中心极限定理的验证与探究

SJTU-SEIEE　cny123222

摘要

本文聚焦于数理统计的基础性定理——中心极限定理的验证与探究。通过使用 KL 散度衡量收敛精度以及 KS 检验判断显著性水平，深入分析了中心极限定理在不同条件下的收敛特性。研究表明，样本量、峰态、偏态以及长尾对收敛精度有着显著影响。其中，峰态通过不同自由度的 t 分布进行研究，偏态利用不同自由度的 χ^2 分布，长尾则借助对数正态分布和 Pareto 分布展开讨论。研究结果不仅深化了对中心极限定理的理解，也为实际应用提供了理论支持。

1　研究背景

中心极限定理是统计学中的重要基础理论，描述了在特定条件下大量独立随机变量的和逼近正态分布的现象。该理论被广泛应用于各个领域，从金融领域的风险评估、股票市场的波动分析，到生物学中对种群特征的研究、医学领域的疾病诊断等。

本研究旨在探究中心极限定理的适用性，并通过 KL 散度和 KS 检验等方法，评估其在不同总体分布下的收敛表现。同时，我们将着重考察样本量、峰态、偏态以及长尾对收敛精度的影响。这一研究将有助于加深对中心极限定理的理解，并为实际数据分析提供更为准确和有效的方法。

2　研究方法

2.1　总体思路

根据独立同分布的中心极限定理，设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的简单随机样本，且总体期望为 μ，方差为 σ^2，则样本均值 $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ 在样本量足够大时近似服从 $N(\mu, \frac{\sigma^2}{n})$，即归一化后的 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 近似服从标准正态分布 $N(0,1)$。

在实验中，我们给定样本量及总体分布，随机生成简单随机样本，计算样本均值并进行标准化处理。重复该抽样 10000 次，将得到的均值分布与标准正态分布进行比较。为了评估收敛精度，我们采用了 KL 散度 [1] 和 KS 检验方法 [2]。

2.2　KL 散度

KL 散度（Kullback-Leibler Divergence）是一种用于衡量两个概率分布相似度的指标。在我们的研究中，KL 散度被用来衡量样本均值分布与正态分布之间的差异程度。从信息论的角度来看，KL 散度可以表示为总体分布和近似分布之间的信息差异。

对于两个概率分布 $P(x)$ 和 $Q(x)$，KL 散度定义如下：

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{1}$$

其中，$p(x)$ 和 $q(x)$ 分别代表总体分布和近似分布的概率密度函数。

在我们的研究中，KL 散度将帮助我们量化标准化的样本均值分布与标准正态分布之间的差异，为评估收敛精度提供重要的指标。具体来说，KL 散度越小，说明两个分布越接近，收敛精度越高。

2.3 KS 检验

KS 检验（Kolmogorov-Smirnov Test）是一种非参数检验方法，用于比较样本分布与参考分布之间的差异。在本研究中，KS 检验用于评估标准化样本均值分布是否与标准正态分布一致。

KS 检验基于分布函数的最大差异定义统计量：

$$D_n = \sup_x |F_n(x) - F(x)| \tag{2}$$

其中 $F_n(x)$ 是样本分布的分布函数，$F(x)$ 是参考分布的分布函数。

在本研究中，KS 检验的原假设 H_0 是标准化的样本均值分布与标准正态分布相同，备择假设 H_1 是它们不同。我们设定显著性水平 $\alpha = 0.05$，计算 KS 统计量 D_n 及 p 值，并将 p 值与临界值比较。若 $p < 0.05$，则认为样本均值分布与正态分布的差异显著，否则认为二者相符。通过 KS 检验，我们能够从统计显著性的角度判断标准化样本均值分布与标准正态分布的接近程度，与 KL 散度形成互补，为中心极限定理的验证提供更全面的证据。

3 研究过程及结论

3.1 样本量对收敛精度的影响

由中心极限定理可知，当样本量 n 足够大时，样本均值将近似服从正态分布。因此不难想到，样本量 n 的大小会对样本均值的收敛精度产生影响，即样本量越大，收敛精度越高。

