

# 基于区间型数据的城市家庭收入分布与基尼系数测算方法研究<sup>①</sup>

张金宝<sup>1</sup> 廖理<sup>2</sup>

(1. 北京第二外国语学院财务管理系; 2. 清华大学五道口金融学院)

**【摘要】** 本文提出在不完全信息条件下, 估计家庭收入分布和测算基尼系数的方法。利用清华大学中国金融研究中心 24 个城市的家庭消费金融调查数据, 估计我国地级以上城市家庭的收入分布。研究表明, 复合概率分布更适合描述我国城市家庭的收入分布, 我国城市居民家庭群体具有一定的复杂性。我国地级以上城市居民家庭收入的基尼系数在 0.36267~0.38216 之间, 略高于现有文献对城镇居民收入的基尼系数估计, 表明基于数值型数据或分组数据计算的基尼系数存在一定程度的低估。

**关键词** 区间型数据 收入分布 基尼系数

**中图分类号** F064.1 **文献标识码** A

## Research on the Methods of Estimating the Income Distribution and Gini Coefficient of Urban Household base on the Interval Data

**Abstract:** This paper proposes a method of estimating income distribution and Gini coefficient under incomplete information. Using the data derived from the household survey of 24 cities, we study the form of income distribution of urban household and find that complex distribution is more suitable to describe the income distribution of urban household in China, which testifies that urban household of China is made of different groups. The Gini coefficient of urban household is between 0.36267~0.38216, which is slightly higher than the published research results. It suggests that the existing calculation by using numeric data or group data maybe underestimated Gini coefficient to some extent.

**Key words:** Interval Data; Income Distribution; Gini Coefficient

<sup>①</sup> 本文获国家自然科学基金重点项目“中国城市居民家庭/消费者金融研究”(71232003)的资助。感谢匿名审稿人提出的宝贵意见, 但文责自负。

## 一、问题的提出

家庭收入分布状况是收入分配制度运行的结果,研究家庭收入的分布可以为客观评价现行的收入分配制度,促进公平分配提供决策依据。目前,开展这项研究所需要的数据主要是通过调研的方式获得,但调研数据的准确性一直受到争议。其原因不外乎两个方面:一是家庭对自身收入难以准确计量。目前我国居民家庭收入结构逐渐呈现多元化特征,除非家庭成员从事收入稳定的职业,其工薪收入比较容易报告,其他不稳定的工薪收入一般都会存在一定的误差。经营性收入和财产性收入这两项,如果家庭对自己收入进行核算的时间与调研的时点存在差异,受访者很难在较短的受访时间内对家庭收入进行精确计量,这导致“家庭的收入结构越复杂,收入数据的误差就越大”。二是出于谨慎动机和不愿露富的心理,多数家庭并不愿意将家庭的真实收入报告给调研人员,这意味着调研数据的真实性无法保证。换句话说,尽管形式上受访者可以报告一个很“精确”的数字,但并不可靠。若依此类数据开展研究,有可能扭曲真实的收入分布状况,得出不当的结论。

为了解决这个问题,一些学者和业界专家进行了有益的探讨和尝试,取得了一定的经验。Campbell (2006) 建议在抽样调查中,并不强求受访者填写收入变量的准确数值,而是将调查变量的取值范围划分成若干个子区间 (Interval) 来供受访者选择。例如,理论上家庭年收入的取值范围可以表述为  $[0, +\infty)$ 。在调研的过程中,以千元为单位,我们可以将家庭年收入的取值范围划分成  $[0, 5)$ ,  $[5, 10)$ ,  $[10, 20)$ ,  $[20, 30)$ ,  $[30, 40)$ ,  $[40, 60)$ ,  $[60, 100)$ ,  $[100, 200)$ ,  $[200, +\infty)$  这样一组连续的区间。这种处理可以消除受访者被要求填列真实收入时戒备和疑虑心理,使其更愿意配合调查工作,不仅可以提高有效问卷的比率,而且比起不愿意透露真实收入而随意填写数值而言,调查所得到的数据可能更具可靠性。其次,对于确实难于精确估算收入的家庭而言,将变量的取值范围划分成一系列子区间供受访者选择,具有一定的合理性。相对于报告一个精确的数值而言,家庭居民可能更有把握回答收入所属的区间。

在居民家庭的抽样调查实践中,这种“将调查变量的取值范围划分为一系列子区间,然后供受访者选择”的方法已成功地运用到美国居民家庭的健康与退休状况 (Health and Retirement Survey) 调查中 (Thomas 和 Smith, 1997)。美国的消费金融调查 (Survey of Consumer Finance, SCF) 也采用类似的方法。为区分数值型的收入数据,本文将这种数据定义为家庭收入的区间型数据。尽管它形式上类似于胡祖光 (2004)、程永宏 (2008) 等提到的分组数据,但却有本质不同。分组数据严格地说,它不是原始数据,是对原始的数值型数据进行加工后得出的。而区间型的数据是在调研过程中获得的原始数据,虽然在调研过程中容易获得,也比较可靠,但也损失掉部分信息,就是说调研获得的收入信息是不完全的。

收入信息的不完全,给研究居民家庭收入分布带来困难。以基尼系数为例,理论上精确地计算基尼系数,必须利用收入的准确数值进行计算。即便是如胡祖光 (2004) 那样利用分组数据计算基尼系数,每组收入所占全部收入的比重也必须事先知道。但区间型收入数据,无法提供这样的研究条件。尽管如此,这种不完全信息毕竟给我们提供了家庭收入分布的一份图景,考虑区间型数据在数据可靠性方面有其独特的研究价值,利用区间型数据研究家庭收入分布是有意义的。

2010年清华大学中国金融研究中心开展对全国24个城市的家庭调研工作,获得调查样本家庭的区间型的收入数据,为本文利用区间型数据开展家庭收入分布的研究和基尼系数的

测算提供了数据条件。为此,我们先对家庭收入分布和基尼系数测算的研究进行梳理,找到适合于区间型数据的研究方法,通过清华大学中国金融中心的调研数据,对我国城市居民家庭的收入分布和基尼系数测算进行实证研究。

## 二、文献回顾

理论上居民家庭的收入分布对居民家庭收入反映的信息最完全。尽管为便于直观比较,学者们也根据家庭的收入分布计算基尼系数、泰尔熵系数、收入分布方差等,然而作为衡量居民家庭收入不平等的指标(迟巍等,2006),居民收入分布是计算这些指标的基础,所以我们重点梳理关于家庭收入分布的研究。实际上如果我们把家庭收入研究看成是一个整体工作的话,这个研究包含数据和方法两个部分,离开了调查数据的具体形式来谈论方法可能并不恰当。因此,在讨论每种方法时我们也关注它对数据的适应性问题。考虑基尼系数是衡量家庭收入分配不平等的一个重要指标,我们也对基于收入分布的基尼系数计算作相应的说明。

