Statistics and Application 统计学与应用, 2021, 10(2), 256-268
Published Online April 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.102025

Hans汉斯

基于大数据背景下的正态分布研究

——以浙江省居民人均收入分布为例

张凌邦，徐秀丽

燕山大学理学院，河北 秦皇岛
Email: 1961522539@qq.com

收稿日期：2021年3月28日；录用日期：2021年4月12日；发布日期：2021年4月25日

摘　要

运用大数据推动经济转型升级、完善社会治理、提升政府服务和管理能力已成为当今社会发展的必然趋势。如何分析、挖掘、解读数据，从而让数据说真话，一直是学术界、政界和产业界关注的重要问题。众多经济问题中，居民人均收入分布一直是衡量一个地区发展的重要指标之一。本文通过对浙江省居民人均收入分布研究为例，同时通过使用MATLAB，参考多种分布模型拟合收入分布，对不同分布模型的适用程度进行研究，得到了在不同情况下不同分布的适用范围和拟合优度。同时通过对浙江省居民人均收入分布的相关数据的解读，为政府相关政策的制定提出合理的建议，为较为深入及复杂的人均收入分布问题提供了一些必要的研究基础。

关键词

大数据时代，分布模型，居民人均收入，MATLAB

Study on Normal Distribution in Large Data Background

—A Case Study of per Capita Income Distribution in Zhejiang Province

Lingbang Zhang, Xiuli Xu

School of Science, Yanshan University, Qinhuangdao Hebei
Email: 1961522539@qq.com

Received: Mar. 28th, 2021; accepted: Apr. 12th, 2021; published: Apr. 25th, 2021

文章引用: 张凌邦, 徐秀丽. 基于大数据背景下的正态分布研究[J]. 统计学与应用, 2021, 10(2): 256-268.
DOI: 10.12677/sa.2021.102025

Abstract

It has become an inevitable trend of social development to use big data to promote economic transformation and upgrading, improve social governance and improve government service and management ability. How to analyze, excavate and interpret data so as to tell the truth is always an important issue in academia, politics and industry. In many economic problems, the distribution of per capita income has always been one of the important indicators to measure the development of a region. This article through the study of residents' per capita income distribution in Zhejiang province as an example, through the use of MATLAB, refer to a variety of distribution model fitting of income distribution, studied the applicability of different distribution models. The applicability and goodness of fit of different distributions in different cases are obtained. According to Zhejiang province to study the distribution of the per capita income, for further on the more complex economic value per capita income distribution problem research foundation, and also puts forward suggestions on related to study the distribution of the income.

Keywords

Age of Large Data, Distribution Model, Per Capita Income, MATLAB

1. 引言

1.1. 研究背景及意义

在大数据迅猛发展的当代，许多反映经济问题及经济现象的模型也同样随之不断完善与改进。在基于传统正态分布研究的基础上，更多的是根据数据特点，重点解决存在于其中的两个问题。第一个问题是如何拟合关于某一经济现象的分布函数，第二个问题是对于不同方法进行参数估计的精确度孰好孰坏。对于此研究背景提出了关于人均收入分布函数拟合优度与方法的研究。从而进一步得出拥有不同经济发展水平的地区更适合以何种模型拟合分布函数的结论。如今大数据发展迅速，传统使用参数估计与传统计算的方法显然已经无法满足于数据繁琐复杂的今天。对于使用 MATLAB，R 语言等数学与统计软件对居民人均收入的数据进行拟合并对拟合分布进行研究就显得尤为重要。

本文也通过运用当下使用较为广泛的 MATLAB 软件以及 MATLAB 中的统计工具箱来进行统计与数学分析，一方面更好的阐述问题并进行研究，另一方面更是适应大数据发展的背景与研究趋势。对于浙江省居民人均收入分布的研究，不仅可以更清楚的认识到地方经济结构和收入结构，在拟合与研究的过程中，对于研究数据选择、数据拟合、代码操作分布函数选择等理论问题都具有一定的推动的作用。其次，在大数据的背景下，能够更合理的拟合居民人均收入分布，将不同分布类型与经济现实相结合并逐步进行研究，可以在地区经济问题分析上起到重要的辅助作用。

1.2. 研究思路及研究方法

本文旨在根据浙江省统计年鉴 2013~2016 年浙江省居民人均收入以及浙江省居民人均收入情况等数据，运用常见的五种适用于对于人均收入分布的分布函数对数据进行拟合。它们分别是对数正态分布、

Γ分布、威布尔分布、SM分布和B2分布。然后，得到不同的浙江省居民人均收入分布模型，并对分布的拟合优度与分布中参数估计的精确度进行比较，结合各分布的一系列经济指标对各个分布的拟合优度和分布的代表情况进行评估。从而比较出各个分布模型的优势与劣势，使我们得到最优的拟合居民人均收入分布的方法与分布模型，并可以在全国居民人均收入分布的研究做进一步的推广。

