第 30 卷第 1 期
2013 年 1 月
统计研究
Statistical Research
Vol. 30, No. 1
Jan. 2013

基尼系数的区间估计及其应用

戴平生

内容提要:本文给出了收入为离散分布的三种计算基尼系数的新方法。利用收入份额法导出了基尼系数协方差算法的离散形式,并因此产生了计算基尼系数的回归系数法。文章重点讨论了对基尼系数进行区间估计的两种方法,这些方法也适用于集中度指数,因而它们在测度社会经济领域的不平等中具有着十分广泛的用途。实际应用表明,新算法有效地简化了对基尼系数区间估计的标准差估算。

关键词:基尼系数;区间估计;收入份额法;回归系数法;抽样分布法
中图分类号:C812　　文献标识码:A　　文章编号:1002 - 4565(2013)00 - 0083 - 07

Interval Estimation of Gini Index and Its Application

Dai Pingsheng

Abstract: This paper promotes several new approaches for computing Gini index based on discrete distribution. The discrete form of covariance that computes Gini index is derived from the income share method, and it leads to the regressive coefficient method, whist sample estimation of Gini index and statistical inference is discussed in two ways. The new approaches can also be applied to analyze the concentration index; they will have a wide use to measure inequality in the socioeconomic field. It is confirmed that the new approaches are effective to simply estimating standard deviation in internal estimation of Gini index.

Key words: Gini Index; Interval Estimation; Income Share method; Regressive Coefficient Method; Sample Distribution Method

一、引言

基尼系数是测度收入不平等最常用的指标之一。通常研究对象构成了一个离散的总体,即个体收入为离散型随机变量,而统计结果一般以样本或组数据形式出现,使得研究者不得不用样本数据估算总体收入的基尼系数。拟合收入分布函数估算总体基尼系数是一种十分常用的办法(胡志军,2012)[1],在总体容量很大的情况下该方法可以得到基尼系数的一个很好估计,与基尼系数真值具有强相合性(陈家鼎、陈奇志,2011)[2],但对应的洛伦兹曲线与离散数据的折线相比,在大多数情况下会出现对基尼系数的高估(王亚峰,2012)[3]。同时,要对基尼系数进行区间估计,本身也需要直接利用总体的抽样数据估计方差。因此利用拟合收入分布以外的其他方法计算总体的基尼系数(徐宽,2003)[4],是本文讨论的基本出发点。另一个关注点是总体基尼系数的区间估计问题,虽然陈希孺(2004)给出了总体基尼系数区间估计的基本思想以及方差估计的计算公式[5],陈家鼎、陈奇志(2011)证明了样本基尼系数与总体基尼系数差值收敛于正态分布以及方差估计的强相合性,但由于方差估计的计算公式十分复杂,常常使人望而却步。于是研究者给出了一些变通的方法,如王春雷、黄素心(2007)给出了基尼系数的上限和下限计算公式[6],陈家鼎等(2012)也给出了一个推理更为严密的基尼系数下限值计算公式[7]。通过基尼系数新的计算公式,本文给出了估计方差的两种方法,它们可以方便地进行方差的计算处理。

二、收入份额法

在基尼系数的各种计算方法中洛伦兹曲线的几何算法、基尼的平均差法可以认为是基尼系数的本源,其他算法都是派生的。但在不同的数据条件下

各种派生的算法各具优势，可以极大地简化人们的计算。对于离散分布的收入数据尤其如此，几何算法、基尼的平均差法都显得过于繁杂，最终常常要转化为代数算式。

1. 基尼系数的代数算式。

设 $y_1, y_2, \cdots, y_n$ 依次为第 $1, 2, \cdots, n$ 组人群的平均收入（不妨假定已按递增排列），$q_1, q_2, \cdots, q_n$ 为相应的人口数，q 为总人口数；记 $f_i = q_i/q$ 表示第 i 组的人口占总人口的比例，简称为人口份额；记 $F_0 = 0$，F_i 称为至第 i 组的累计人口份额（$i = 1, 2, \cdots, n$），它实际上就是收入的分布函数，显然有 $F_n = 1$。几何算法和基尼的平均差都可以转化为以下算式

