的了解之中。依据两个函数空间的对偶性，纳入两个已正确识别的概率分布集合是冗余的，因此，过度参数化并不会带来关于总体分布结构的新信息。然而，当参数分布只是近似得到、而非确切已知时，纳入第二个正交函数空间，便能使欧几里得距离空间上的偏差得到恰好识别。因此，对互补测度空间进行对偶优化，必须仅凭一种度量就能增加相对信息。考虑到对于许多不同的学习问题而言，不包含尺度参数和位置参数的排序在功能上毫无意义（例如，n 个单位区间有限元素的正确顺序分布包含无穷多个符合“正确”排序的变量实现值），因此，自然的出发点应当是：我们希望获得唯一正确的排序，同时也希望它对欧几里得距离作出最准确的经验近似。当然，鉴于这种对偶性，在参数情形下，度量对偶性渐近而言是冗余的。然而，有限样本中随机扰动的不确定性，使得即使在存在正确概率分布的情况下，只要随机抽样均匀发生，仍有可能获得信息增益。

参考文献

Bogachev, V. I.（2007）。测度论。Springer Berlin Heidelberg。https://doi.org/10.1007/978-3-540-34514-5

Buldygin, V., & Kozachenko, Y. V.（2000）。随机变量和随机过程的度量刻画。美国数学学会。Diaconis,P., & Graham, R. L.（1977）。Spearman足规作为错乱度量。《皇家统计学会期刊：B辑（方法论）》，*39*，262–268。https://doi.org/10.1111/j.2517-6161.1977.tb01624.xHollander, M., Wolfe, D. A., & Chicken, E.（2014）。非参数统计方法（第3版）。约翰·威利父子公司。Kemeny, J. G.（1959）。广义随机变量。《太平洋数学杂志》，*9*（4），1179–1189。https://doi.org/10.2140/pjm.1959.9.1179Kendall, M. G.（1948年）。秩相关方法。Le Cam, L. M.（2012年）。统计决策理论中的渐近方法。施普林格科学与商业媒体。Nelder, J. A., & Wedderburn, R. W. M.（1972）。广义线性模型。《皇家统计学会期刊：A辑（综合）》，*135*（3），370。https://doi.org/10.2307/2344614Schoenberg, I. J.（1938）。度量空间与正定函数。《美国数学学会汇刊》，*44*（3），522–536。https://doi.org/10.1090/s0002-9947-1938

Semrl, P. (1996). 关于Cochran统计定理的矩阵版本 [《线性代数与统计学：庆祝 C. R. Rao 75 岁生日（1995 年 9 月 10 日）》]。《线性代数及其应用》，*237-238*，477–487。https://doi.org/https://doi.org/10.1016/0024-3795(95)00395-9

A 附录

定理 A.1. *The Kemeny norm space is a pre-Hilbert space with the property that for finite positive a, holds the following equivalence, using κ^* to denote the signed function as given in Kemeny (1959):*

$$\rho_\kappa(a \cdot \kappa^*(X_A), \kappa^*(X_B)) = a\rho_\kappa(\kappa^*(X_A), \kappa^*(X_B)).$$

Proof. 通过标量乘法，根据引理 0· $a = 0$，存在下界 A.2，并且对于所有有限的 A.4，存在一个实数上界（引理 n。因此，Kemeny 距离函数的典范形式按 $a^{-1}a = 1$缩放，所以线性距离总是等价于闭区间

$$[0, 1, \quad , n^2 - \cdots$$

n] $\forall n \geq 0 \quad a = 1$
()，若，否则则与该序列成正比，比例系数为正有限标量 $a \in \{(0, \infty^+)\} \setminus 1$。由此可见，Kemeny 巴拿赫范数空间也是正齐次的，但缺乏内积构造，因此是一个拟希尔伯特空间。 ■

推论 A.1.1. *From Theorem A.1 we have justified the valid existence of an inner-product formulation for the Kemeny pre-Hilbert metric space, by definition, which is provided in equation 1a.*

引理 A.2. *The Kemeny metric space is non-negative under the condition that $0 < a < \infty^+$. When the domain of a univariate random variable in $\mathbb{R}^{n\times 1}$ is extended to a bivariate pair (X_A, X_B) the function $\kappa(X)$ is convex, for the arbitrary arbitrary real constant α.*