基于对浙江省居民人均收入分布研究，所研究数据包含三个类别：即浙江省2013~2016年城镇居民人均收入、浙江省2013~2016年农村居民人均收入、浙江省2013~2016年全体居民人均收入。基于大数据的背景，所选择研究的数据来源于中国统计年鉴和浙江省统计年鉴。通过使用MATLAB进行操作，在选定拟合数据的函数分布类型时，现有研究证明拟合分布函数的分布类型大致包含六种，选取双参数，三参数，和多参数分布模型。拟合后的分布函数通过A-D检验和K-S检验。比较不同分布模型的拟合优度，最后得出结论。通过横向纵向比较，比较不同经济发展水平地区的收入分布模型适用情况。从而推广至全国居民人均收入分布的拟合模型。

2. 收入分布模型的选择

2.1. 研究思路及研究方法

在拟合居民人均收入分布时，一般存在着两个问题，第一个是如何选择合适的分布函数，第二是如何进行参数的估计。其中较为重要的就是分布函数的选择。使用城市和乡村的收入数据推断该地区居民总体收入状况依赖于收入分布函数的选择。收入分布函数的研究由来已久，一个多世纪前，帕累托提出了帕累托分布(Pareto Distribution)。吉布拉指出收入分布可以很好地由对数正态分布(Log-normal Distribution)拟合。不过随后的研究表明该分布会低估高收入阶层的收入状况[1]。在国内，也有许多专家对收入分布函数有所研究。其中，部分学者运用伽马分布拟合了中国农村居民的收入分布。同时，通过提出运用帕累托分布、正态分布和指数分布构成混合分布对居民人均收入进行估计[2] [3]。随着非参数估计的估计方法近几年开始有所发展，许多学者也提出了更多不同类型分布对数据的拟合方法，并采用威布尔分布、对数正态分布以及第2类beta分布(B2)拟合了中国城乡居民的收入[4]。

通过已经进行过的研究不难发现，在比较不同分布函数和参数估计方法的拟合效果之后，如果分析复杂类型的分布，由于参数有限，参数估计调节形态的能力会明显劣于非参数方法。需要特别指出的是，在分析收入分布这类分布曲线为单峰而且右偏，同时分布曲线较为光滑的分布时，使用非参数方法从而产生过多的参数反而会产生大量冗余信息同时会形成“噪声”，进而影响最终的拟合效果，因此，在本科学习阶段，使用参数估计要明显优于非参数估计。鉴于前人的研究以及相关文献，我们选择了五种分布分别对浙江省居民人均收入数据进行拟合，包括：对数正态分布、Γ分布、威布尔分布、SM分布和B2分布。

2.2. 分布模型概述

2.2.1. 对数正态分布

对数正态分布(Logarithmic-normal Distribution)若一个随机变量X的对数服从于正态分布，那么就称X服从于对数正态分布。对数正态分布从短期来看，与正态分布非常接近。但长期来看，对数正态分布向上分布的数值更多一些。

设X是取值为正数的连续随机变量，若$\ln X \sim N\left(\mu,\sigma^2\right)$，则$X$的概率密度为：

$$f\left(x,\mu,\sigma\right)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\dfrac{1}{2\sigma^2}\left(\ln x-\mu\right)^2\right], & x>0\\ 0, & x\le 0\end{cases}$$

则称随机变量 X 服从对数正态分布，记为 $X \sim N\left(\mu,\sigma^2\right)$

设 X 服从对数正态分布，其密度函数为：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma x}\mathrm{e}^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

2.2.2. 伽马分布

伽马分布(Gamma Distribution)是概率统计中的一种重要的分布类型，同时也是概率论中特别强调的一种分布。“指数分布”和“卡方分布”都是伽马分布的特例。该分布中的参数 α 称为形状参数(Shape Parameter)，β 称为尺度参数(Scale Parameter)。

假设随机变量 X 为等到第 α 件事发生所需之等候时间，密度函数为：

$$f\left(x,\beta,\alpha\right)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x},x>0$$

2.2.3. 威布尔分布

威布尔分布是可靠性分析和寿命检验的理论基础。特别是适用于某一类产品经磨损损耗后累计失效的分布形式。因其在对分布中的未知参数进行推断时较为简单，所以在许多领域都可以看到该分布的使用。研究表明威布尔分布也可以用来拟合收入数据。其概率密度函数为：

$$f\left(x;\lambda,k\right)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{k}}, & x\geq 0\\ 0, & x<0\end{cases}$$

2.2.4. 多参数分布模型

上述介绍的三个分布模型均为双参数分布。在众多分布模型的实证研究之中，多参数分布模型被越来越多的人用于拟合数据。此处引用两个分布进行分析，分别是 SM 分布和 B2 分布。引入多参数模型对研究居民收入分布同样具有重要作用。双参数分布中，两个参数分别为形状参数和尺度参数。早期研究阶段，帕累托认为，收入分布完全可以由帕累托分布进行拟合[5]。但经研究发现，帕累托分布不能很好地拟合全体居民人的收入分布而更侧重于拟合高收入人群的收入分布[6]。与此同时，多参数分布同样被应用于拟合居民收入分布。调查表明，针对某地区的居民人均收入分布研究进行收入数据拟合时，四参数模型拟合效果往往要优于三参数模型，三参数模型要优于双参数模型的拟合效果，所以我选择了多参数模型对居民人均收入分布进行拟合。但由于多参数模型对于参数估计的难度有了显著性的提高，所以我选择了较为简单的两种多参数分布模型进行拟合，保证研究的可行性。同时也降低了拟合过程中对参数估计的难度要求，使研究在可操作的前提下，使用 MATLAB 能够实现最后的结果。同时，参数过多不仅拟合难度大大增加，还会造成过多的冗余信息，影响最终拟合结果。