图1、图2、图3和图4分别展示了不同样本量 $n = 1, 5, 10, 50, 100$ 下均匀分布总体 $U(0,1)$、t 分布总体 $t(5)$、χ^2 分布总体 $\chi^2(3)$ 和指数分布总体 $E(1)$ 的样本均值分布的收敛情况。其中蓝色条形图表示标准化后的样本均值的分布，橙色曲线是标准正态分布的概率密度曲线。图5展示了这些总体分布下 KL 散度随样本量 n 的变化关系。

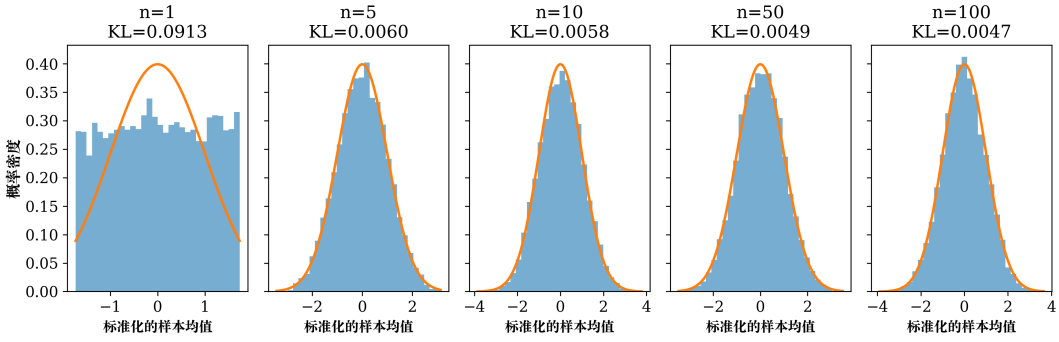

图 1　不同样本量 n 下均匀分布总体 $U(0,1)$ 样本均值分布的收敛情况

由上述图像可以看出，随着样本量 n 的增加，各总体分布下样本的 KL 散度逐渐减小，标准化的样本均值分布逐渐逼近标准正态分布，表明收敛精度随着样本量的增大而提高，这与我们的猜想相符。同时，可以注意到，当样本量 $n = 100$ 时，各标准化的样本均值分布已经与标准正态分布