Proof. 依次应用次可加性、对称性和不可区分对象同一性，可推出非负性，但这仅在对 $\kappa(\cdot)$ 的特定定义下成立，由此得出要求 $0 < a < \infty^+$:

$$\begin{aligned} \rho_{\kappa^*}(a,b) + \rho_{\kappa^*}(b,a) &\geq \rho_{\kappa^*}(a,a) \\ \rho_{\kappa^*}(a,b) + \rho_{\kappa^*}(a,b) &\geq \rho_{\kappa^*}(a,b) \\ 2 \cdot \rho_{\kappa^*}(a,b) &\equiv 2 \cdot \rho_{\kappa^*}(\kappa^*(X_A, \kappa^*(X_B)) \geq 0 \\ \rho_{\kappa^*}(\kappa^*(X_A, \kappa^*(X_B)) &\geq 0 \end{aligned} \tag{20}$$

■

定义 A.1（完备空间）。度量空间 (X, ρ) 若其扩张常数为 $\leq 2.$，则称为完备空间

引理 A.3。 *The Kemeny distance function has an expansion constant μ less than 2.*

Proof. 设存在两个实数区间族 $\{A_i\}_{i \in I}$ 和 $\{B_j\}_{j \in J,\ i,\ j=1,\ \ldots\ldots,\ n}$，使得 $a_i \leq b_j$，因此

$$\begin{aligned} \rho_{\kappa^*}(\kappa^*(A_i), \kappa^*(B_j))\, \forall i, j \implies & \sup_{i \in I}(\kappa^*(A_i)) \equiv \mu \\ & = \sup\{-1 \cdot a, 0 \cdot a\} \leq \inf_{j \in J}(\kappa^*(B_j)) \equiv \inf\{0 \cdot a, 1 \cdot a\} \\ & = \sup\{\mu\} \leq 1. \end{aligned} \tag{21}$$

因此可以得出，对于任意区间族 A_i，B_j 只要两个有限集的交集始终非空，$\cap$[上确界 A_i，下确界 $B_j] \neq \varnothing$就具有扩张常数 $\mu =$ 上确界$\{0,\ 1a\} \leq 2a\ \forall a > 0$。因此，定义 A.1 适用于 Kemeny 距离函数，因为其等距扩张常数始终小于 2；由于$X_A = A_i$，$X_B = B_j$，此证明适用于 $\mathcal{X}$ 中的所有得分向量；X 是由有限长度扩展实数向量构成的空间 n。■

引理 A.4. *There exists a unique finite upper-bound upon the Kemeny metric space for any collection of n element length vectors in the extended reals whose lower bound , is 0.*

Proof. 对于任意 $\kappa^*(X_j)\ \forall X_j, j = 1, \ldots, p$ 结果得到一个最大值为 (n^2-n) 2 的映射向量，并为其赋予三个不同值中的一个:

$$\begin{array}{ccccc} a_{ij} \in & 1 & -1 & 0 & X_i \quad i = \{\ ,\ ,\ \}。如果所 \\ & & & & = n^n - n \\ 1 & & & & \forall 0 < \\ n^n = & n_1 & n_2 & \overline{\quad n_n} & \end{array}$$

有的 $,\ ,\ ,$ 均构成单调递增序列，则存在 M，$n < \infty^+$，表示所有扩展实数得分向量，$\{\ ,\ ,\cdots\}$。每个向量在运算 κ^* 下得到一个维数为 (n^2-n) 2 的映射；约化后的映射集合由度量空间的对称性给出。因此，对于固定的 n，沿单调序列 X_j 存在 $I_n = \{1, 2, 3, \ldots\} \in X_n$，由此得到 $\sum_{i=1}^n \kappa^*(X_j) = \sum\{1a, 1a, \ldots, 1a\}$。根据方程 1a，进而可知，在这样的域上，可达到的最大距离由相应单调排序序列中差值的绝对值序列确定，并代入值 $1a \to -1a$；唯一的例外是 $a_{i=j} = 0$:

$$\sum_{i=1}^n \sum_{j=1}^n |a_{ij}^* - (-1 \cdot a_{ij}^*)| = |\sum_{i=1}^n \{2_i a, \ldots, 2_n a\}| = \frac{n^2 - n}{2} \cdot 2a = a(n^2 - n).$$

唯一性证明与前述引理相同，其中对所有映射求和 κ^* 含有一个或多个 a 根据方程 1a 其距离按定义大于观测值，因此对于任意固定的有限正 a 以及有限正整数 n，均具有唯一的极值。■

