Statistics and Application 统计学与应用, 2016, 5(1), 58-63
Published Online March 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.51007

Hans汉斯

A Class of Generalized Normal Distribution and Application

Xiaoqing Liu, Weihua Zhao

School of Science, Nantong University, Nantong Jiangsu
Email: 1067217478@qq.com

Received: Mar. 9th, 2016; accepted: Mar. 25th, 2016; published: Mar. 31st, 2016

Abstract

In this paper, we study a new class of distribution, called as generalized normal distribution. It can not only include the standard normal distribution, but also expand a new class of distribution which can describe the data with the features such as asymmetry, multimodal, heavy-tail and has good flexibility. We firstly investigate the generalized normal distribution by the images of distribution and density functions, and study its properties. Then we apply the new class of distribution to the new housing price index data, and the usefulness of generalized normal distribution can be examined by the real data analysis.

Keywords

Generalized Normal Distribution, Density Function, New Housing Price Index

广义正态分布族及其应用

刘晓庆，赵为华

南通大学理学院，江苏 南通
Email: 1067217478@qq.com

收稿日期：2016年3月9日；录用日期：2016年3月25日；发布日期：2016年3月31日

摘 要

本文研究了一类新的分布函数族，称之为广义正态分布族。广义正态分布族不仅包含了原有的正态分布，还拓展出了一族新的分布，且能刻画具有非对称、多峰、厚尾等特征的数据，具有很好的灵活性。本文首

文章引用: 刘晓庆, 赵为华. 广义正态分布族及其应用[J]. 统计学与应用, 2016, 5(1): 58-63.
http://dx.doi.org/10.12677/sa.2016.51007

先研究了广义正态分布族的分布函数以及密度函数图像，通过观察图像形状的变化，并研究广义正态分布族的性质。然后将这一新分布应用到房价指数数据分析中，通过分析说明了广义正态分布族的有用性。

关键词

广义正态分布，密度函数，新房价指数

1. 引言

正态分布是最常用的分布，也是概率统计中研究得最多的分布，在许多实际问题中有大量的应用。当实际数据服从或近似服从正态分布，基于正态分布提出的统计方法及其统计量具有很多良好的性质，如无偏性、一致性、有效性等。但是，当实际数据的分布偏离正态分布或服从非正态分布时，仍按照传统方法假定其服从正态分布，就会产生诸多问题，如统计推断精度显著降低，甚至产生错误结论，所提的统计方法容易受个别数据的影响，不具有“稳健性”。

为克服正态分布假定的弱点，许多研究者开始研究一些非正态分布，如 T 分布、对数正态分布、拉普拉斯分布等，并研究他们各自的性质及其在实际问题中的应用。本文在正态分布的基础上定义出一类广义正态分布族，该分布具有更好的灵活性，能拟合尖峰、厚尾、非对称数据。

2. 正态分布

正态分布又通常被称为高斯分布，是一个在数学、物理及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。

正态分布的密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{(x-\mu)^2}{2\sigma^2}}$；

其标准化后的概率密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\frac{x^2}{2}}$。

通过图 1 可以发现，正态分布曲线为一钟型曲线，形状优雅，形态对称，具有轻尾性[1]。

正态分布具有稳定性，使得其被广泛的使用。但是对于很多杂乱无章的数据，运用正态分布进行数据分析检验时，得到的结论往往偏差较大，与实际情况不符。在这种情况下，我们就需要一种新的分布对这些数据进行拟合和分析。

3. 广义正态分布族

本文研究一类新的分布函数族，我们称之为广义正态分布族，其分布函数定义如下：

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\ln\left(\frac{G(x;\theta)}{1-G(x;\theta)}\right)}\mathrm{e}^{\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t$$

其中 $G(x;\theta)$ 是任一随机变量的分布函数。

其概率密度函数为：

$$f(x;\mu,\sigma^2,\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(\ln\frac{G(x;\theta)}{1-G(x;\theta)}-\mu\right)^2}{2\sigma^2}\right\}\cdot\frac{g(x;\theta)}{G(x;\theta)\cdot(1-G(x;\theta))}$$