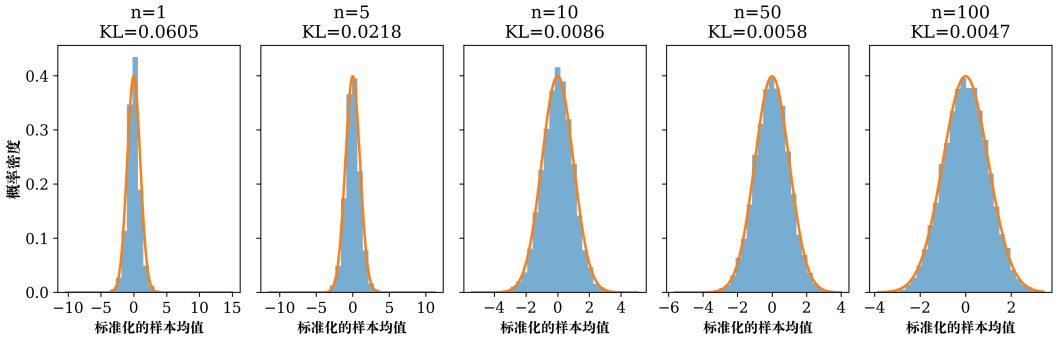

图 2　不同样本量 n 下 t 分布总体 $t(5)$ 样本均值分布的收敛情况

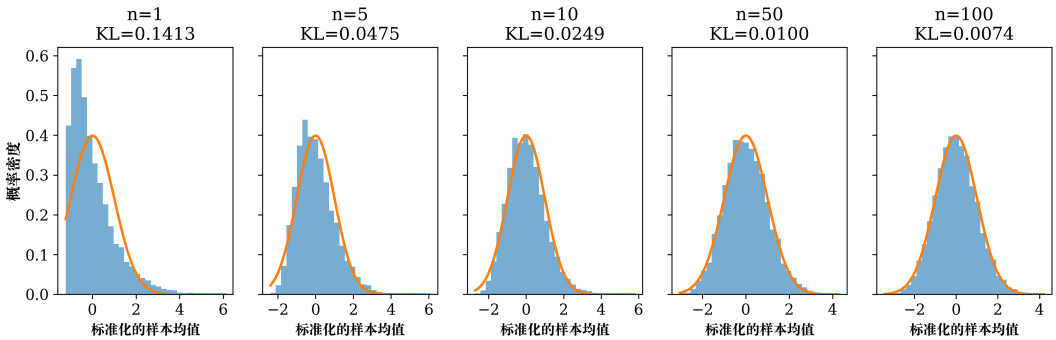

图 3　不同样本量 n 下 χ^2 分布总体 $\chi^2(3)$ 样本均值分布的收敛情况

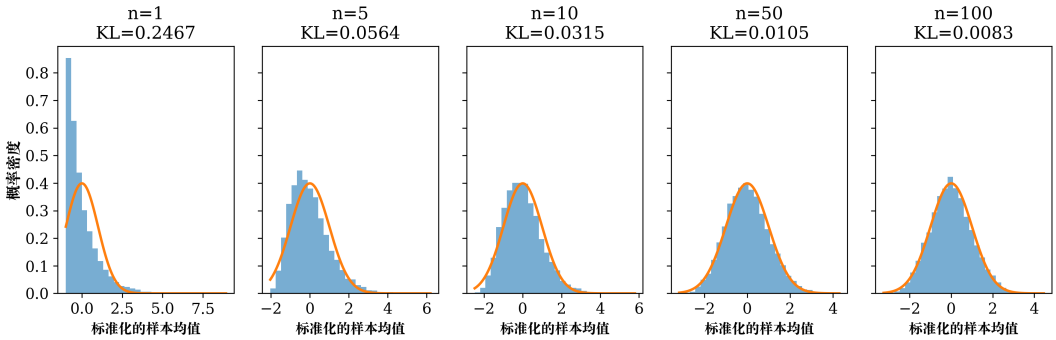

图 4　不同样本量 n 下指数分布总体 $E(1)$ 样本均值分布的收敛情况

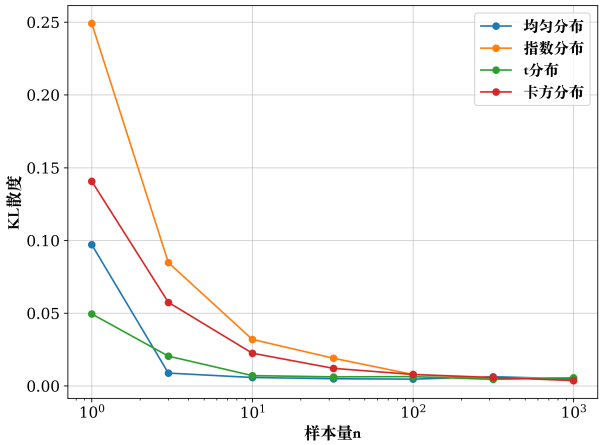

图 5　不同总体分布下样本量对收敛精度的影响

十分接近。进一步从图5可以看出，此时各分布总体下的 KL 散度已经接近于 0，且样本量继续增大对 KL 散度降低效果不明显，可以认为样本均值已经基本收敛到正态分布。

这些观察结果进一步验证了随着样本量的增加，样本均值分布逐渐趋近于正态分布，同时也强调了收敛精度对于样本量增大而提高的重要性。这些发现为我们深入理解中心极限定理提供了直观的支持，同时也为统计推断中样本量选择提供了有益的参考。

3.2 峰态对收敛精度的影响

由图5可以发现，在我们选取的四种总体分布中，t 分布总体的样本均值收敛速度较快，这可能与其与正态分布的相似性有关。图6展示了 t 分布的特性：随着自由度 m 的增大，其峰逐渐变尖，尾逐渐变薄，在 $t \to \infty$ 时，将收敛到标准正态分布。因此我们猜想，样本均值的收敛精度可能与总体的峰态有关。对于 t 分布而言，自由度 m 越大，相同样本量下的收敛精度越高，也即收敛到相同精度所需的样本量越少。