引理 A.5。 *The Kemeny metric space is a complete metric space.*

Proof. 根据 Kemeny 度量的引理 A.2 和引理 A.4 的有效性，对于所述距离函数，其最大可能距离以及任意有限序列 X长度为 n 的距离为 $a(n^2-n)$，这是由包含 $\frac{n^2-n}{2}$ 个元素的可加序列得出的，每个元素的值均为 $2a$，由此确立了对所有有限 r 均为正且有限的定义n。因此，根据构造性映射可得出柯西收敛性：对于每个

$X_A\ X_B \in \qquad I_n\ I'_N \qquad\qquad 0 \leq |\rho_\kappa\ \kappa^*\ X_A\ - \kappa^*\ X_B\ | <,\ \mathcal{M} \setminus \{,\ \},$
$a\ \ r < a\ n^2 - n$

都存在有限距离 $((\,)\,(\,))\cdot(\,)$，并且根据定义 A.1，可证明 Kemeny 度量空间是柯西收敛的。由此立即可得出一个推论：Kemeny 距离的上界和下界由有限 n 和 a的合取选择确定，其中

$$0 \cdot a \leq \rho_\kappa(a\kappa^*(X_A), \kappa^*(X_B)) \leq a(n^2 - n),$$

由于已证明 Kemeny 度量空间对 $X_n \in \mathbf{X}$ 是完备的，因此引理得证。■

定义 A.2（马尔可夫不等式）。马尔可夫不等式适用于 M 中仅取非负值的随机变量：此定义与完备度量函数的定义相一致，因此适用于所有 n，并由此定义子空间 M。马尔可夫不等式指出，在对总体进行穷尽观测后，随机变量 $\rho_\kappa(x_n)$ 满足：对于任意 $\epsilon > 0$，观测到的实现值与该变量之间的距离不大于 ϵ 的概率由不等式比率定义

$$\Pr(x \geq \epsilon) \leq \frac{E(\rho_\kappa(x_n))}{\epsilon}.$$

为便于表达该不等式，我们定义 $x \in M$，使得相对于原点置换 $I_n = 1,\ 2,\ \ldots\ldots,\ n$，有 M 个不同的排列，其距离必定落在闭区间 $[0a, \ldots, a(n^2-n)]$ 内，且发生概率为正；否则，其发生概率为 0。

引理 A. **6** *Let X be a random variable with a probability density*（马尔可夫不等式）。 *function f_X which arises as a consequence of the Kemeny distribution cumulative* $(\cdot)$ *,distribution function $F_{\kappa(X)}$ and let $X_1\ X_2 \quad X_N$ be a random sequence* $(\cdot)$ *or* $(,,\cdots)$ *,sample drawn upon said distribution. By Definition A.2 it is observed that for all,, real numbers $\rho_\kappa(x_n) \geq \epsilon$ upon which the first moment-expectation $\mu_\kappa \in \mathbb{R}$ exists, Markov's inequality holds in expectation,*

$$P(\rho_\kappa(x_n) \geq \epsilon) \leq \min\left[\frac{\mu_\kappa}{\epsilon}, a\right], \tag{22}$$

and for which a strong upper-bound upon the distance holds such that the absolute distance between the expectation and any other observable point, for known $(n\ a$ *as in equation 4 is never greater than $0 \leq \epsilon \leq a\ n^2 - n$ (Lemma A.5).* $,),(0$

Proof. 接下来是

$$
\begin{aligned}
E(x_n) &= \sum_{i=\infty^-}^{\infty^+} x_i p(x_i), s.t., \sum_{i=\infty^-}^{\infty^+} p(x_i) = 1 \\
E(x) &= \sum_{i=0}^{\epsilon} x_i p(x_i) + \sum_{i>\epsilon}^{a(n^2-n)} x_i p(x_i) \geq \sum_{i>\epsilon}^{a(n^2-n)} \epsilon \cdot \Big(x_i p(x_i)\Big) \\
E(x) &= \epsilon \sum_{i>\epsilon}^{a(n^2-n)} p(x_i) = \epsilon P(|x| \geq E(x) \geq \epsilon) \\
\frac{E(x)}{\epsilon} &= P(x \geq \epsilon).
\end{aligned}
$$