3. 参数估计

3.1. 参数估计方法概述

选择好合适的收入分布模型以后，是否能成功对数据进行拟合得到最终的拟合结果，使用何种参数估计方法得到分布模型中的相关参数就显得更为重要。常用的参数估计方法包括点估计法、极大似然估计方法、区间估计方法、最小二乘法、一致最小方差无偏估计法等。每一种参数估计方法都具有各自的特点，具体选择哪一种参数估计方法要考虑数据类型与数据量的大小以及区间估计与点估计的精确度，

同时也要考虑无偏性、有效性、一致性这些统计特性。

3.2. 参数估计方法的选择

矩估计方法是参数估计中较为常见的。矩估计同时也称为点估计，点估计由于其计算简单，估计结果直观的特点被普遍运用。对于任意的随机变量而言，矩是含义较为清晰，应用较为广泛的数字特征，总体的各阶矩同总体分布中的总体参数有关，有的甚至就等于总体参数。常见的矩包括：一阶样本原点矩、二阶样本中心矩等。由大数定律可知，随机抽样所获取的样本中计算出的样本原点矩可以依概率收敛至相对应的总体原点矩。使我们得到了新的参数估计思路，可以用样本矩来近似总体矩，当样本量逐渐增大的过程中，二者的实际观测值就越来越接近，这种寻找总体中未知的参数的方法就被称为矩估计法，称之为矩估计。极大似然估计是点估计中所包含的另一种参数估计方法，在点估计基础上，极大似然估计要依赖于一个新的概念：似然函数。其核心思想是在参数估计过程中，根据似然函数与对数似然函数得到最终的参数估计值。同时，极大似然估计具有不变性,在计算一些分布包含多个参数时，与其他参数估计方法相比会便捷许多[7]。

最终，结合研究问题选择了极大似然估计对分布中的参数进行估计。首先，基于已学知识，对于所研究的年度数据等时间序列数据，使用极大似然估计是较好的选择。基于对参数估计的相关研究，极大似然估计的方法的估计效果较好。其次，虽然最小二乘估计、置信区间估计等方法在参数估计时都是很好的选择，但其中大部分估计方法依赖于总体分布。反观所研究的人均收入分布问题，具体收入分布状况未知；将数据用五种分布类型进行拟合，研究所选取的数据有限，则参数估计过程中重复次数也有限，进行区间估计的精度就会有所下降。因此，选择使用极大似然估计方法进行参数估计。

3.3. 极大似然估计方法的实现

极大似然估计方法的思想是：得到样本值 $x_1, x_2, x_3, x_4, \cdots, x_n$ 时，选取使似然函数 $L(\theta)$ 取得最大值的 $\hat{\theta}$ 作为未知参数 θ 的估计值。由于在选择对数据进行拟合的五种分布均为连续型分布，则此处只写出连续型分布进行极大似然估计的理论方法。

以对数正态分布的参数估计为例，此分布中的待估参数为总体均值 μ 和总体方差 σ^2,参数估计的计算过程为：

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

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

$$\text{令}\begin{cases}\dfrac{\partial\ln L\left(\mu,\sigma^2\right)}{\partial\mu}=\dfrac{\sum_{i=1}^{n}\left(\ln x_i-\mu\right)}{\sigma^2}=0\\[2ex]\dfrac{\partial\ln L\left(\mu,\sigma^2\right)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}-\dfrac{\sum_{i=1}^{n}\left(\ln x_i-\mu\right)^2}{2\sigma^4}=0\end{cases}$$

根据计算过程，可以解出总体均值 μ 和总体方差 σ^2 两个参数的极大似然估计值，它们分别为 $\mu=\frac{1}{n}\sum_{i=1}^{n}\ln x_i$ 、 $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}\left(\ln x_i-\mu\right)^2$ 。在 MATLAB 中，可以根据已经分组的数据，计算出其似然函数，然后使用迭代算法，计算出似然函数值达到最大时其所对应的参数值，使用极大似然估计的方法得到所要估计的参数。

4. 拟合数据选择

4.1. 拟合数据来源

使用 MATLAB 进行数据拟合时，最为重要的就是选取所需要拟合的数据。在进行研究设计之初，通过查阅一部分关于居民人均收入的相关文献。阅读相关文献的同时查阅了各个网站与数据库的关于浙江居民人均收入的数据。在基于大数据的背景与前提下，查找到关于浙江省居民人均收入数据大致分为以下几个类型[8]。第一个类型就是关于浙江省居民人均收入的总体数据，类似于表 1。