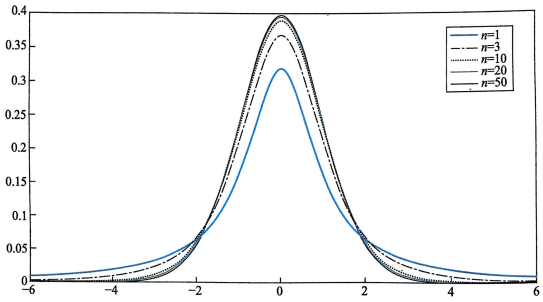

图 6　不同自由度 t 分布的概率密度曲线 [3]

为此，我们从不同自由度的 t 分布总体中分别抽样，用 KS 检验的方法计算出不同自由度下使得标准化样本均值分布与标准正态分布差异不显著（$\alpha = 0.05$）的最小样本量 $n_{\min}$，观察到 $n_{\min}$ 与 t 分布总体的自由度 m 的关系如图7所示（注意纵轴采用对数坐标）。需要注意的是，为了获得更加准确的 $n_{\min}$，我们需要在相同自由度和相同样本量下多次取样并进行 KS 检验。然而，多次检验会使得我们犯第一类错误的概率增大，因此我们采用了 **Benjamini-Hochberg 方法** [4] 进行校正。

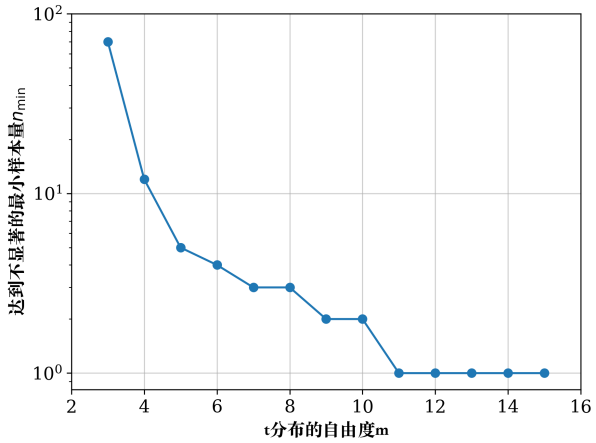

图 7　t 分布总体达到不显著水平的最小样本量与自由度 m 的关系

由图7可以看出，随着 t 分布总体的自由度 m 增大，即 t 分布的峰不断变尖并接近标准正态分布，使得标准化样本均值分布与标准正态分布差异不显著的最小样本量 $n_{\min}$ 逐渐减小。特别地，

当自由度 $m \geq 11$ 时，只需要一个样本就可以让样本均值分布与正态分布的差异不显著，也就表明了总体分布已经和正态分布十分接近。这一发现进一步验证了峰态对收敛精度的影响，并强调了在统计推断中样本量和总体分布特性的重要性。

3.3 偏态对收敛精度的影响

由图5可以发现，在我们选取的四种总体分布中，指数分布和 χ^2 分布这两个偏态总体收敛速度最慢，这符合我们的直觉，即将不对称分布的总体“矫正”到正态分布，需要抽取更多的样本。为了研究总体的偏态对收敛精度的影响，我们选取 χ^2 分布作为研究对象。图8展示了 χ^2 分布的特性：当自由度 m 很小时，其密度函数偏向左侧，而当 m 很大时，其逐渐逼近正态分布，即由偏态过渡到正态。

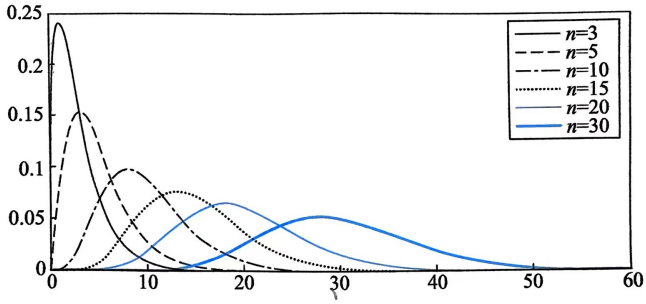

图 8　不同自由度 χ^2 分布的概率密度曲线 [3]

类似对峰态的研究，我们从不同自由度的 χ^2 分布总体中抽样，用 KS 检验的方法计算出不同自由度下使得标准化样本均值分布与标准正态分布差异不显著（$\alpha = 0.05$）的最小样本量 $n_{\min}$，观察到 $n_{\min}$ 与 t 分布总体的自由度 m 的关系如图9所示（注意纵轴采用对数坐标）。