对居民家庭收入分布的描述有两种方法:离散分布和连续分布。假定家庭的收入  $X$  是离散分布的,在调查中抽取了  $n$  个家庭,用  $x_i$  ( $0 \leq x_i < +\infty, i=1, \dots, n$ ) 代表第  $i$  个家庭收入,如果将  $n$  个家庭的收入  $x_i$  按照非递减的顺序排列成一个向量  $x$ , 即  $x = [x_1, x_2, \dots, x_n]$ , 那么  $x$  就代表家庭收入的离散分布形式。多数文献对离散分布的基尼系数计算进行了介绍(陈昌兵,2006;田为民,2012),指出调查样本的基尼系数  $G_{mi}$  可以用式(1)来表示。

$$G_{mi} = \frac{1}{2\mu^2} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j| \quad (1)$$

其中,  $\mu$  为样本家庭收入的均值。但通常情况下,这种对家庭收入的离散描述并没有考虑样本的代表性问题。换句话说,式(1)计算的基尼系数只是反映调查样本的收入不公平程度。Thomas等(2000)给出考虑样本代表性的基尼系数的计算。假定调查样本中收入为  $x_i$  的家庭在总体中所占的权重为  $w_i$ , 那么反映总体收入平等的基尼系数  $G_{mi}$  的计算则可表示为:

$$G_{mi} = \frac{1}{\mu} \sum_{i=1}^n \sum_{j=1}^n w_i |x_i - x_j| w_j \quad (2)$$

用离散分布的方法描述家庭收入分布,由于运用每个样本家庭的收入信息,所以对调查信息运用得较为充分。这种方法的不足是:它只适用于数值型的数据,要求家庭报告的收入数据是对家庭收入的点估计,但由于本文开篇讨论的原因,数值型数据的准确性常常受到质疑,如果把数据因素考虑进去,结果未必能真实反映居民家庭的收入不平等情况。

另一种方法是用连续分布函数来描述家庭的收入分布。从分布密度函数估计的方法角度又可细分为参数估计方法和半参数估计方法两种。常见的半参数估计方法是用核函数平滑方法来估计家庭的收入分布。核函数平滑方法中代表性的文献是Dinardo等(1996)在 *Econometrica* 上发表的估计家庭收入的一篇论文。此后,我国学者迟巍等(2008)、陈昌兵(2007)采用这种方法估计我国居民收入的分布函数。用核平滑化方法估计的收入分布密度函数的公式为:

$$\hat{f}(x) = \sum_{i=1}^n \frac{w_i}{h} K\left(\frac{x-x_i}{h}\right) \quad (3)$$

其中,  $\hat{f}$  代表估计的收入分布密度函数。  $x_1, x_2, \dots, x_n$  代表调查样本所包含的  $n$  家庭的收入,  $w_i$  代表第  $i$  个家庭在总体中的权重。  $K(\cdot)$  代表核函数,  $h$  代表带宽。该方法的优点是不必拘泥于分布函数的特定形式。不足在于, 虽然回避了参数估计方法中分布密度函数的选择问题, 但会面临核函数和带宽选择的问题。同时, 在接近家庭收入  $x$  观察值的边界, 由于观察点较少会降低密度函数的精度。问题是在家庭收入两端的观察值, 分别代表最低收入和最高收入的家庭, 而基尼系数对这种误差比较敏感。此外, 从数据条件的角度看, 核函数平滑方法同样要求数值型的数据。

收入分布密度函数的参数估计方法, 是假定家庭收入  $x$  的概率密度具有特定的函数形式, 用调查获得的数据来估计分布函数的参数, 从而获得家庭收入的分布函数。相当多的学者采用这种方法来估计收入的分布函数。例如, 胡志军 (2012)、Rao 和 Tang (2006)、程永宏 (2007、2008)、王海港和周开国 (2006)、Schmittlein (1983)、McDonald 和 Bartell (1979)、Singh 和 Maddala (1976) 等。也有学者对该种方法提出批评 (王亚峰, 2012), 认为多数概率密度函数的选择缺乏理论基础。本文对这种看法持保留意见, 从国外的研究看 (Singh 和 Maddala, 1976; Schmittlein, 1983), 概率密度函数的选择是基于对家庭收入增长规律的认识, 所采用的技术类似于寿险中的死亡率模型。假定家庭收入用  $X$  表示,  $F(x)$  为其分布函数,  $f(x)$  为分布密度函数。令  $\bar{F}(x) = 1 - F(x)$ , 那么家庭收入超过  $x$  的条件下, 收入终止于区间  $(x, x + \Delta x]$  内的概率为:

$$P(x < X \leq x + \Delta x | X > x) = (F(x + \Delta x) - F(x)) / \bar{F}(x) = \frac{f(x)}{\bar{F}(x)} \Delta x + o(\Delta x) \quad (4)$$

所以当  $\Delta x \rightarrow 0$  时, 定义:

$$T(x) = \lim_{\Delta x \rightarrow 0} P(x < X \leq x + \Delta x | X > x) / \Delta x = f(x) / \bar{F}(x) \quad (5)$$

那么  $T(x)$  可以用来表示一个收入为  $x$  的家庭, 其收入终止于  $x$  附近可能性的大小。欧阳植和于维生 (1994) 称  $T(x)$  为停滞率, 假定它是家庭收入  $X$  的线性函数, 能较好地拟合我国城镇居民家庭 1990 年的收入分布。当然也可以根据研究中样本家庭的具体情况, 对  $T(x)$  做更精确的描述。例如, Singh 和 Maddala (1976)、Schmittlein (1983) 假定  $T(x)$  是  $\ln(x)$  的函数来描述家庭收入的增长规律。通过求解含有  $F(x)$  和  $F'(x)$  即  $f(x)$  的微分方程, 即式 (5) 来获得家庭收入的分布密度函数。这意味着, 只要我们能够发现  $T(x)$  的规律性, 用参数估计的方法描述家庭收入的概率密度函数是有理论基础的。这也是本文利用区间型数据估计家庭收入分布时, 考虑采用参数估计方法的原因之一。当然在研究过程中, 考虑到微分方程的可解性问题, 有些概率密度函数形式所对应的  $T(x)$  对家庭收入增长规律的描述与实际会有些偏差, 从而导致在收入的某个区间对实际数据的拟合误差较大的情况<sup>①</sup>。本文尝试用复合分布的函数形式 (Composite Distribution) 加以克服。