■

引理 A.7（切比雪夫不等式）。 *The first expectation, expressed* $E(\rho_\kappa(x)) = \mu_\kappa =$ $a\ \frac{n^2-n}{2}$ *and the variance of* $\rho_\kappa\ x$ *is expressed as* σ_κ^2 *for an arbitrary non-negative* ( ) ( ) , , *sequence of* t*. Then*

$$P(|\rho_\kappa - \mu_\kappa| \geq t) \leq \frac{\sigma_\kappa^2}{t^2}, \tag{23}$$

by Markov's inequality, as

$$P(|\rho_\kappa(x_n - \mu_\kappa) \geq t) = P((\rho_\kappa(x_n) - \mu_\kappa)^2 \geq t^2) \leq \frac{E\big((\rho_\kappa(x_n) - \mu_\kappa)^2\big)}{t^2} = \frac{\sigma_\kappa^2}{t^2}.$$

Proof. 这在构造上很容易理解，并且直接符合所考察的 Kemeny 度量下有限值域空间的预期行为：大于距实数 t 个标准差的距离的值，其比例不会超过某个实数上限；这表示对称覆盖的概率界为 $1 - \frac{1}{t^2}$。

由于线性函数 n 的有限距离严格已知不超过一个实数有限距离 $\frac{1}{2}\big(n^2 - n\big)$，也就是距期望的对称最大距离 $\max\ \epsilon$，因此证明了 Kemeny 度量的上尾界和下尾界，对于任何一组来自共同分布、独立观测到的 n 个实数，均满足切比雪夫不等式。■

定义 A.3（希尔伯特空间）。希尔伯特空间是一个完备度量空间（巴拿赫赋范空间），其上存在一个内积 $\mathrm{h}x,\ y\mathrm{i}$，并满足以下性质：

1. 内积具有共轭对称性，因此一对元素的内积等于交换这两个元素后所得内积的复共轭：

$$\langle y, x\rangle = \overline{\langle x, y\rangle}.$$

2. 所述内积关于其第一个（实数）变量是线性的，即对所有复数 a 和 b均成立。

$$\langle ax_1 + bx_2, y\rangle = a\langle x_1, y\rangle + b\langle x_2, y\rangle.$$

3. 元素与自身的内积是正定的

巴拿赫赋范空间上的正齐次性这一性质的存在，确立了内积的存在。由这些性质可推出度量 ρ 上柯西-施瓦茨不等式的存在，即：

$$|\langle x, y\rangle| \leq \|x\|\|y\|$$

等号成立时， *iff* x 和 y 线性相关。因此，希尔伯特空间是满足进一步条件——正共轭齐次性的完备度量空间（巴拿赫空间）：

$$f(cx) = c \cdot f(x),\ c > 0.$$

引理 A. **8** *The κ^2 function is an even function as are all con-* （偶函数）。，*structed affine linear transformations thereupon.*

定义 A.4（偶函数与奇函数）。如果对于每个输入 x，$\rho_\kappa^2\ x\ = \rho_\kappa^2\ -x$ $x\ f\ x$ ( ) ( )，则函数为偶函数。相应地，如果对于每个输入 ( ) $-\rho_\kappa^2\ -x$ $\rho_\kappa^2\ x\ = \rho_\kappa^2\ -x$ ( )，则函数为奇函数。如果函数满足 ( ) ( )，则它是偶函数。相应地，如果函数满足 $\rho_\kappa^2(x) = -\rho_\kappa^2(-x)$，则它是奇函数。如果函数不满足这两种规则中的任何一种，则它既不是偶函数，也不是奇函数。

Proof. 考虑将 $\kappa(\cdot)$ 函数应用于长度为 n 的实向量 x $= \{1, \ldots,$n$\}$。对于这种映射 $\kappa(x)$，会得到一个 $n \times n$ 阶方阵。根据这种矩阵的构造性定义（方程 1b），我们得到四个条件，其中三个是独立的：$\{x_i \leq x_j \to -1, x_i > x_j \to 1, x_i = x_j \to 0\}$，所有 n 元素均由此得出。对于条件 3，注意到 $0 = -1$(0)，因此， x 上相同元素之间的比较既非偶也非奇。对于其他条件， $\kappa(x) = -1\kappa(x)^\intercal$，因此单独来看， $\kappa(\cdot)$ 函数是奇函数。然而， $\kappa^2(x) = \kappa^2(-x)$，平方元素因此只能是 {0, 1}；由于除非函数为奇函数，否则不会出现 0，而且实数平方后不会得到负元素，因此任意两个 κ 函数的仿射线性函数处处为偶函数。■

引理 A.9 *We next proceed to prove that the second moment, the variance, upon the Kemeny measure space is always finite subject solely to the assumption of a collection of independent observations.*