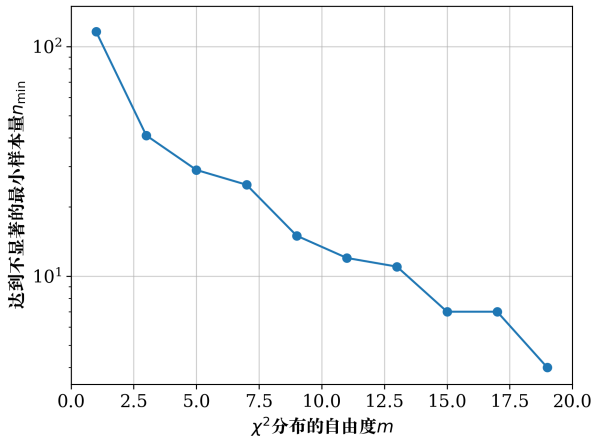

图 9　χ^2 分布总体达到不显著水平的最小样本量与自由度 m 的关系

由图9可以看出，随着 χ^2 总体的自由度 m 增大，即 χ^2 分布由偏态逐渐转为正态，使得标准化样本均值分布与标准正态分布差异不显著的最小样本量 $n_{\min}$ 逐渐减小。我们可以推测，偏态对样本均值的收敛精度有显著影响，且对于强偏态分布，纠正总体形状对正态逼近的影响大于增加样本量的直接效应。具体而言，当 $m < 10$ 时，$n_{\min}$ 随 m 的增加快速下降，而当 $m > 15$ 时，$n_{\min}$ 的变化趋于平缓。这一发现对实际抽样设计具有重要意义。对于强偏态总体（如低自由度的 χ^2 分

布），需要显著增大样本量才能获得与正态分布足够接近的样本均值分布，这为偏态数据的统计推断提供了指导。

3.4 长尾对收敛精度的影响

长尾对收敛精度的影响是统计研究中一个重要的课题。长尾分布通常指尾部较厚、尾部数据点频率较低的分布，其特点是在尾部存在极端值或异常值，因此这种分布可能会对样本均值的收敛速度产生影响。

为了深入研究长尾对收敛精度的影响，我们选择了**对数正态分布** [5] 总体和 **Pareto 分布** [6] 总体作为研究对象。首先简单介绍这两种分布的特点及其概率密度函数。

对数正态分布是指随机变量 X 的对数 $\ln(X)$ 服从正态分布 $N(\mu, \sigma^2)$ 的分布。其概率密度函数为：

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), \quad x > 0 \tag{3}$$

Pareto 分布是一种典型的长尾分布，常用于描述富豪分布、城市人口等尾部较重的数据。其概率密度函数为：

$$f(x;a) = \begin{cases} ax^{-a-1}, & x \geq 1, \\ 0, & x < 1, \end{cases} \tag{4}$$

其中 $a > 0$ 是形状参数（控制尾部厚度）。

这两种分布都具有长尾特性，即随着 x 的增大，尾部衰减较慢，尾部的极端值可能显著影响样本均值。其概率密度函数如图10所示。

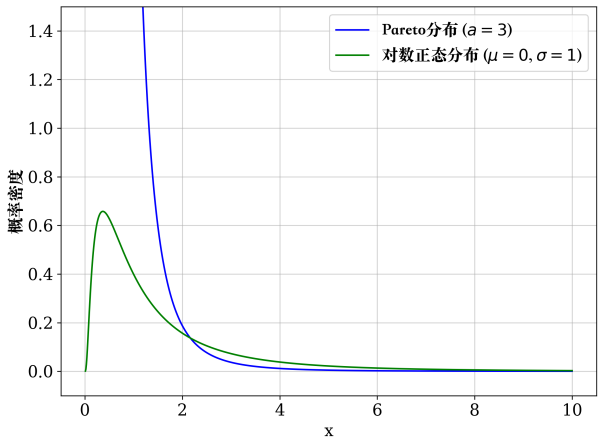

图 10　对数正态分布和 Pareto 分布的概率密度曲线

我们将探讨这两种总体分布在不同自由度下的收敛速度，观察它们的峰值变化以及是否会逐渐接近标准正态分布。这将有助于我们更好地理解长尾分布对样本均值收敛精度的影响，以及样本量选择在该类分布下的重要性。