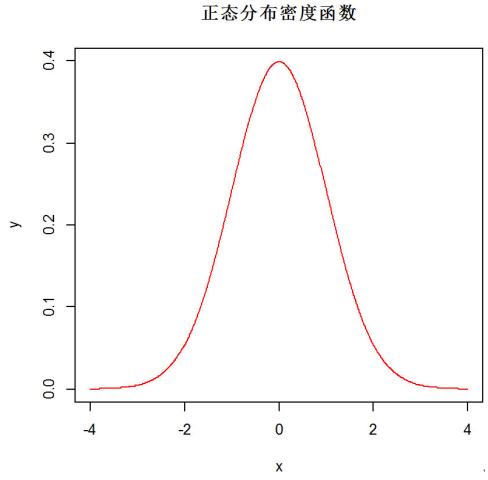

Figure 1. Curve: normal distribution
图 1. 正态分布曲线

当 $G(x;\theta)=\dfrac{e^{x}}{e^{x}+1}$ 时，$F(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$，即为正态分布函数。而当 $G(x;\theta)$ 变化时，$F(x)$ 也在不断的变化，即拓展出很多新的分布。

因此该广义正态分布族函数不仅包含了原有的正态分布，同时拓展出了一族新的分布。通过选取不同的分布 $G(x;\theta)$，广义正态分布族可以刻画很多更一般的数据，具有很好的灵活性。

为了对广义正态分布族有更深刻的了解，我们选取了不同的 $G(x;\theta)$，画出其分布函数及其密度函数图像，直观地去感受广义正态分布族的灵活性和可适用性。

图 2 中我们能展示了具有明显特征和研究意义的一些概率密度图像，如多峰、厚尾、偏态的多种情况。

由于定义中的广义正态分布函数中参数较多，对于随机给出的一组数据，我们首先需要估计出这些参数的值，得到确切的密度函数，才能对其进行数据分析。对此，我们采用极大似然估计方法对参数进行估计。

广义正态分布族中参数的极大似然估计[2]步骤如下：

1) 写出似然函数：

$$\begin{aligned}L&=L\left(\mu,\sigma^2,\theta;x_1,x_2,\cdots,x_n\right)\\&=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^2}\cdot\exp\left\{-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}\right\}\cdot\frac{e^{\sum_{i=1}^{n}y_i}}{\prod_{i=1}^{n}\left(1+e^{y_i}\right)^2};\end{aligned}$$

其中 $y_i=\log\dfrac{G(x_i)}{1-G(x_i)}$。

2) 求出参数 μ,σ^2,θ 的最大似然估计 $\hat{\mu},\hat{\sigma}^2,\hat{\theta}$，取其对数似然函数 $\ln L\left(\mu,\sigma^2,\theta\right)$：

$$\ln L\left(\mu,\sigma^2,\theta\right)=-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}+\sum_{i=1}^{n}y_i+\ln\left(2\pi\sigma^2\right)^{-\frac{n}{2}}+\ln\left(\frac{1}{\prod_{i=1}^{n}\left(1+e^{y_i}\right)^2}\right)$$