Proof. 对于 $x \in X^{n\times 1}$，注意到 $\kappa(x) \forall x \in$ M 始终是一个方阵。根据度量空间的定义，任意度量定义域中的给定 x 到其自身的距离始终为 0。然而，方差的二阶矩是基于 κ 而定义的，而非其交叉乘积，因此这些元素实际上表示长度为 n 的实向量上的表列差异率（唯一性）。因此，令 $\kappa^r(\cdot)$ 的多项式函数表示向量 x 的 r^{th} 矩；其一阶矩已证明为 0，由此可通过中心极限定理得到 Kemeny 度量上的原始矩和中心矩。因此，逻辑上可得，Kemeny 方差存在有限上下界，这正如任何全有界空间（例如 Kemeny 度量）所预期的那样。

我们首先确立 $\kappa(x)$ 存在一个下界，后文将证明该下界为 0。二阶矩的这个定义等价于一个有限实常数向量，其中所有 n 元素都相同。因此，在 x时，观测到的得分不会有任何此类变异。于是，对于任意 $\kappa(x) \in$ M，都存在一个对 $i, j = 1, \ldots, n$ 进行边缘化的向量。

$$\sum_{j=1}^{n} \kappa_{ij}(x) = \sum_{i=1}^{n} \kappa_i(x) = 0 \equiv \sum_{i=1}^{n} -1 \sum_{j=1}^{n} \kappa_{ij}(x)$$

存在 n 个向量，其中 x 中的所有元素都相同，因此，任何 n元组的所有映射都是由 n 0组成的向量n 因为根据 $\kappa(\cdot)$ 函数的第三个条件，所有 $x_1 \quad x_n \rightarrow a \quad 0_1, \quad 0_n{}^{\intercal} \quad n$ 两两比较都由满足 $(,\cdots)\cdot(\cdots)$ 的映射构成。因而可以看出，这种序列的平方始终为 0，因为 $\kappa(\cdot)$ 的 $r=2$ 统计功效的一般表达式可用于表示二阶矩。

$$\sigma_{\kappa}^{2}(x)=\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{ij}^{2}(x)\geq 0. \tag{24}$$

■

引理A.10. *We next proceed to prove that the upper-bound, for any n is also finite and must sum to a constant real which will be shown to be proportionate to a scalar function of n such that the upper bound of the variance is defined to be always in the , , interval ring of $\kappa \propto [0,1]$. It will also be shown that when σ_{κ} is maximised the sub-space of the domain is equivalent to that of the traditional S_n group from which the original order-statistics concept arises.*

Proof. 首先请注意，对于长度为 n 的向量 x 的所有 M 种排列，存在 M 个 κ_i 映射，每个映射的期望均为 0。随着实数重复次数减少，预期方差会达到最大值；这一点可从方差的定义看出：当实数得分构成一个常量向量时，方差为 0，所有 n 元素都具有相同的值。随着可观测得分的多样性增加，方差也必然增加。因此，当斜对称矩阵的对角线外不存在 $\kappa\left(x\right)=0$ 时，向量 x 的排序的方差达到最大值；此时，由 κ 函数生成的映射中，实数大小总是等于自身。

假设该反对称 κ 矩阵中有 $2(n**2-n)$ 个正值，其最大值 $\kappa_i^2(x)$不大于$a^2(n^2-n)$，则此类矩阵的平方的期望值为.....。要观测到这样的测量结果，必须有一个向量 x，其中的某个元素 x 满足以下两个条件之一：(1) $(a_{ij})^{\intercal}\cdot\ a_{ij}\equiv -(a_{ij})^{\intercal}a_{ji}$；或 (2) $(a_{ij})^{\intercal}\cdot\ a_{ij}\neq -(a_{ij})^{\intercal}a_{ji}.$。我们首先注意到，只有在$\pm a$的符号——即向量中两个元素的乘积——必定为 0 的情形下，才会出现并列值；根据公式 1b，此时便会出现并列值。对于映射 $x\rightarrow\kappa(x)$中的其他所有元素，κ矩阵的反对称性质会迫使$\sqrt{.5}\cdot\ n$ 个元素的和大于 $a(n^2-n)$。这要求 $-(a_{ij})^{\intercal}\cdot(a_{ij})=-(a_{ij})^{\intercal}\cdot(a_{ij})$，因而对于 $(-a)=a, a\in\{-a,0,a\}\setminus 0=\varnothing$。这是一组空解，因此不可能发生：由此证明，对于 $a>0, n\geq 0,\ 0\leq\sigma_{\kappa}^{2}\leq a(n^2-n)$的上界最大值为.....。令$a=\sqrt{.5}$，则在 $\frac{n^2-n}{\sqrt{.5}^2}$的支撑集上可得到最大方差。因此，对于所有 n，根据 Kemeny 距离的刻画，可以观察到不存在这样的实数 x 向量集合 n，使得其方差不存在于 0 到 1 2 (n^2-n) 之间。■