Table 1. Per capita income of urban and rural residents in Zhejiang province, 2010~2016
表 1. 浙江省 2010~2016 年城镇及农村居民人均收入表

年份	2010	2011	2012	2013	2014	2015	2016
城镇收入/元	27,359	30,971	34,550	37,080	40,393	43,714	47,237
农村收入/元	11,303	13,071	14,552	17,494	19,373	21,125	22,866

对于上述的三个数据表，浙江省 2010~2016 年城镇居民人均收入和浙江省 2010~2016 年农村居民人均收入两个数据表来自于浙江省统计信息网，浙江省 2013~2016 全体居民人均收入的数据来自于中华人民共和国国家统计局网站。对于这一部分的数据，在最初的研究设计时，初步设计思路是将其分别拟合，根据拟合结果研究居民人均收入分布的相关问题。但是在使用 MATLAB 进行初步拟合的过程中发现了许多问题，例如出现无法进行拟合，拟合出的结果非正态等问题。综合考虑数据来源后，放弃了对于该类数据的拟合。

第二类数据就是各城市数据。对于年度的总体数据而言，由于数据点太少，不能将其进行分组，在利用分布进行数据拟合时，困难太大。考虑到这个问题之后，在原有研究思路的基础上，我考虑到使用各城市的年度数据，如表 2。预实验结果表明对该类数据再进行分类，更加适用于进行分布模型的拟合。

Table 2. Per capita income of all residents in 11 cities of Zhejiang province, 2013~2016
表 2. 浙江省 11 市 2013~2016 年全体居民人均收入表

年份	2016	2015	2014	2013
杭州市	46,116	42,642	39,237	35,763
宁波市	44,641	41,373	38,074	34,657
舟山市	41,564	38,254	35,330	32,027
嘉兴市	41,532	37,139	34,318	31,315
绍兴市	41,506	38,389	35,335	32,191
温州市	39,601	36,459	33,478	30,602
金华市	37,159	34,378	31,599	28,673
台州市	36,915	33,788	30,950	28,215
湖州市	33,966	34,251	31,510	28,717
丽水市	26,757	24,402	22,426	20,418
衢州市	26,745	24,460	22,436	20,342

根据浙江省 11 个城市关于城镇、农村以及全体居民的相关平均收入数据。在进行拟合时发现，由于数据点较少，在使用五种分布进行数据拟合时，很难拟合出预期的效果。即便是根据平均收入的多少对

数据进行分组，还是根据数据的其他特点，比如地区等因素分组，都不能拟合出满意的结果。这一点与上述的使用年度总体数据的结果十分相似，在使用 MATLAB 进行拟合时，更倾向于拟合出增长率函数，更倾向于说明人均收入在随年份的增加而逐渐增长之类的经济问题。但是对于研究居民人均收入分布的问题没有更深层次的实际研究意义，对所要研究的问题没有起到一定辅助作用。

第三类数据就是已有的分组数据[9]。这类分组数据虽然已进行分组，但是已有的分组数据不提供收入的上下组限，在拟合时，虽然可以呈现正态分布的形状，但对于实际研究的意义不大，相关分组数据来自于浙江省统计年鉴，如表 3 所示。

Table 3. Basic information of all households by income in Zhejiang province, 2013~2015
表 3. 浙江省 2013~2015 年按收入分全体居民家庭基本情况表

指标	2013 年	2014 年	2015 年
居民人均可支配收入	29,775	32,658	35,537
人均可支配收入中位数	25,210	28,580	31,499
按人均可支配收入多少分组			
最低 20%户	8773	9910	11,574
较低 20%户	17,545	20,330	22,730
中间 20%户	25,271	28,504	31,443
较高 20%户	35,737	39,567	43,085
最高 20%户	68,028	72,159	75,072
最高和最低收入组收入差距倍数	7.75	7.28	6.49

在对上述数据进行拟合时，多数情况下都不能够很好的拟合于研究所选定的五种分布类型。为使得分布拟合效果更优，通过检验的概率有所提高，我们需要尝试使用其他数据继续拟合分布。

4.2. 拟合数据的选择

在选择分组数据进行拟合试验时，由于数据量过少，拟合效果未能达到研究所预期的最终结果与目标。最终，我选择进行拟合的数据来自于浙江省 2015~2017 统计年鉴目录下，各市县国民经济主要经济指标中的浙江省市区县居民人均收入数据，如表 4 所示。

Table 4. Per capita disposable income of residents in Zhejiang province, 2014~2016 (partial data)
表 4. 浙江省 2014~2016 年居民人均可支配收入表(部分数据)

市县名称	2016 年城镇居民人均可支配收入(元)	2016 年农村居民人均可支配收入(元)	2015 年城镇居民人均可支配收入(元)	2015 年农村居民人均可支配收入(元)	2014 年城镇居民人均可支配收入(元)	2014 年农村居民人均可支配收入(元)
杭州市区	52,185	27,908	48,316	25,719	44,632	23,555
萧山区	55,712	31,849	51,490	29,354	47,195	26,758
余杭区	53,215	31,608	49,273	29,159	45,329	26,581
富阳区	47,339	27,236	43,510	25,010	39,954	22,840
临安市	**44,858**	**25,849**	**41230**	**23,736**	**37,860**	**21,578**
建德市	41,531	21,896	38,102	20,051	35,117	18,295
桐庐县	**42,496**	**24,619**	**39,348**	**22,504**	**36,366**	**20,627**