在确定收入分布函数的具体形式之后, 通常可以采用极大似然估计和最小二乘估计的方法来估计分布函数的参数。两种方法均适用于数值型的数据以及分组数据, 当然也适用于本

① 如对数正态分布在描述家庭收入的低端部分拟合效果较好, 而对高端收入部分拟合效果不好就是该原因。

文提出的区间型数据。但利用最小二乘估计方法，可以便于我们考察不同的函数形式究竟在家庭收入分布的哪个区间拟合的效果更好，从而更好地判断不同分布函数描述家庭收入分布时所表现出来的性质，有助于我们思考复合分布函数的构成问题。

与离散分布一样，当收入分布是连续分布时同样可以计算出基尼系数。Dorfman (1979)、Kendall 和 Stuart (1977) 在严格意义上给出利用收入分布函数精确计算的基尼系数  $G_{mi}$  的公式：

$$G_{mi} = 1 - \frac{1}{\mu} \int_0^{+\infty} (1 - F(x))^2 dx \quad (6)$$

其中， $\mu$  代表收入变量的均值， $F(x)$  为家庭收入的分布函数。当家庭收入数据是分组数据时，也有研究文献直接利用分组数据计算基尼系数<sup>①</sup>。例如胡祖光 (2004) 给出一种近似计算方法为：

$$G_{mi} = \frac{1}{n} \sum_{i=2}^n \sum_{j<i} (p_i - p_j) \quad (7)$$

式 (7) 中  $p_i$  代表第  $i$  组家庭收入占全部家庭收入的比重，考虑到分组数据是根据调查的原始数据加工后给出的， $p_i$  通常是已知的。但本文由于采用的是原始区间型数据， $p_i$  是未知的，这意味着本文的数据条件并不适合该方法。

通过讨论家庭收入分布中常见的方法可以看出，不同方法对数据的要求不尽相同。本文采用的区间型数据，由于无法获得家庭收入的准确数值，一些处理数值型数据的研究方法，如离散分布中的基尼系数计算和家庭收入的概率密度函数的核平滑估计方法等并不适合本文的研究。区间型数据甚至缺乏分组数据所能提供的一些信息，如每个组别的收入占全部样本家庭的比重等，因此适合分组数据的基尼系数计算方法，如式 (7) 也无法采用。但区间型的调研数据能够提供属于相应的收入区间的家庭频数，这为利用参数估计的方法来估计家庭的收入分布创造数据条件。相关的研究文献多是采用单一的函数形式进行收入分布研究，本文在家庭收入分布的具体形式上做更多的尝试，包括采用已知分布的复合分布形式，以便找到最优的分布函数表达形式。

### 三、收入分布估计与基尼系数测算的基本思想

#### 1. 关于符号的定义

在叙述估计家庭收入分布和测算基尼系数的方法之前，我们先对区间型数据的符号做出规定。我们用变量  $x$  表示家庭的收入，不妨假定在入户调研时，我们将家庭收入的取值范围划分成  $g$  个首尾相连的区间： $[x_1, x_2)$ ， $[x_2, x_3)$ ， $[x_3, x_4)$ ， $\dots$ ， $[x_{g-1}, x_g)$ ， $[x_g, +\infty)$ 。当受访者选择第  $i$  个区间时，表示该户家庭的收入  $x$  满足  $x_i \leq x < x_{i+1}$ 。为简单起见，我们将这样一个划分用向量  $X_{point}$  来表示：

$$X_{point} = (x_1, x_2, \dots, x_{g-1}, x_g) \quad (8)$$

<sup>①</sup> 程永宏 (2006)、金成武 (2007) 分别给出在连续分布和离散分布条件下，基尼系数按组分解的计算公式，但侧重于基尼系数分解的经济意义，并且前提是连续分布和离散分布均是已知的，这同样也不适合本文的区间型数据。

其中,  $x_1=0$  表示家庭收入的下限为 0<sup>①</sup>。假定入户调研抽取 N 个家庭, 我们用  $n_i$  表示落入第  $i$  个区间的家庭的个数。用  $h_i$  表示样本家庭落入第  $i$  个区间的频率, 则:

$$h_i = \frac{n_i}{N} \quad (9)$$

正是  $X_{point}$  和  $h_i$  给我们描述了家庭收入分布的初步轮廓。

### 2. 收入分布估计和基尼系数测算的基本思想

无论我们是否能够估计家庭收入的真实分布, 事实上它都是客观存在的。假定真实的家庭收入分布为  $F(x/\theta)$ , 其中  $\theta=(\theta_1, \theta_2, \dots, \theta_l)$  为描述家庭收入分布的一组参数。因此, 理论上落入第  $i$  个区间的家庭的频率  $h'_i$  可以通过式 (10) 进行计算:

$$h'_i(\theta) = \int_{x_i}^{x_{i+1}} f(x/\theta) dx \quad (10)$$

其中,  $f(x/\theta)$  为家庭收入的概率密度函数。尽管分布参数  $\theta$  并不知道, 但给定具体的收入分布形式后,  $\theta$  的取值应该使理论上计算得出的频数  $h'_i$  与利用实地调研数据计算得出的频率  $h_i$  非常接近, 这种接近我们用下面的平方和 (SSE) 来表示:

$$SSE = \sum_{i=1}^g (h_i - h'_i)^2 \quad (11)$$

于是, 通过区间型数据估计家庭的收入分布可以转化成这样一个数学问题:

$$SSE = \text{Min}_{\theta \in \Theta} \sum_{i=1}^g (h_i - h'_i(\theta))^2 \quad (12)$$

即给定一个具体的收入分布形式, 我们在所有可能的参数组合  $\Theta$  中找到一个最优的参数向量  $\theta^*$ , 使得通过式 (12) 估计的收入分布最接近调研获得的实际结果。我们把这个分布  $F^*$  看作是城市居民家庭收入分布  $F$  的最佳估计。这就建立了利用区间型数据估计家庭收入分布的模型。这里有必要强调一下式 (12) 的经济意义, 式 (12) 所表示的误差 SSE 越小, 意味着我们估计家庭收入分布时, 对调研获得的收入信息利用得越充分, 也就是说损失的信息就越少。