图11和图12分别展示了不同样本量 $n = 1, 5, 10, 50, 100$ 下对数正态总体样本（$\mu = 0, \sigma = 1$）和 Pareto 分布（$a = 3$）总体的样本均值分布的收敛情况。可以看出，当样本量 n 增大到 100 时，两个样本均值的峰仍然右偏，KL 散度仍较大，收敛效果并不好。这是由于长尾分布的尾部极端值出现概率高，在小样本量下，样本均值容易受到极端值的显著影响，导致其偏离总体均值较远。尽

管样本量 n 增大后，均值会逐渐向总体均值收敛，但由于长尾特性，这种收敛的速度明显慢于对称分布或短尾分布。

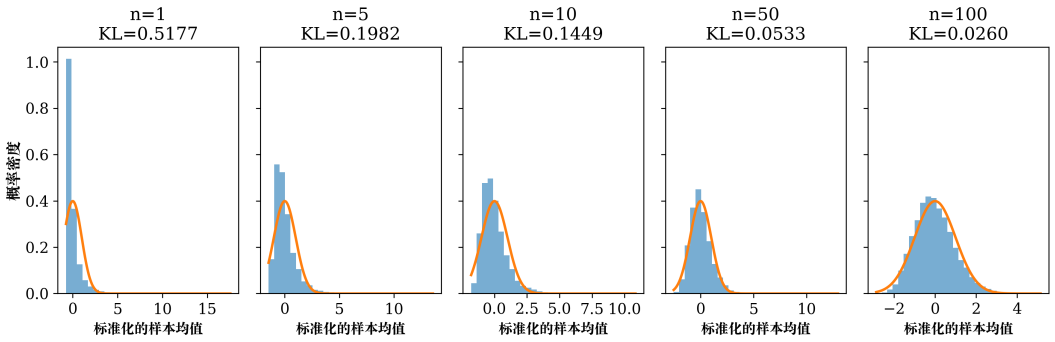

图 11　不同样本量 n 下对数正态分布总体（$\mu = 0, \sigma = 1$）样本均值分布的收敛情况

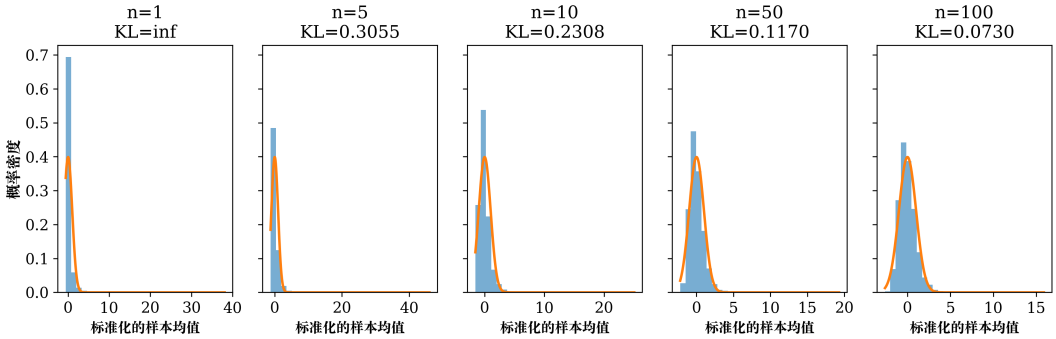

图 12　不同样本量 n 下 Pareto 分布总体（$a = 3$）样本均值分布的收敛情况

图13展示了对数正态分布总体和 Pareto 分布总体在不同样本量 n 下的样本均值分布的收敛精度，并与其他分布总体进行了对比（各分布参数均与之前一致）。从图中可以看出，对于均匀分布、指数分布、t 分布和 χ^2 分布等总体，样本均值在样本量较小时便快速接近标准正态分布，其 KL 散度迅速下降，表明收敛效果较好。然而，对于长尾分布（如对数正态分布和 Pareto 分布），总体的样本均值分布收敛速度显著慢于其他分布。其中，对数正态分布总体的样本均值即使在样本量 $n > 1000$ 时才勉强接近收敛，而 Pareto 分布总体的样本均值在 $n = 10000$ 时仍未能良好收敛，KL 散度仍然保持下降趋势。这表明，长尾分布在尾部的极端值对收敛精度的影响十分显著。