Continued

淳安县	36,708	16,110	33,432	14,632	30,559	13,278
宁波市区	**51,560**	**28,572**	**51,201**	**28,087**	**47,190**	**25,815**
鄞州区	55,151	30,343	50,215	29,110	46,324	26,682
奉化区	**45,371**	**25,885**	**45,359**	**26,500**	**41,921**	**24,312**
余姚市	48,831	28,589	47,182	27,295	43,526	25,041
慈溪市	**50,828**	**29,547**	**41,894**	**23,950**	**38,755**	**22,033**
象山县	46,836	26,113	43,565	24,228	40,189	22,146
宁海县	**47,702**	**26,233**	**44,324**	**24,319**	**40,664**	**22,209**
温州市区	47,785	22,985	44,026	21,235	40,510	19,394
洞头区	**37,827**	**19,907**	**34,674**	**18,365**	**43,208**	**21,682**
瑞安市	50,904	25,570	46,949	23,671	42,610	22,668
乐清市	**50,263**	**26,943**	**46,352**	**24,891**	**31,730**	**16,617**
永嘉县	38,246	18,391	35,161	16,938	32,330	15,404

在选取拟合数据时，对多组数据使用 R 进行过拟合试验。对于上述提到的第一类、第二类与第三类数据，使用五种分布进行拟合时的拟合效果较差，不能清晰直观地看出收入分布的形态与特征，同时也缺乏一定的研究意义。根据所研究的研究背景与研究目标，在大数据挖掘的要求下，数据点过少则无法体现出数据量在拟合分布时所起到的作用。同时，少量数据点更多倾向于展现年度增长率、定基发展速度、环比发展速度等经济指标，而对于数据拟合，无论分组还是未分组数据，数据量达到一定的要求是最终能够得到较为合理和优质结果的前提[10]。选择浙江省 2014~2016 年城乡居民人均可支配收入的数据在 R 中进行拟合试验时，预试验输出结果如图 1。针对于对数正态分布，该组数据的拟合效果，频率分布与理论分布的差异相较于其他组数据都体现出了更为理想的拟合效果，同时也具备了一定的研究意义。所以，最终选取该组数据作为研究对象进行拟合与研究。

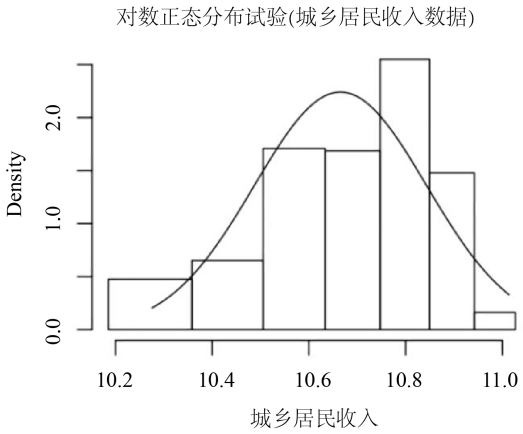

Figure 1. Using R to fit the experimental effect
图 1. 使用 R 语言进行拟合试验效果图

5. 数据拟合及相关检验过程

5.1. 实验数据初步拟合

使用 MATLAB 对数据进行拟合[11]，以对数正态分布的拟合过程为例，数据拟合后的曲线结果可以

初步看出，收入数据是否初步符合所将要进行拟合的分布模型。运行程序后得到的数据拟合对数正态分布曲线如图 2。初步拟合结果显示，居民人均收入分布数据与对数正态分布的拟合曲线具有一定的相似程度。