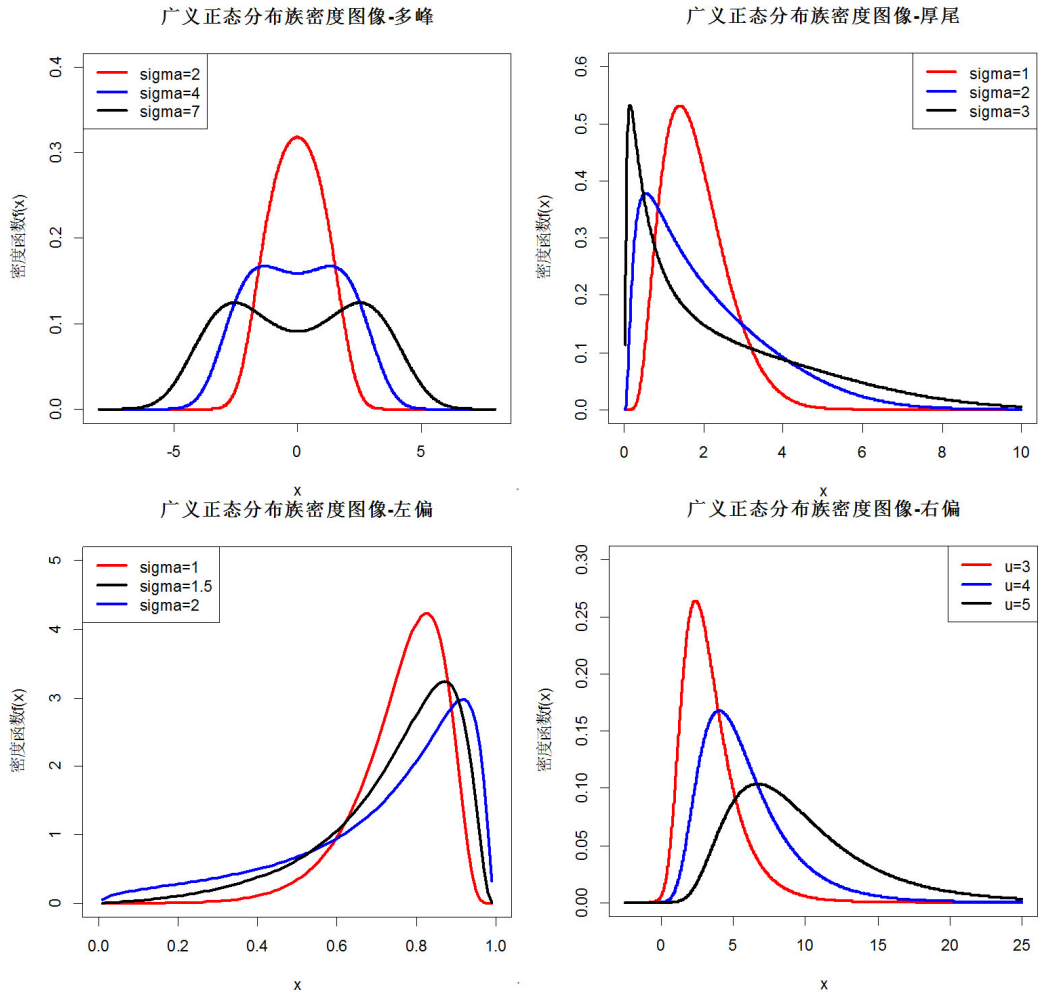

Figure 2. The density function image of the generalized normal distribution
图 2. 广义正态分布族的密度函数图像

3) 通过对数似然函数求偏导并令其等于 0，可得到如下似然方程：

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = -\dfrac{1}{\sigma^2} \cdot \sum_{i=1}^{n} \left(y_i - \mu \right) = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{1}{2\sigma^4} \cdot \sum_{i=1}^{n} \left(y_i - \mu \right)^2 - \dfrac{n}{2\sigma^2} = 0 \end{cases}$$

计算得出 μ 和 σ^2 的似然估计：

$$\begin{cases} \hat{\mu} = \dfrac{1}{n} \sum_{i=1}^{n} \log \dfrac{G\left(x_i \right)}{1 - G\left(x_i \right)} = \overline{y} \\ \hat{\sigma}^2 = \dfrac{1}{n} \sum_{i=1}^{n} \left(\log \dfrac{G\left(x_i \right)}{1 - G\left(x_i \right)} - \overline{y} \right)^2 \end{cases}$$

4) 基于上述估计值，我们再极大化似然函数得到 $\hat{\theta}$。由于 $G\left(x;\theta \right)$ 可以是任意的分布函数，在指定分布下，我们可以得到相应的表达式。

5) 迭代上述两步直至收敛。

4. 广义正态分布族的应用

本文选取了近年来人们关注较多的房地产市场中的房价指数数据。房价指数反映了一定时期内房屋销售价格变动程度和趋势的相对数，反映房价在不同时期的涨幅程度，其中二手房价格指数，就是反映一定时期内二手房价格水平变动情况的统计指标。通过价格指数，可以了解二手房价格的变动程度和综合观察二手房价格变动对房价总水平的影响。房价指数是决策者、投资者的重要参考资料之一，但是目前我国各类房地产指数在应用上存在着很大的发展空间。因此研究房价指数的变化趋势并对其进行预测有着其实际意义[3]。

我们对 2011 年 1 月 1 日至 2015 年 9 月 1 日北京地区的每月的二手住宅价格定基指数进行研究[4]。

首先我们对数据进行了标准化处理，以减少量纲的影响程度。接着我们使用 R 语言画出了二手住宅价格定基指数的概率密度图像(如图 3)，以确定房价指数的分布，并以此来作进一步分析[5]。显然，该指数的密度函数是多峰的情况。如果运用正态分布去进行拟合分析的话，具有较大的误差且与实际情况不符。

因此，我们尝试使用新定义的广义正态分布族进行拟合。在前期的研究中我们发现当 $G(x;\theta)\sim N(\mu,\sigma^2)$ 时，出现多峰的情况。因此我们选取 $f(x;\mu_2,\sigma_2^2,\theta)$，$G(x;\theta)\sim N(\mu_1,\sigma_1^2)$ 进行估计和拟合。