–

引理 A. 11 *Let F_{κ} be a distribution function for the Kemeny metric of a random.·· variable realised upon $\overline{\mathbb{R}}$. For each $\epsilon>0$ there exists a finite partition of the extended real line such that for an orderable sequence $\infty^{-}\leq t_1\leq \quad \leq t_k\leq\infty^{+}$ by Lemma 1.6 . ,and there exists $0\leq j\leq k-1$*

$$F_{\kappa}(t_{j+1})^{-}-F_{\kappa}(t_j)\leq\epsilon.$$

Proof. 设 $0 < \epsilon$ 满足存在单调收敛的条件。允许$t_0 =$ 无穷大$\overline{\mathbb{R}}$，对于该情况， $j \geq 0$ 我们定义 $t_{j+1} =$ 上确界$\{z : \ F(z) \leq F(t_j) + \epsilon\}$。根据右连续性，存在一个有限的步骤序列，使得该定义在这些步骤处不连续。我们观察到，对于我们的 κ 函数定义，这种情况不会出现在任何可数有限总体中。因此，定义了一个从 F的期望出发的单调递减距离序列的转移状态，从而对于紧致且全有界的上确界 Kemeny 距离 $n^2 - n$，任意有限样本都具有有限距离。■

引理 A. **12** *The Kemeny estimator functions satisfy the Cramèr-Rao lower bound* 。M *upon the population constructed of asymptotic limit on* n.

Proof. 估计量函数的无偏性已在引理 1.2 中得到证明，并且根据引理 1.5，观察到该估计量函数也服从渐近正态分布。由于对于所有有限的 n，估计量函数的方差严格为次高斯，该方差是一个标量常数比，并且作为所有数据分布的线性函数收敛于 1。在这些条件下，根据高斯-马尔可夫定理（定理 1.3），渐近方差的增长趋近于上述正态分布的渐近方差；当相对于 n 的极限为 0 时，由此得出结论，获得 Cram`er-Rao 下界

$$\lim_{n \to \infty^+} \frac{n^n - n}{n^n} \sigma_\kappa^2 \to \sigma^2,$$

正如中心极限定理所要求的那样。引理 A.13 保证了收敛性：对于 Kemeny 度量上的任意随机变量，其距离将收敛于 0，并且这与等距欧几里得距离下的解等距。可测性上界确保，即使收敛不趋于 0（如模型设定错误，且样本上的条件贝叶斯错误率大于 0 时），误差仍始终能够被唯一识别。

这些性质表明，两个估计量的对偶构造是互补的，而非相互排斥的，因此，根据定义，对有限样本上的两个度量空间进行考察必然会降低熵，从而增加通过两个函数空间获得的、趋向于相互一致的最优解的信息。■

引理 A.

13 *The total variational distance between any Euclidean linear function and* 。*an unbiased Kemeny linear function is lower-bounded by 0.*

Proof. 设来自次高斯场和高斯场的随机变量的所有有限集合之和是稳定的，允许 Z 为长度为 n 的实现变量随机集，并令 X 为一个独立随机变量，其分布函数为F_κ，且其中 $N \in U(\mathsf{M})$ 是定义 F_κ 的紧且全有界支撑集的任意有限数值。因此， X 超过 N 的尾概率就是 $X \notin U(\mathsf{M})$ 的概率。若 F_κ 为高斯分布，则 $N \ni F$ 的概率为 0，因此 N 关于 F_κ 不可测：

$$e_G = \Pr(|X| > 0, \ldots, N) = \Pr(X \notin U(\mathcal{M}) = F(N) - \lim_{\varepsilon \to 0^+} F(-N - \varepsilon).$$

现在考虑包含 G 的有限伽罗瓦域 G，它由同样基于随机变量 X 测量的贝塔-二项分布组成。全

变差距离 $Y = G(Z)$ 与 $X = F(Z)$ 之间的变差距离，是它们的概率分布之间的全变差距离，其中 F 是定义在 F 上的（概率测度）博雷尔集σ代数，而 G 是定义在 G 上的相应集合，由此可得:

$$\|X - Y\|_{TV} = \sup_{A \subset \mathcal{F}} |\Pr(X \in A) - \Pr(Y \in A)| .$$