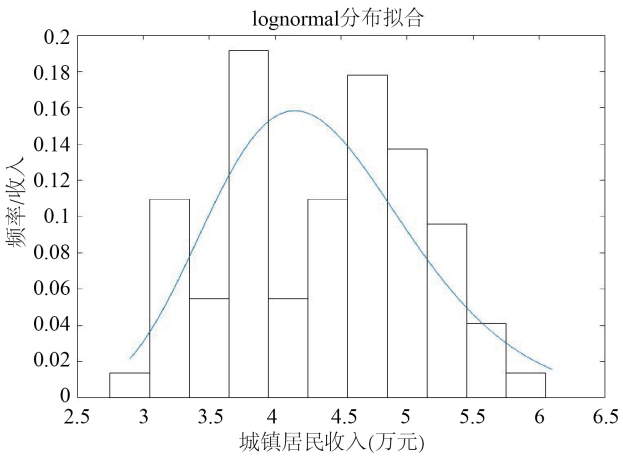

Figure 2. Fitting effect of lognormal distribution
图 2. 对数正态分布曲线拟合效果图

5.2. 分布拟合效果相关检验

5.2.1. K-S 检验及代码实现

K-S (Kolmogorov-Smirnov)检验是通过比较频率分布 $f(x)$ 与理论分布 $g(x)$ 分布之间差别的检验方法。其原理是当原假设 H_0 为两个数据分布一致或者数据符合理论分布时，$D=\max\left|f(x)-g(x)\right|$，当实际观测值 $D>D(n,\alpha)$，则拒绝 H_0，否则接受原假设。特别强调的是 K-S 检验具有分布无关的性质，在总体分布未知的情况下，K-S 检验的应用优势就显露无遗。K-S 检验在 MATLAB 中的程序代码如图 3 所示。

```
p=logncdf(x,para.mu,para.sigma);
[H1,p1]=kstest(x,[x,p],0.05)
if H1==0
disp('K-S检验该数据源服从对数正态分布')
else
disp('K-S检验该数据源不服从对数正态分布')
end
```

Figure 3. K-S inspection procedure chart
图 3. K-S 检验程序图

5.2.2. A-D 检验及代码实现

A-D (Anderson-Darling)检验是另一种对拟合优度进行检验的检验方法，同样的，其检验值越小则表示最终得到的数据拟合效果更为优秀，A-D 检验是在 K-S 检验的基础上进一步发展所得到的，相比 K-S 检验，A-D 检验的灵敏度更高。因为 K-S 检验对不同的分布是在各个分布无关的状态下进行检验，不同的分布并不影响其临界值的计算和最终的检验结果。相比之下，A-D 检验在计算临界值时更依赖于某一

特定分布，这使得 A-D 检验具有更灵敏且更适应结构复杂变化的数据结构的优势。考虑到最终拟合效果的准确检验，所以同时使用这两种检验方法并将其检验结果作为数据拟合效果的分析依据。A-D 检验在 MATLAB 中的程序代码如图 4 所示。

```
dist=makedist('lognormal','mu',para.mu,'sigma',para.sigma);
[H2,p2]=adtest(x,'Distribution',dist)
if H2==0
disp('A-D检验该数据源服从对数正态分布')
else
disp('A-D检验该数据源不服从对数正态分布')
end
```

Figure 4. A-D inspection procedure chart
图 4. A-D 检验程序图

6. 关于收入分布的实证分析

6.1. 浙江省居民人均收入分布的拟合结果

根据所选取的数据，使用 MATLAB 利用常见的五种分布类型对数据进行拟合，得到五组收入分布曲线。这五组分布同样都基于极大似然估计的参数估计方法。在使用同样的参数估计方法的前提下，五个分布对数据的拟合优度有不同的结果与表现。这五种分布的拟合效果如图 5。

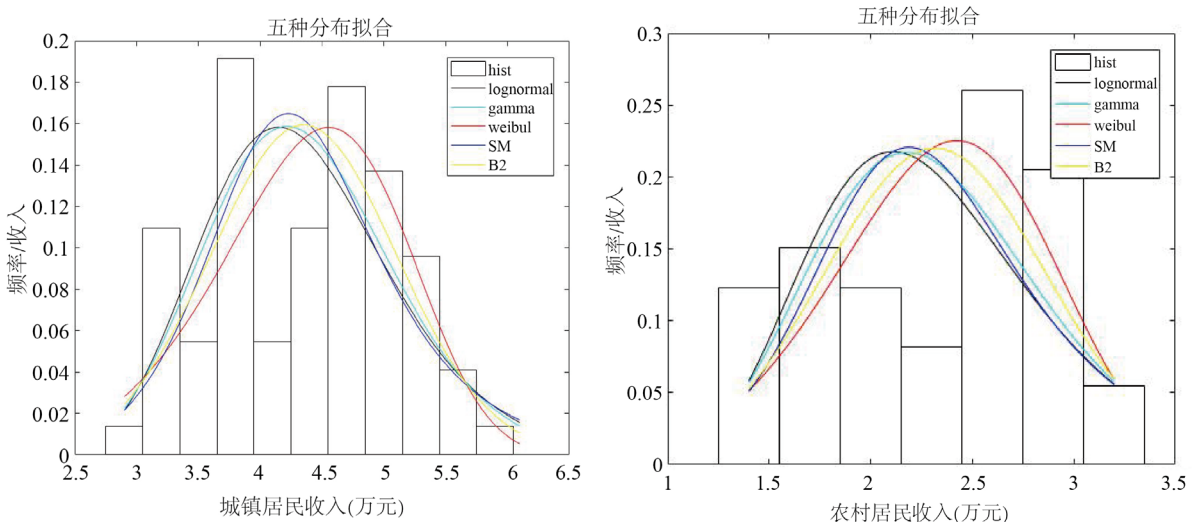

Figure 5. Comparison of fitting effects of five distribution
图 5. 五种分布曲线拟合效果对比图

对于所研究的数据，全都服从于这五种分布，数据拟合结果全部通过了 K-S 检验与 A-D 检验。基于对浙江省居民人均收入的研究，具体观察五种分布的 K-S 检验与 A-D 检验结果如表 5。