家庭收入分布问题解决后, 基尼系数的计算问题迎刃而解。根据式 (6) 我们给出基尼系数  $G_{mi}$  的计算公式:

$$G_{mi} = 1 - \frac{1}{\mu} \int_0^{+\infty} (1 - F^*(x))^2 dx \quad (13)$$

其中,  $\mu$  代表收入变量的均值,  $F^*$  为利用式 (12) 估计得出的家庭收入的分布函数。

### 3. 家庭收入分布函数的选择

在式 (10) 中, 我们没有给出家庭收入分布函数的具体形式。究竟什么样的概率分布函数适合描述中国家庭收入的分布, 理想的解答方式应是采用家庭收入的原始数据, 通过对中国居民家庭收入分布的停滞率  $T(x)$  的精确描述, 再求解式 (5) 所表达的微分方程来求出分布函数的具体形式。但原始数据要求家庭对收入的报告是一个确定的数值。这种性质的

<sup>①</sup> 极端的情况下, 若家庭全部的收入来源于有风险投资或经营活动, 有可能出现家庭收入小于 0 的情况, 但这种情况较为少见。

原始数据在国内鲜有公开发表，可以借鉴的研究成果几乎没有。

因此，我们借鉴国外关于家庭收入分布函数方面的研究成果，来选取适当收入分布函数。通常家庭收入的分布具有明显的“厚尾”特征，因此一些有偏的分布函数被学者们用来描述家庭的收入分布。Aitchison 和 Brown (1957) 利用对数正态分布函数研究收入分布时发现，对数正态分布对收入分布的高端部分拟合效果并不好。Mandelbrot (1960) 利用 Pareto 分布拟合收入分布时，发现除了的家庭收入的高端部分拟合较好外，其他部分拟合的效果较差。Salem 和 Mount (1974)、McDonald 和 Bartell (1979) 利用 Gamma 分布进行实证研究，得出的结论是 gamma 分布在收入分布的两端拟合效果优于对数正态分布。前述的这 3 种分布都是具有两个参数的分布函数。Singh 和 Maddala (1976) 利用一种新的分布函数（以下简称 Sm 函数）研究美国家庭的收入分布数据，发现 Sm 分布函数的拟合效果好于对数正态分布和 Gamma 分布。Schmittlein (1983) 运用一种被称为 Burr 分布的函数对美国人口普查局提供的 47500 个美国家庭的收入分布进行拟合，发现 Burr 分布对较大样本的数据拟合效果较好。Rao 和 Tang (2006) 采用广义的第二类 Beta 分布估计中国城镇、农村和城乡一体化的收入分布，此外他们也尝试采用 Weibull 分布来估计中国居民的收入分布，所采用的数据来自《中国统计年鉴》，虽然数据特征有别于本文提出的区间型数据，但收入信息同样是不完整的。

以上我们对参数估计方法中收入分布函数的具体形式进行了总结。事实上由于国情有别，居民家庭收入的结构也有所不同。上述分布函数在拟合家庭收入分布方面所体现的特点，未必在拟合中国家庭的收入分布时也成立。为严谨起见，在后续的实证研究中，我们逐一采用现有的研究文献提及的分布函数，利用区间型数据进行收入分布研究。这些分布函数的具体表达形式见表 1。

表 1 描述家庭收入分布的几种概率分布函数

分布 (缩写)	参数	概率密度函数	累积分布函数
Beta-II <sup>①</sup> (Be)	$b, p, q$ $p, q > 1$ $b > 0$	$\frac{x^{p-1}}{b^p B(p, q) \left(1 + \frac{x}{b}\right)^{p+q}}$	$\frac{1}{B(p, q)} \int_0^{x/(b+x)} t^{p-1} (1-t)^{q-1} dt$ $= B_{x/(x+b)}(p, q)$
Burr (Bu)	$r, \alpha, \beta$ $r, \alpha, \beta > 0$	$\frac{r\beta}{\alpha} x^{\beta-1} \left(\frac{\alpha}{\alpha+x^\beta}\right)^{r+1}$	$1 - \left(\frac{\alpha}{\alpha+x^\beta}\right)^r$
Gamma (Ga)	$\alpha, \beta$ $\alpha, \beta > 0$	$\frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}$	$\frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t/\beta} dt$
Lognormal (Ln)	$\mu, \sigma$	$\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$	$\Phi\left(\frac{\ln x - \mu}{\sigma}\right)$ <sup>②</sup>
Pareto (Pa)	$\kappa, x_m$ $\kappa > 1, x_m > 0$	$\frac{\kappa x_m^\kappa}{x^{\kappa+1}}$	$1 - \left(\frac{x_m}{x}\right)^\kappa$

① Beta-II 的分布函数和密度函数中  $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 t^{p-1} (1-t)^{q-1} dt$ 。

②  $\Phi(\cdot)$  为标准正态分布的累积分布函数。

(续)

分布 (缩写)	参数	概率密度函数	累积分布函数
Singh-Maddala (Sm)	$\alpha_1, \alpha_2, \alpha_3$	$\frac{\alpha_1 \alpha_2 \alpha_3 x^{\alpha_2 - 1}}{(1 + \alpha_1 x^{\alpha_2})^{\alpha_3 + 1}}$	$1 - \frac{1}{(1 + \alpha_1 x^{\alpha_2})^{\alpha_3}}$
Weibull (Wb)	$\alpha, \beta$ $\alpha, \beta > 0$	$\frac{\alpha x^{\alpha - 1} e^{-(x/\beta)^\alpha}}{\beta^\alpha}$	$1 - e^{-(x/\beta)^\alpha}$

#### 四、实证研究

##### 1. 关于数据

本文选用清华大学中国金融研究中心在 2010 年对全国地级以上城市居民家庭进行消费金融调查所收集的数据。调研涉及家庭的基本信息、理财行为、资产负债情况、收入和支出、投资和融资、家庭的住房、医疗和养老保险、家庭的遗产规划等内容。此次调研根据我国地级以上城市规模、经济发展水平的差异，调查将这些城市分为 3 类。即经济发达的城市，经济较发达的城市和经济发展水平一般的城市（廖理和张金宝，2010）。考虑我国地区间经济发展的不平衡，本次调查将全国的城市（不含港澳台地区）在地理位置上按东北、华北、华东、华南、华中、西南、西北 7 个大区进行划分，抽样家庭的样本数量在各大区之间根据家庭户数按比例分配。除华东地区每个类别的城市各遴选两个外，其余的各大区遴选的城市中，每类城市各包含一个，共计 24 个城市<sup>①</sup>。在每个抽取到的城市中，通过随机抽样的方式抽取居民居住的小区 and 访问家庭的样本。调研获得的数据基本满足随机性要求，基本反映我国地级以上城市家庭的经济和金融状况。