马尔可夫（引理 A.6）和切比雪夫（引理 A.7）不等式显然适用于任何以 Kemeny 距离函数度量的次高斯函数。这是因为其具有有限期望（紧且全有界，引理 1.1），因为 $|X|$ 的期望不超过共同总体的积分，而该积分是有限的（定义 1.2），因此对于任何出现的随机变量，它们始终是可测的。

假设 N 足够大，使得 Pr($X \in U$(M)) 显然非零，并满足 $n > 2$ 所需的所有条件。将 X 截断于 N 的一个有理分数处，由此消除其超过 N 的所有可能性；对于 G，新的分布函数在 x 处对 $x < -N$ 为 0，在 $x \geq N$ 处对 $F(x)$ 中所定义的值为 1，其余情况下则等于

$$F_{[N]}(x) = \frac{F(x) - \lim_{\varepsilon \to 0^+} F(-N - \varepsilon)}{1 - e_F(N)}.$$

反之，假设 $F(x)$ 是一个有偏函数，允许 $0 < p < 1$ ，并允许 γ_M^2 为任意正数，表示我们希望将 X 在由 Kemeny 度量定义的期望渐近总体上的分布期望平移的幅度。若 X 是任意具有有限期望的随机变量，则令 $\varepsilon > 0$。选取一个 N，使得 $e_F(N) \leq \frac{\varepsilon}{2}$，并在 N 处截断 X。由此可得， $E(X_{[N]}) = 0 + \gamma_M^2$ 的均值可表示为一个 $\frac{\varepsilon}{2}$ 混合，这会使全变差距离至多改变 $\frac{\varepsilon}{2}$。由次可加性可得

$$\|X - Y\|_{TV} \leq \|X - X_{[N]}\|_{TV} + \|X_{[N]} - Y\|_{TV} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \tag{25}$$

因为不存在元素 $N \notin U$(M)。

因此，这证明对于任何 X 都存在有限期望，并且对于 Kemeny 度量下具有有限期望且独立分布的随机变量，总有办法截断 X 并将其均值平移至 γ_M^2，无论 γ_M^2 取何值，都不会使其在全变差距离上的变化超过 ε 。这些构造为 kYk 给出了一个上界：它不大于|$\frac{N}{2}$|中的较大者，|$\frac{N}{2}$|表示位于 $2^{(\gamma_M^2 - E_{(X))\varepsilon}}$处的原子位置的绝对值。因此， Y的尾部为零，这使其成为次高斯。由于 ε可以任意小，距离唯一可能的下界是零；而任何紧且全有界或完备度量空间的仿射线性不变性，都使估计参数几乎必然位于 U(M)，对所有有限的 n均如此。■

引理 A.14. *The total variational distance between any Euclidean linear function and an unbiased Kemeny linear distance function is upper-bounded by a finite function of* $\Pr_\kappa$ (X) *corresponding to a finite distance of* $|\frac{n^2 - n}{2}| \geq 0$.

Proof. 对于度量空间上的任意仿射线性函数以及任意点集，下界始终有限，且不失一般性可视为 0。对于任意度量，系统 $\rho(\hat{Y},\ Y)$ 的性能上界 ρ 可能无法确定（这是因为扩展实数轴上缺少紧且全有界的定义域）。对于欧几里得度量空间，这一问题通过使用

“近似正确系统”来解决，它们将空间的测度限制为有限值。对于性能取值于扩展实数轴的情形，结合 Kemeny 度量 ρ_κ 可使任意测度空间的矩有限；我们证明，对于任意齐次函数空间（即针对共同总体的仿射线性函数空间），全变差距离几乎必然存在有限上界。

允许 X 不具有有限期望，同时也是一个随机变量。由于X 不可测，因此其矩在欧几里得度量空间中不属于实数直线，概率界只有在 Kemeny 度量空间中才几乎必然（以概率 1 可测）得到保证，其中 Kemeny 距离支撑集定义为 U(M)且以 0 为中心。因此，对于任意非常量向量 X， X 与 Y之间所有可测距离都位于 U(M)上，实现误差为 sup $|\frac{n^2-n}{2}|$.。于是，最大误差几乎必然位于以 0 为中心的邻域内，Pr(2|($\frac{n^2-n}{2}$)| $\in$ U(M)) $=$ 1，从而保证收敛。■