比较五种分布的检验值，城镇居民人均收入数据均可服从于这五种分布。但是，基于相同的 MLE 参数估计方法，威布尔分布的 A-D 检验值过大，数据拟合的效果最不显著。通过比较也发现作为多参数分布的 SM 分布对数据拟合的效果最好。其中，对数正态分布与伽马分布的拟合效果类似，SM 分布与 B2

分布的拟合效果类似。对于农村居民人均收入，SM 分布与威布尔分布的拟合效果更为显著，对数正态分布与伽马分布由于 A-D 检验值过大，分布模型与数据间的拟合效果较差。

Table 5. K-S and A-D test values of five distribution
表 5. 五种分布的 K-S 与 A-D 检验值

	城镇居民人均收入分布		农村居民人均收入分布	
分布类型	K-S 检验	A-D 检验	K-S 检验	A-D 检验
对数正态分布	0.3243	0.3154	0.4627	0.4633
伽马分布	0.3875	0.3623	0.3666	0.4143
威布尔分布	0.1983	0.6170	0.1275	0.2122
SM 分布	0.2526	0.2753	0.1189	0.1467
B2 分布	0.2622	0.2831	0.3495	0.1775

比较城镇与农村的居民人均收入分布拟合情况可以得出，多参数 SM 分布模型的拟合性能都拥有较好的表现。SM 分布作为拟合效果较优的模型针对不同类型的居民人均收入数据都拥有较强的拟合能力。根据检验值结果也可以看出，城镇与农村居民人均收入数据对于不同的分布模型，表现结果也有所不同。传统对数正态与伽马分布对于城镇数据的拟合效果优于其对于农村数据的拟合效果。威布尔分布对于农村数据的拟合表现也优于其对城镇收入的表现。

针对分布模型对于城镇及农村数据拟合效果不同的问题，观察图 5 及相关数据可以发现，城镇数据“分级”现象较为明显。相比与农村，城镇数据的峰值与极大极小值之间的差距较大，收入结构区别更为明显。基于当今社会农业经济的发展，农村居民的收入峰值与极值相差小，农业经济也体现出其较为简单的收入结构。所以，考虑现实经济意义，针对不同的收入结构选择不同的分布模型进行研究成为解决农村发展的关键问题之一。

6.2. 分布拟合结果分析及相关建议

通过比较表明，多参数模型的拟合效果要优于双参数模型与单参数模型。从分布的拟合过程来分析，则认为对分布拟合效果影响较大的有以下几个方面：

首先，数据量的大小对分布的拟合效果有较大的影响。进行数据的拟合之前，我设想了数据基于分布的最终拟合效果，最终的结果符合我当初对研究所进行的假设与研究方向设定。对于少量数据来说，威布尔分布的拟合优度要好于其他分布，但通过增加数据量，威布尔的拟合效果则大大下降。通过八个数据点再到上百个数据点，威布尔分布不再适用于对大量收入数据分布的拟合。对数正态分布以及伽马分布在增加数据量的同时并未出现拟合效果的明显变化。对于数据量的增大，多参数模型的拟合效果显著增加，以 SM 分布为例，增大数据量的检验结果如表 6。

Table 6. A-D test value of SM distribution with increasing data volume
表 6. 增大数据量时 SM 分布的 A-D 检验值

数据个数	A-D 检验值
8	0.5653
15	0.4281
432	0.2753

A-D 检验值的逐渐下降表明分布的拟合效果逐渐增强，可以说多参数分布模型在今后拟合全国或地方居民人均收入分布时，将发挥更为重要的作用。同时，许多参考文献以及学者的研究中也表明，多参数模型相比较于单参数模型和双参数模型具有更多的优点，拟合效果也更优更明显。

其次，对于数据拟合效果来说，数据分组也对其有较大的影响。在对数据逐步进行拟合的过程中，分组方法的不同使得数据拟合效果的不同也引起了我的注意。数据分组分的越密集，分布的正态拟合效果越不明显。根据实验的数据拟合结果来看，以城镇居民收入为例，对于以人口收入层次的 100%、50%、25%、20%为分界线进行划分的四组数据集，拟合 K-S 检验值随着分组指标间距的减小而增大，以数据集合 20%为分界线进行划分的数据集拟合效果最优，细化数据集的指标缩小指标间距有利于提高数据的拟合效果与分布的拟合性能。数据分组越分散，K-S 检验值逐渐减小，频率分布与理论分布更为接近。所以，在追求分布拟合效果接近于正态分布的同时，减小数据分组的组距对于浙江省的居民人均收入分布拟合更为理想。

再次，对于居民人均收入分布拟合效果产生影响的就是参数估计的方法。虽然我选择了使用极大似然估计的参数估计方法，但是同时还有许多其他可能估计精度更为优秀的估计方法，碍于本人在本科阶段所学的内容，此处不再展开，但是居民人均收入分布的参数估计是一个值得更加深入思考的问题所在。