表1进一步定量展示了不同总体分布下标准化样本均值与标准正态分布差异不显著所需的最小样本量。可以看出，偏态分布所需的样本量远大于正态的分布，而长尾分布所需的样本量又远大于普通偏态分布，其中 Pareto 分布所需的最小样本量达到了 57082，远超其他分布。

均匀分布	t 分布	χ^2 分布	指数分布	对数正态分布	Pareto 分布
5	10	474	512	5183	57082

表 1　不同总体分布下达到不显著水平（$\alpha = 0.05$）的最小样本量

综合来看，长尾分布对中心极限定理的收敛速度有着极其显著的影响。因此，在实际应用中，对于长尾分布的样本分析尤其需要关注样本量的选择，盲目使用小样本量可能导致统计推断的偏

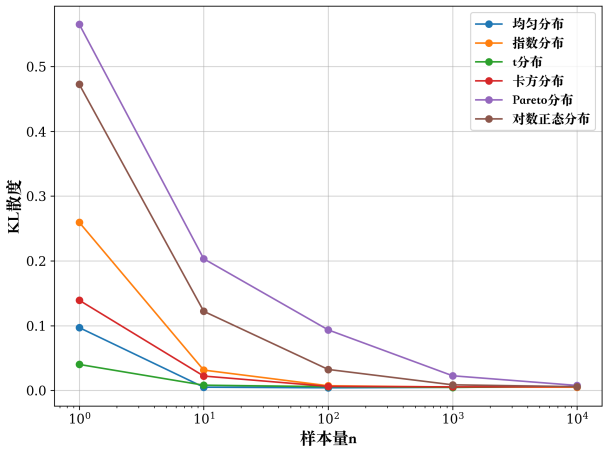

图 13　对数正态分布总体和 Pareto 分布总体的收敛精度与其他分布总体的对比

差，无法正确反映总体特征。这一结果强调了在长尾分布数据处理中的审慎性，以及对大样本量需求的不可避免性。

4　总结及展望

本研究围绕中心极限定理的验证与探究展开，通过理论分析和实验模拟，全面考察了样本量、峰态、偏态以及长尾对收敛精度的影响。

- **样本量：**样本量的增加使得样本均值分布逐渐趋近于正态分布，收敛精度提高。
- **峰态：**通过对不同自由度的 t 分布研究发现，自由度越大，样本均值收敛速度越快，收敛精度越高。这表明峰态与收敛精度之间存在密切关系。
- **偏态：**以 χ^2 分布为例，研究表明偏态对样本均值收敛精度有显著影响。随着自由度增大，偏态逐渐转为正态，所需样本量减小。
- **长尾：**对数正态分布和 Pareto 分布等长尾分布在样本均值收敛过程中表现出较慢的收敛速度，极端值对收敛精度影响显著。

本研究对于深入理解中心极限定理具有重要意义。通过对中心极限定理的验证和探究，我们不仅揭示了不同因素对收敛精度的影响，还为实际应用提供了理论指导。如在处理偏态或长尾数据时，必须充分考虑样本量的大小。

基于本研究，未来可进一步定量刻画收敛精度与偏度、峰度之间的关系。例如，通过建立数学模型，将偏度、峰度与样本均值收敛精度的量化指标进行关联。具体而言，对于不同偏度和峰度的分布，研究其在不同样本量下的收敛精度变化规律，从而更准确地描述中心极限定理在不同条件下的收敛特性。

参考文献

[1] Wikipedia contributors. Kullback–leibler divergence, 2024.

[2] Wikipedia contributors. Kolmogorov–smirnov test, 2024.

[3] 卫淑芝，熊德文，皮玲. 大学数学概率论与数理统计：基于案例分析. 北京：高等教育出版社, 2020.

[4] Wikipedia contributors. False discovery rate, 2024.

[5] Wikipedia contributors. Log-normal distribution, 2024.

[6] Wikipedia contributors. Pareto distribution, 2024.