此次调研共收集 5273 个家庭的数据，其中家庭收入变量调查的是家庭的可支配收入，采用区间型的数据形式。剔除掉无效样本后，共获得 5122 个家庭收入的数据。调研将家庭收入  $x$ （单位：元）的取值范围划分成 14 个区间，区间划分点的向量表示为：

$$X_{point} = 10^3 \times (0, 5, 10, 20, 40, 60, 100, 150, 200, 300, 400, 500, 600, 800)$$

向量的第  $i$  个分量代表第  $i$  个区间的左端点，家庭收入落入每个区间的频数见表 2。

表 2 样本家庭收入区间统计

$i$	1	2	3	4	5	6	7
$h_i^s$ (%)	1.9328	1.7962	6.9309	25.0488	26.5131	25.2245	8.688
$n_i$	99	92	355	1283	1358	1292	445
$i$	8	9	10	11	12	13	14
$h_i^s$ (%)	2.0305	1.0933	0.3514	0.1952	0.0781	0.0976	0.0195
$n_i$	104	56	18	10	4	5	1

<sup>①</sup> 在调研所涉及的城市中，第一类城市为：北京、沈阳、上海、济南、广州、重庆、西安和武汉；第二类城市有：包头、吉林、徐州、南昌、海口、昆明、乌鲁木齐、洛阳；第三类城市为：朔州、伊春、安庆、泉州、桂林、攀枝花、白银、株洲。



2. 关于样本和总体

表 2 给出的是调研样本数据的统计结果。也就是说，表 2 中计算的每个家庭落入相应区间的频率，并没有考虑样本家庭的代表性问题，为了以示区分，我们用  $h_i^p$  来表示。实际上，在经济发达的城市、经济较发达的城市和经济发展水平一般的三类城市中，家庭经济状况的差异性是不一样的。经济发展水平较低的一类城市，家庭之间的经济差异也较小。这意味着通常调研时可以采用相对较小的样本数量来反映这类家庭的经济状况，同时也说明这类样本的代表性较强。

因此，当我们通过样本数据研究全国地级以上城市家庭的收入分布时，我们必须考虑抽样数据中每个样本家庭的代表性问题。我们定义落在第  $i$  区间的来自第  $j$  类城市的第  $k$  个家庭所代表的家庭数目为  $w_{i,j,k}$ <sup>①</sup>，则考虑样本家庭的代表性，对落在每个收入区间的家庭频数进行重新计算的公式为：

$$h_i^p = \frac{\sum_{j=1}^3 \sum_{k=1}^{n_{i,j}} w_{i,j,k}}{\sum_{i=1}^g \sum_{j=1}^3 \sum_{k=1}^{n_{i,j}} w_{i,j,k}} \tag{14}$$

其中， $i \in (1, 2, \dots, g)$  代表特定的区间， $j \in (1, 2, 3)$  代表三类城市中的某一类城市， $n_{i,j}$  代表收入落在第  $i$  个区间，来自第  $j$  类城市家庭的数目， $\sum_{j=1}^3 n_{i,j} = n_i$ 。经过调整后，每个收入区间的对应的家庭频数见表 3。我们将根据这个频率表对我国地级以上城市家庭收入的分布情况进行研究。

表 3 城市家庭收入区间统计

$i$	1	2	3	4	5	6	7
$h_i^p$ (%)	2.8420	3.0520	9.5510	27.8040	23.9640	22.8100	7.2490
$n_i$	99	92	355	1283	1358	1292	445
$i$	8	9	10	11	12	13	14
$h_i^p$ (%)	1.3350	0.8520	0.2610	0.1040	0.1120	0.0540	0.0100
$n_i$	104	56	18	10	4	5	1

3. 城市家庭收入分布状况分析

表 4 给出采用数值方法求解公式 (12) 后得出我国地级以上城市居民家庭收入分布和基尼系数的测算结果。在分布参数确定的情况下，我们也给出利用区间型的家庭收入数据对家庭平均收入的测算结果。从表 4 可以看出，Beta-II 型的分布函数描述地级以上城市家庭收入分布的效果最好，表示 Beta-II 型的分布函数与实际调研结果的接近程度的指标 SSE 达到 0.000838。这表明该分布描述城市家庭的收入分布，对调研获得的收入信息利用的最充分。据此测算的家庭平均收入为 52884 元，基尼系数为 0.35062。Weibull 分布和对数正态分布的效果略差些。Singh-Maddala 分布、Gamma 分布和 Pareto 分布的效果更差些，测算得出家庭收入均值和基尼系数参考意义不大。

① 详细计算方法的说明请参见《中国消费金融调研报告》(廖理和张金宝, 2011)。

表 4 地级以上城市家庭收入分布测算结果

分布类型	参 数	SSE_l	SSE_m	SSE_h	标准差	均 值	G <sub>mi</sub>
Beta-II <sup>①</sup> (Be)	p 2.481	0.54941	0.002718	5.3114	0.000838	52884	0.35062
	q 42.75						
	b 8.89×10 <sup>5</sup>						
Weibull (Wb)	α 1.6261	0.23937	0.005573	6.2816	0.001061	49131	0.34706
	β 54877						
Lognormal (Ln)	μ 10.766	1.37250	0.006272	3.3561	0.002257	59893	0.37177
	σ 0.6848						
Burr (Bu)	α 6.28×10 <sup>8</sup>	0.22532	0.008929	5.0643	0.013664	53330	0.36466
	β 1.6963						
	γ 5.7517						
Singh-Maddala (Sm)	α <sub>1</sub> 1.65×10 <sup>-6</sup>	7.7083	0.39214	2738.6	0.034296	70178	0.57839
	α <sub>2</sub> 1.0118						
	α <sub>3</sub> 3.5012						
Gamma (Ga)	α 145.37	12.7550	3.26900	8.0000	0.31292	20435	0.04675
	β 140.57						
Pareto (Pa)	κ 1.65×10 <sup>5</sup>	1170.7	3.0000	8.0000	1.1463	657.67	0.00001
	χ <sub>m</sub> 657.67						