图 3 中：红色曲线为用广义正态分布族拟合出来的密度曲线图，蓝色曲线为标准正态分布拟合出来的密度曲线图。可以明显发现正态分布拟合出来的密度函数与事实情况严重不符，而广义正态分布族的拟合效果较好，具有良好的适用性，运用此拟合出来的概率密度函数进行后续的研究准确度也较高。

另外，通过查阅相关资料，我们也对该房价指数进行了一些研究。2011 年，中国房地产市场受政策环境的影响，二手房在成交方面和价格方面都出现了大幅度的下跌。而 2012 年房地产价格止跌反弹，增长迅速，尤其是一线城市的回暖力度很快，成交成为当时近三年内的最高。2013 年房价持续增长，尤其是二手房市场创新高。虽然 2014 年以来，我国国房景气指数逐月下滑，房地产投资增速显著放缓，70 个大中城市中房价下跌城市不断增加。但是房地产作为一热门市场，尤其是北京这个一线城市，其二手住宅一直走在涨幅榜的前列。2015 年北京楼市又创下新高，在新房供应紧缺的背景下，北京的二手房成交热度超过新房，在进入明显的二手房时代[6]。

因此 2011 年 1 月 1 日至 2015 年 9 月 1 日的二手房的房价指数呈现出两极分化的趋势，但是在政府

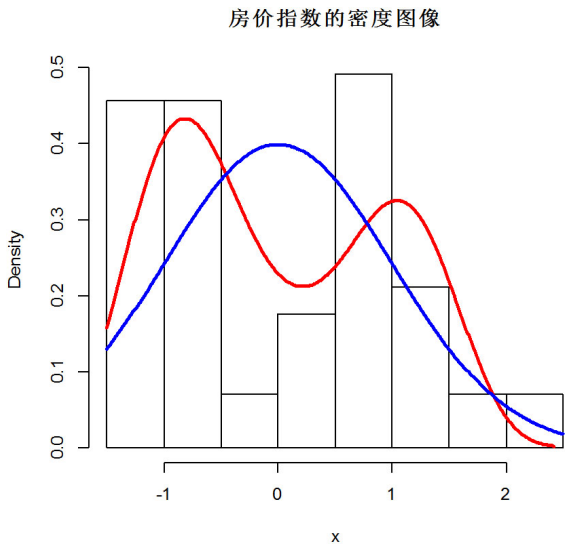

Figure 3. The density image of house price index
图 3. 房价指数的密度图像

不断的宏观调控和房地产市场本身的自我调整，我国的房价虽有涨有跌，却依旧集中在一定的范围之内。

5. 结语

本文在正态分布的基础上，引进广义正态分布族的概念。该广义正态分布族，可以避免正态分布的对称性、轻尾性、尖峰性等问题，在实际应用中具有很好的可适应性，并为研究其他分布提供了理论和方法。

致　谢

本论文是在指导老师赵为华副教授的细细指导下完成的。导师渊博的专业知识，严谨的治学态度，精益求精的工作作风，诲人不倦的高尚师德，严以律己、宽以待人的崇高风范，朴实无华、平易近人的人格魅力对我影响深远。不仅使我树立了远大的学术目标、掌握了基本的研究方法，还使我明白了许多待人接物与为人处事的道理。本论文从选题到完成，每一步都是在导师的指导新完成的，倾注了导师大量的心血。在此谨向导师表示崇高的敬意和忠心的感谢！

本论文的顺利完成，离不开各位老师、同学和朋友的关心和帮助。在此感谢赵为华老师的指导和帮助；感谢辅导员钱宗霞老师和班主任罗秀花老师的支持和鼓励；感谢项目小组成员葛雅雯和曹阳的努力和坚持，没有他们的帮助和支持是没有办法完成该论文。

由于本人理论水平比较有限，论文中的有些观点和阐述难免有疏漏和不足的地方，欢迎老师和专家们指正。

基金项目

本文受大学生创新实践项目“广义正态分布族的研究”(201510304051Y)资助。

参考文献 (References)

[1] Rickjin (靳志辉). 正态分布的前世今生[Z], 2012.

[2] 魏宗舒, 等. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2008.

[3] 百度百科. 房价指数[EB/OL]. http://baike.baidu.com/view/2307070.htm

[4] 新房价指数数据[EB/OL]. http://data.eastmoney.com/cjsj/newhouse.html

[5] Kabacoff, R.I. R 语言实战[M]. 高涛, 肖楠, 陈钢, 译. 北京: 人民邮电出版社, 2013.

[6] 中国指数研究院. 中国房地产市场形势总结与展望[R]. 中国指数研究院, 2015.