推论 A. **14 1** *Suppose* X_1 X_n *is a sequence of independent random vari-*. . {, .. . , }, *ables each with finite expected value* μ_i *and variance* σ_i^2. *Define* $s_n^2 = \sum_{i=1}^n \sigma_i^2$; *then for some* $\delta > 0$, *Lyapunov's condition is satisfied, then a sum of* $\frac{X_i - \mu_i}{s_n}$ *converges in distribution to a standard normal random variable, as* $n \to \infty^+$:

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0.$$

Proof. 可以看出，对于任意有限的 n $=$ 0 $\forall n = 1 \ldots \mathbb{N}^+ < \infty^+$ κ ，Lyapunov条件都成立。这是因为，任意希尔伯特空间上的函数及其任意仿射线性变换（定义A.3；引理A.8和引理A.3）都具有偶函数性质，因此 Kemeny 距离的任意分布 $\gamma_\kappa^3(\mathcal{M})$,,,的偏度为零。

由于 X_1， X_2， ..... X_n $\in$M 相互独立同分布， $\sum_{i=1}^{\mathcal{M}} E[|X_i|^3] = nE(|X_n|^3)$。根据 Lyapunov不等式， $\infty > E(|X_n|^3) \geq [E(|X_n|^2)]^{\frac{3}{2}} \geq (\sigma_n^2)^{\frac{3}{2}} = \sigma_n^3$，这由偶函数的有限性和全紧致性可知成立。由于 $s_n^3 = (\sum_{i=1}^n \sigma_i^2)^{\frac{3}{2}} = n^{\frac{3}{2}}\sigma^3$，因此可得 $\frac{nE|X_1|^3}{\sqrt[1.5]{n}\sigma_\kappa^3} = C\frac{1}{\sqrt{n}} \to 0$。

对于任意正整数 n，可以得出该函数的分布是对称的。由此可知，对于大于 2 的矩，边界条件的期望 $2 + \inf\ \delta = 2 + 1 \implies \gamma_\kappa^{(2+\inf\delta)} = 0$ 根据该 κ 函数在 n 实数上的偶性（引理 A.8），中心极限定理因此对任意一列 n 独立分布的实现值均成立。■

推论 A.14.2. *Kolmogorov's law applied to the Kemeny metric holds that the sample average converges almost surely to the expected value over the field of permutations*M*:*

$$\lim_{m\to\infty^+} \overline{x}_m \xrightarrow{a.s.} \mu \equiv \Pr\left(\lim_{m\to\infty^+} E(x_m)_\kappa = \mu_\kappa\right) = 1.$$

Proof. 强大数定律适用于具有可排序期望值的独立同分布随机变量；如果这些变量相互独立且同分布（例如在 Kemeny 度量下），则它们必须具有期望值；引理 A.10），而如果加数相互独立但不同分布，且具有有限的二阶矩（引理 A.9 和引理 A.10），则可得:

$$\lim_{m\to\infty^+} \overline{x}_m - E\left[\overline{x}_m\right] \xrightarrow{\text{a.s.}} 0,$$

因此，对于来自包含 M 个排列的总体的任意独立分布样本，Kemeny 度量空间上的柯尔莫哥洛夫收敛定律均有效成立。■

B 排列空间 M 的密度

$$n^n < \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}}$$

$$n\log(n) < \frac{1}{2}\log(2\pi n) + n(\log(n) - \log(e)) + \Big(\log(1) - \log(12n+1)\Big)\log(e) < n!$$

$$< n\log(n) < \frac{1}{2}\log(2\pi n) + n(\log(n) - \log(e)) + \Big(\log(1) - \log(12n)\Big)\log(e)$$

$$n\log(n) < \frac{1}{2}\log(2\pi n) + n(\log(n) - 1) + \Big(0 - \log(12n+1)\Big) < n!$$

$$< \frac{1}{2}\log(2\pi n) + n(\log(n) - 1) + \Big(0 - \log(12n)\Big)$$

$$n\log(n) < \frac{1}{2}\log(2\pi) + n(\log(n) + \log(n) - 1) \equiv (\log(n)) + \Big(0 - \log(12n+1)\Big)$$

$$n^n < \sqrt{2\pi n}\cdot(\frac{n}{e})^n$$

$$n\log(n) < \log(2\pi n) + n\Big(\log(n) - \log(e)\Big)$$

$$n\log(n) < \frac{1}{2}\log\Big(2\pi n\Big) + n\log(n) - n\log(e)$$

$$0 < \frac{1}{2}\log\Big(2\pi n) - n$$

$$2n < \log\Big(2\pi n\Big)$$