为了深入研究各种概率分布函数描述家庭收入的特点，我们也考察表 1 中所列的各函数在描述家庭收入的低端、中端和高端时的表现。关于收入的低、中、高端的定义，目前并没有统一的标准。《中国城市（镇）生活与价格年鉴》将“收入最低的 20%家庭的年收入的最大值”作为低端收入和中端收入的分界点，将“收入最高的 20%家庭年收入的最小值”作为高端收入与中端收入的分界点。然而，区间型的家庭收入数据由于无法提供准确的收入信息，并不适合这种定义。为此，我们将第 1、2、3 个区间作为家庭收入分布的低端，将第 4、5、6 个区间作为家庭收入分布的中端，其余的第 7~14 个区间作为收入分布的高端。分属这 3 个区间的家庭数目依次占家庭总数的 15.5%、74.5%、10%左右。用 SSE\_l、SSE\_m、SSE\_h 这 3 个指标，分别表示给定的收入分布函数在低、中、高端收入部分与实际调研的结果的接近程度，与式（11）有所不同，我们给出的是相对意义上的定义，目的是为了放大 3 个指标的数值，更加便于比较。3 个指标的定义见式（15）：

$$SSE_l = \sum_{i=1}^3 \left( \frac{h_i^p - h_i'}{h_i^p} \right)^2 \quad SSE_m = \sum_{i=4}^6 \left( \frac{h_i^p - h_i'}{h_i^p} \right)^2 \quad SSE_h = \sum_{i=7}^{14} \left( \frac{h_i^p - h_i'}{h_i^p} \right)^2 \quad (15)$$

由表 4 可以看出，Beta-II 型的分布函数整体上比较适合描述城市家庭的收入分布，这个结论与胡志军（2012）相似。但仔细分析表 4 的计算结果就会发现，它对家庭收入的高端部分的描述却并没有对数正态分布的效果好。同样，它对收入低端部分的描述也没有

① Beta-II 的分布函数和密度函数中  $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 t^{p-1}(1-t)^{q-1} dt$ 。

Weibull分布和Burr分布的表现好,说明用Beta-II描述城市居民家庭的收入分布未必是最理想的选择。同时也说明不同概率分布函数在描述城市家庭收入分布不同区间段时,表现各有优劣。这个结果启示我们,若采用这些分布构成一个复合分布或许能够更好地描述我国城市家庭的收入分布。

4. 利用复合分布描述我国城市居民家庭的收入分布

(1) 复合概率分布的含义。考虑一种最简单的复合分布,即假定城市居民家庭收入的分布是由两个概率分布 $F_1$ 和 $F_2$ 复合而成:

$$F(x) = \lambda F_1(x) + (1-\lambda) F_2(x) \tag{16}$$

从经济意义上说,式(16)实际上意味在考虑城市居民家庭收入时,我们可以认为城市居民家庭来自两个子总体, $\lambda$ 、 $1-\lambda$ 分别代表两个子总体占家庭总量的比重。当选定 $F_1$ 和 $F_2$ 的具体形式后,我们仍然可以将式(16)代入到式(12)中,对城市居民家庭的收入分布做出估计。我们从表1所列的概率分布函数中,任选两个函数作为 $F_1$ 和 $F_2$ ,探索不同的复合分布描述城市居民家庭收入的效果,这样的组合共有 $2^7$ 个。限于篇幅,表5给出SSE排名前14位的复合分布描述城市居民家庭收入分布所得出得计算结果。

表5 复合分布函数描述城市家庭收入分布测算结果

复合分布	参 数		$\lambda$	标准差	均 值	$G_{mi}$		
	$F_1$	$F_2$						
Bu-Bu	$\alpha$	$1.70 \times 10^{10}$	$\alpha$	$4.76 \times 10^8$	0.9607	0.000022	52495	0.36267
	$\beta$	2.1068	$\beta$	2.6844				
	$\gamma$	2.0811	$\gamma$	0.26491				
Sm-Bu	$\alpha_1$	$1.0 \times 10^{-11}$	$\alpha$	$2.91 \times 10^9$	0.0181	0.000115	51727	0.36438
	$\alpha_2$	2.6234	$\beta$	1.8969				
	$\alpha_3$	68.873	$\gamma$	3.3191				
Ln-Ln	$\mu$	9.8055	$\mu$	10.82	0.20483	0.000134	53772	0.37523
	$\sigma$	1.1186	$\sigma$	0.5715				
Wb-Be	$\alpha$	1.6295	$p$	15.779	0.25312	0.000199	53966	0.37439
	$\beta$	22827	$q$	5.1446				
			$b$	17160				
Bu-Be	$\alpha$	$2.60 \times 10^6$	$p$	11.959	0.33875	0.000205	54496	0.38216
	$\beta$	1.2181	$q$	4.6804				
	$\gamma$	7.4575	$b$	19533				
Bu-Ln	$\alpha$	$3.33 \times 10^8$	$\mu$	10.792	0.08373	0.000206	54381	0.37104
	$\beta$	1.7658	$\sigma$	0.6802				
	$\gamma$	28.627						
Wb-Ln	$\alpha$	1.00	$\mu$	10.818	0.16149	0.000210	53350	0.37379
	$\beta$	20981	$\sigma$	0.5957				
Pa-Be	$\kappa$	$1.0 \times 10^8$	$p$	2.529	0.02439	0.000224	50876	0.35603
	$\chi_m$	1.4959	$q$	197.8				
			$b$	$4.06 \times 10^6$				

(续)

复合分布	参 数				$\lambda$	标准差	均 值	$G_{mi}$
	$F_1$		$F_2$					
Pa-Ga	$\kappa$	$9.92 \times 10^{11}$	$\alpha$	2.4959	0.02400	0.000236	50778	0.35603
	$\chi_m$	7.9128	$\beta$	20845				
Wb-Bu	$\alpha$	48582	$\alpha$	$1.38 \times 10^9$	0.01436	0.000315	50657	0.36238
	$\beta$	$1.0 \times 10^{-11}$	$\beta$	1.7850				
			$\gamma$	5.1517				
Sm-Ln	$\alpha_1$	$3.46 \times 10^{-5}$	$\mu$	10.797	0.15938	0.000343	55128	0.38408
	$\alpha_2$	0.92496	$\sigma$	0.6010				
	$\alpha_3$	2.8232						
In-Be	$\mu$	10.519	$p$	7.9823	0.51221	0.00060	57535	0.39395
	$\sigma$	0.94953	$q$	11.881				
			$b$	77659				
Pa-Sm	$\kappa$	4.7684	$\alpha_1$	$1.0 \times 10^{-8}$	0.14175	0.00072	53179	0.37251
	$\chi_m$	35245	$\alpha_2$	1.4499				
			$\alpha_3$	12.583				
Be-Be	$p$	6.7096	$p$	39.93	0.26049	0.00073	58204	0.40063
	$q$	1.7691	$q$	4.0986				
	$b$	4645.8	$b$	5000				