最后，结合浙江省居民人均收入的情况来分析。以 2016 年为例，从城镇来看，人均可支配收入比 2015 年增长了 3523 元，同比增长 8.1%，农村常住居民人均可支配收入同比增长 8.2%，人均收入结构发生了较为深刻的变化。农村与城镇差距逐渐缩小，那么在对收入数据进行拟合时，选取威布尔分布这个分布模型时，就必然要考虑其对于高收入人群的拟合效果较好，而对一般收入居民人均收入和较低居民人均收入的拟合效果较为一般这个特性。在考虑选取基本的人均收入分布模型时，考虑收入人群变化和收入结构的改变，也是进行研究的重要落脚点。从浙江省居民人均收入的分布研究可以看出，收入结构的变化在全国范围内也具有同样的特点。贫富差距逐渐缩小，大数据背景下增加数据点就可以继续拟合全国居民人均收入分布。由于数据源与数据量较为匮乏，对于居民人均收入分布的研究依然具有重要的现实意义。同时，针对具有较大贫富差距的地区，在对城镇和乡村居民人均收入分别拟合之后，可以根据拟合结果进行模型的进一步选择和细化，为特定的社会经济问题研究提供案例支持与数据分析。

基于所得到的研究结果，提出以下几点建议：

1) 当更完善拟合效果更好的分布函数出现之前，没有任何一个分布的模型能完全拟合所有数据并保证其最终的拟合效果为最优。所以，在拟合人均收入分布时，应充分考虑研究的收入群体、收入类型以及其他一些重要的经济指标。根据不同的收入结构考虑使用不同的分布函数进行拟合。例如在研究经济发展较为发达的地区的高收入人群收入分布时，可以考虑使用威布尔分布。但当所研究地区的收入结构具有集中趋势时，对数正态分布与伽马分布进行拟合的效果就要优于威布尔分布，同时也应考虑多参数分布模型，针对不同的经济情况与研究目的，选取不同的分布函数。

2) 通过对居民人均收入分布的研究，我们可以得到，在研究不同的问题时对于不同的数据选择应有所侧重。特别是，如果有更完善的数据能够应用于人均收入分布的研究，那么对收入分布模型的细化分类与实际应用将起到极大地辅助作用。例如，更新包括根据人口在收入群体中的比重进行分组的数据、平均收入类型的分组数据、收入包含上限与下限的分组数据以及根据收入的来源不同进行分组的分组数据。如果这些类型的分组数据能够进一步完善，进一步扩充，那么就会为人均收入分布的研究上提供更多研究方向，提供更多的研究角度，同时提高研究的可行性。

3) 针对政府制定发展政策及规划而言，在进行城镇及农村居民人均收入提升的研究时，应当着力于改变现有的居民收入结构。居民收入结构相比于具体的收入数据来说，有着统揽大局的重要作用。对于城镇居民收入结构来说，应该在细化结构的同时对结构指标的划分标准进行统一。城镇经济随着社会发

展呈现出越来越复杂的结构变化，收入结构变化的加大改变了收入数据的系统性。在进行分布研究时可以发现，城镇数据由于其过于细分，而导致收入分布失去了其对于收入状况的代表性。对于农村居民人均收入结构而言，因其结构过于简单，也导致了收入分布失去了对于具体发展指标的指导意义。所以，政府在对居民人均收入进行相关政策制定时，应该充分考虑对城镇及居民收入的经济结构进行区分，针对不同结构划分不同相应统计指标，对统计指标集合进行一定程度的标准统一及细化。针对具体发展问题，增强经济数据在发展过程中的代表性。

参考文献

[1] 陈建东, 罗涛, 赵艾凤. 收入分布函数在收入不平等研究领域的应用[J]. 统计研究, 2013(9): 79-86.

[2] 王海港. 我国居民收入分配的格局: 帕累托分布法[J]. 南方经济, 2006(5): 73-82.

[3] 张萌旭, 陈建东, 蒲明. 城镇居民收入分布函数的研究[J]. 数量经济技术经济研究, 2013, 30(4): 57-71.

[4] 孙伟, 王琳. 居民收入分布函数的实证研究[J]. 数学的实践与认识, 2017(8): 315-320.

[5] 李实, 史泰丽, 别雍·古斯塔夫森. 中国居民收入分配研究[M]. 北京: 北京师范大学出版社, 2008.

[6] 王亚峰. 中国 1985-2009 年城乡居民收入分布的估计[J]. 数量经济技术经济研究, 2012(6): 61-73.

[7] 成平. 极大似然估计与似然比检验的几点注记[J]. 应用概率统计, 2003(1): 23-25.

[8] 陈建东, 程树磊, 蒲明. 如何准确地拟合居民人均收入分布[J]. 北京工商大学学报(社会科学版), 2017, 32(2): 10-20.

[9] 陈建东, 戴岱, 冯瑛. 居民收入样本分组数与基尼系数测算的关系探讨[J]. 统计与决策, 2011(15): 34-37.

[10] 黄恒君, 刘黎明. 一种收入分布函数序列的拟合方法及扩展应用[J]. 统计与信息论坛, 2011(12): 14-18.

[11] 薛定宇. 高等应用数学问题的 MATLAB 求解[M]. 北京: 清华大学出版社, 2004.