(2) 复合概率分布描述城市家庭收入分布的测算结果。对比表 4 和表 5 的计算结果我们发现, 利用复合分布能够显著提高描述城市家庭收入分布的精度。表 5 中的复合分布描述家庭收入分布的效果, 均好于单一的概率分布函数形式。描述效果最好的 Bu-Bu 复合分布, 能够将标准差提高到  $10^{-5}$ 。这说明利用这种分布描述城市居民家庭的收入, 损失的信息最小。究其原因, 复合分布的参数多于单一的概率分布函数, 这为描述家庭收入分布提供了更多的自由度<sup>①</sup>。从收入的角度看, 也说明我国城市居民家庭群体的构成比较复杂。图 1 给出根据标准差最小的 Bu-Bu 组合分布绘出的我国城市居民家庭收入的洛伦兹曲线。

从表 5 中可以看出, 表 4 中描述家庭收入效果并不好的概率分布函数 (如 Burr 分函数、Sm 分布函数), 当它们与其他函数构成复合分布函数时, 也能很好地描述城市家庭的收入分布函数。因此, 在选择参加复合分布的函数时, 切勿受到单一分布函数测算结果的干扰。

针对特定的复合分布, 表 5 同样给出城市居民家庭收入的平均值和基尼系数的估计。理论上, 区间的划分点  $X_{point}$  和每个区间统计的频率  $h_i$ , 并不能完整刻画城市居民家庭的收入分布。这意味着, 可能存在多个分布函数  $F(x)$  都能够在标准差最小的意义上, 比较充分地描述城市居民家庭的收入分布。因此给定  $X_{point}$  和  $h_i$ , 基尼系数的取值未必是唯一的。这正是由于区间型数据提供的信息不完全造成的。但在尝试多种分布进行测算以后, 我们可以

<sup>①</sup> 较多的参数提供更多的自由度, 只是为更好地描述城市居民家庭收入分布提供可能, 因为它还受到分布函数具体表达形式的限制。我们没有列出的复合分布, 描述家庭收入分布的效果就没有单一的 Beta-II 型的分布函数效果好。

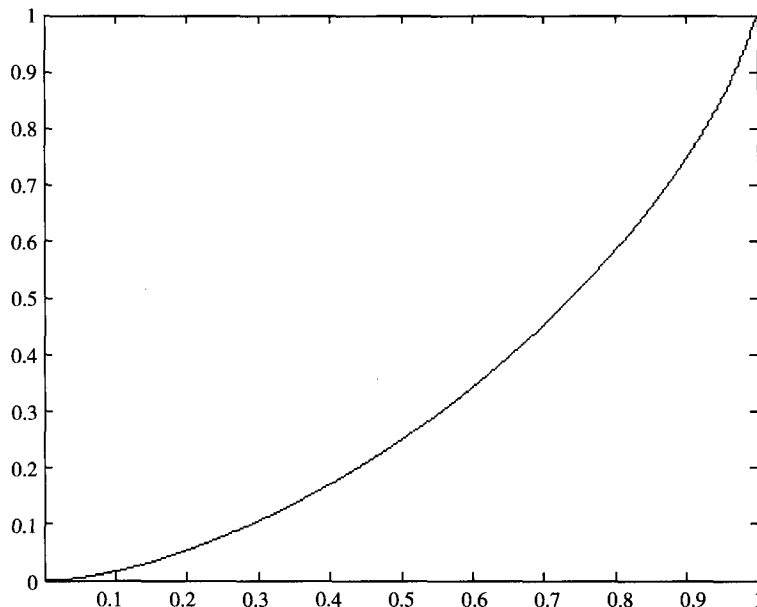


图1 我国地级以上城市居民家庭收入的洛伦兹曲线 (Bu-Bu)

给出一个基尼系数的合理取值范围。比如，在标准差小于0.000210的条件下，我们可以推断城市居民家庭的基尼系数的合理范围为0.36267~0.38216之间。

需要说明的是，本文考察的是地级以上城市居民家庭的收入分布和基尼系数的大小，如果将地级以下城市（镇）的居民家庭考虑在内，则由于这些城市的居民收入相对更低一些，测算所得的基尼系数预计要高于本文测算的结果。同样道理，如果我们将农村的家庭收入放在一起考察的话，测算所得的基尼系数还要更高些。

#### 5. 关于本文测算精度的一个模拟实验

由于调查采集的是区间型的收入数据，我们不可能获得关于居民收入的完整信息。也无法将本文估计的收入分布与真实的城市居民家庭的收入分布进行比较。为了对本文根据区间型数据测算的基尼系数的准确性进行研究，我们设计一个简单模拟试验。假定居民家庭收入的真实分布 $F^0$ 服从简单的对数正态分布，即： $x \sim \text{lognormal}(10.78, 0.7)$ 。先随机生成5000个家庭的收入数据 $x_i^0$  ( $i=1, 2, \dots, 5000$ )，这个数据是数值型的，是调查者观测不到的。为模拟入户调查中收入的区间型数据的采集过程，我们采用清华大学消费金融调查中家庭收入的区间划分方法，记录调查到每个家庭的收入 $x_i^*$  ( $i=1, 2, \dots, 5000$ )，这个 $x_i^*$ 是区间型数据，反映的是家庭收入 $x^0$ 所属的区间，是在调查过程中实际获得的数据。接着我们统计每个区间的频数 $h_i^{\text{①}}$ ，利用本文提供的方法来估计家庭的分布 $F^*$ ，测算出基尼系数 $G_{mi}^*$ ，然后比较调研测算结果 $F^*$ ， $G_{mi}^*$ 与真实分布 $F^0$ 和 $G_{mi}^0$ 之间的差异。

田为民（2012）曾采用分组数据计算分析省际的基尼系数，采用的处理方法是利用每组家庭收入的中值来代替受访者的收入然后利用本文中的式（1）来计算基尼系数。为了便于比较，我们也把本文采用的这种方法与田为民的方法进行对比。

① 为简单起见，我们假定家庭样本的代表性无差异。

我们共进行 100 次这样的模拟试验。在标准差小于 0.0004 的条件下,利用本文提供的方法所测算的基尼系数  $G_{mi}^*$  与根据真实分布测算的基尼系数  $G_{mi}^0$  之间的误差在 0.01%~2.92% 之间。所测算的家庭收入的平均值的相对误差在 0.02%~2.97% 之间。针对利用区间中值代替家庭收入的方法,我们也进行模拟计算,发现该方法测算的基尼系数的相对误差在 0.61%~4.08% 之间,所测算的家庭收入的平均误差为 1.27%~5.57% 之间。针对每次试验产生的随机数据,我们比较两种方法的精度,发现本文的方法所测算的基尼系数的精度平均高于中值替代方法 1.09 个百分点,平均收入的测算精度要平均高于中值替代方法 1.86 个百分点。这说明我们的方法要优于利用区间中值替代家庭收入的方法。

## 五、结 论

收入的区间型数据是现在开展收入调查过程中越来越被普遍采用的数据类型。由于区间型数据的特点,利用它来进行收入分布研究时,只能先通过区间型数据来估计收入的分布,之后再行计算基尼系数等指标。同已有的研究相比,本文在以下几个方面进行新的尝试:第一,搜集汇总了现有的收入分布研究中采用的分布函数的具体形式,逐一考察了每种分布函数的估计误差,这在一定程度上避免我们在选择收入分布函数时的盲目性。同时,利用最小二乘方法研究了这些函数在描述家庭收入的不同阶段所表现出来的性质。第二,采用复合分布的方法来描述家庭的收入分布,通过尝试 128 种不同分布函数的组合形式,发现用两个 Burr 分布的组合描述我国城市居民家庭的收入分布效果最好。这说明复合分布的性能要优于单一函数,同时也说明我国城市居民家庭的群体组成具有一定的复杂性。第三,由于区间型数据的可靠性相对较高,利用它计算的收入分布和基尼系数有望具有较高的参考价值。利用清华大学消费金融的调研数据,实证测算所得的地级以上城市家庭的收入分布的基尼系数介于 0.36267~0.38216 之间,略高于胡志军(2012)、王亚峰(2012)对我国城镇居民收入的基尼系数的估计结果。考虑区间型数据的可靠性,这或许表明基于数值型数据或分组数据计算的基尼系数存在一定程度的低估。

在信息不完全的条件下,测算误差是难免的。理论上除非以下两个条件完全具备,否则不可能得出真实的基尼系数:一是对全部家庭的收入进行普查;二是全部家庭在普查过程中汇报的收入是真实的收入。但上述两个条件均很难满足,尤其是收入的真实性。而采用区间型的数据,能够在一定程度上提高收入信息的真实性和可靠性。通过本文的数值模拟试验,认为本文的测算精度是可以接受的,本文的探索将有助于提升区间型收入数据的研究价值,使通过调研获得的区间型数据得到更广泛的应用。

## 参 考 文 献

- [1] Aitchison J., Brown A. C., 1957, *The Lognormal Distribution* [M], Cambridge University Press, Cambridge.
- [2] Campbell J. Y., 2006, *Household Finance* [J], *Journal of Finance*, 61, 1553~1604.
- [3] Dorfman R., 1979, *A Formula for the Gini Coefficient* [J], *The Review of Economics and Statistics*, 1 (61), 146~149.
- [4] DiNardo J., Nicole M., Thomas L., 1996, *Labor Market Institutions and the Distribution of Wages, 1973~1992: A Semiparametric Approach* [J], *Econometrica*, 64, 1001~1044.
- [5] Kendall, Sir Maurice, Alan Stuart, 1977, *The Advanced Theory of Statistics* [M], Charles Grif-

fin.

[6] Mandelbrot B., 1960, *The Pareto-Levy Law and the Distribution of Income* [J], *International Economic Review*, 2, 79~106.

[7] McDonald J. B., Bartell C., 1979, *An Analysis of Some Properties of Alternative Measures of Income Inequality Based on the Gamma Distribution* [J], *Journal of the American Statistical Association*, 368 (74), 856~860.

[8] Rao D. S., Prasada, Kam Ki Tang, 2006, *Estimating Income Inequality in China Using Grouped Data and the Generalized Beta Distribution* [R], Research Paper, No. 2006/134.

[9] Salem A. B. Z., Mount T. D., 1974, *A Convenient Descriptive Model of Income Distribution* [J], *Econometrica*, 42, 1115~1127.

[10] Schmittlein D. C., 1983, *Some Sampling Properties of a Model for Income Distribution* [J], *Journal of Business & Economic Statistics*, 2 (1), 147~153.

[11] Singh S. K., Maddala G. S., 1976, *A Function for Size Distribution of Incomes* [J], *Econometrica*, 44, 963~970.

[12] Thomas Vinod, Yan Wang, Xibo Fan, 2000, *Measuring Education Inequality: Gini Coefficients of Education* [R], World Bank Policy Research Working Paper, No. WPS2525, Washington D. C., World Bank.

[13] 陈昌兵:《各地区居民收入基尼系数计算及其非参数计量模型分析》[J],《数量经济技术经济研究》2007年第1期。

[14] 程永宏:《改革以来全国总体基尼系数的演变及其城乡分解》[J],《中国社会科学》2007年第4期。

[15] 程永宏:《基尼系数组群分解新方法研究:从城乡二亚组到多亚组》[J],《经济研究》2008年第8期。

[16] 迟巍、黎波、余秋梅:《基于收入分布的收入差距扩大成因的分解》[J],《数量经济技术经济研究》2008年第9期。

[17] 段景辉、陈建宝:《基于家庭收入分布的地区基尼系数的测算及其城乡分解》[J],《世界经济》2010年第1期。

[18] 胡志军:《基于分组数据的基尼系数估计与社会福利》[J],《数量经济技术经济研究》2012年第9期。

[19] 胡祖光:《基尼系数理论最佳值及其简易计算公式研究》[J],《经济研究》2004年第9期。

[20] 金成武:《历算分布收入数据基尼系数的矩阵向量形式及相关问题》[J],《经济研究》2007年第4期。

[21] 黎波、迟巍、余秋梅:《一种新的收入差距研究的计量方法——基于分布函数的半参数估计》[J],《数量经济技术经济研究》2007年第7期。

[22] 廖理、张金宝:《城市居民家庭的经济条件、理财意识和投融资行为》[J],《经济研究》2010年第S1期。

[23] 廖理、张金宝:《中国消费金融调研报告》[M],经济科学出版社,2011。

[24] 田卫民:《省域居民收入基尼系数测算及其变动趋势分析》[J],《经济科学》2012年第2期。

[25] 王海港、周开国:《中国城乡居民收入分配的不平等程度被低估了吗?——基于帕雷托分布的一个检验》[J],《统计研究》2006年第4期。

[26] 王亚峰:《中国1985~2009年城乡居民收入分布估计》[J],《数量经济技术经济研究》2012年第6期。

(责任编辑:陈星星)