$$G = \sum_{i=1}^{n} (L_i F_{i-1} - F_i L_{i-1})$$

$$L_i = \frac{q_1 y_1 + \cdots + q_i y_i}{q_1 y_1 + q_2 y_2 + \cdots + q_n y_n} = \frac{f_1 y_1 + \cdots + f_i y_i}{f_1 y_1 + f_2 y_2 + \cdots + f_n y_n}$$

$$L_0 = 0 \qquad (1)$$

其中 L_i 为至第 i 组的累计收入份额（$i = 1, 2, \cdots, n$），显然有 $L_n = 1$。这样基尼系数的计算就转化为累计人口份额和累计收入份额两组数据的代数运算，只要先计算 $\{F_i, i = 1, 2, \cdots, n\}$ 和 $\{L_i, i = 1, 2, \cdots, n\}$ 两组数由式（1）就可以算出基尼系数。

2. 基尼系数的收入份额法。

根据以下恒等式

$$\sum_{i=1}^{n} (L_i - L_{i-1})(F_i + F_{i-1} - 1) = \sum_{i=1}^{n} (L_i F_{i-1} - F_i L_{i-1})$$

可以得到基尼系数的一个等价定义

$$G = \sum_{i=1}^{n} \frac{q_i y_i}{S} \omega_i,\ S = q_1 y_1 + \cdots + q_n y_n,$$

$$\omega_i = F_i + F_{i-1} - 1,$$

$$F_i = f_1 + \cdots + f_i, (i = 1, \cdots, n) \qquad (2)$$

由于 $L_i - L_{i-1}$ 就是第 i 组的收入份额（$i = 1, 2, \cdots, n$），因此式（2）就是用收入份额的线性组合来计算基尼系数，我们把它称为基尼系数的收入份额法。组合系数等于当前累计人口份额加上前一项累计人口份额再减 1（即 $F_i + F_{i-1} - 1$），因而该组合系数都是纯小数。

组合系数 $\omega_i (i = 1, 2, \cdots, n)$ 具有以下性质：$f_1\omega_1 + f_2\omega_2 + \cdots + f_n\omega_n = 0$。

证明：由组合系数满足以下等式

$$\omega_i = 1 + \frac{(1 - F_i)^2 - (1 - F_{i-1})^2}{f_i}, (i = 1, \cdots, n)$$

容易验证该性质成立。由于收入分布 F_i 是 i 的增函数，还可以导出 $\omega_i (i = 1, 2, \cdots, n)$ 具有单调性、有界性等性质。

3. 收入份额法的特征。

收入份额法将基尼系数用各组收入份额的线性形式表出，方便基尼系数的组群分解和要素分解。

（1）基尼系数的组群分解。设组数据分为 r 个组群，满足 $n_1 + n_2 + \cdots + n_r = n$，记 $N = \{1, 2, \cdots, n\}$，N_k 为 N 的 r 个真子集（$k = 1, 2, \cdots, r$）。我们有

$$G = \sum_{i=1}^{n} \frac{q_i y_i}{S} \omega_i = \sum_{k=1}^{r} \sum_{i \in N_k} \frac{q_i y_i}{S} \omega_i = \sum_{k=1}^{r} S(k)$$

$$= \sum_{k=1}^{r} \frac{S_k}{S} \sum_{i \in N_k} \frac{q_i y_i}{S_k} (\omega_i^k + \omega_i - \omega_i^k)$$

$$= \sum_{k=1}^{r} \frac{S_k}{S} G_k + \sum_{k=1}^{r} \sum_{i \in N_k} \frac{q_i y_i}{S} (\omega_i - \omega_i^k) \qquad (3)$$

其中 $S(k)$ 表示将来自第 k 个组群的项合并，反映该组群对总收入基尼系数的贡献；S_k 表示第 k 个组群的全部收入，它与总收入 S 的比值就是第 k 个组群的收入份额；ω^k 表示第 k 个组群按群内组平均收入递增排序产生的组合系数，它与总体的组合系数 ω 完全不同；G_k 表示第 k 个组群的基尼系数（$k = 1, 2, \cdots, r$）。式（3）表明，基尼系数等于各组群基尼系数收入份额的加权平均，再加上一个因组合系数即排序差异产生的调整项。

（2）基尼系数的要素分解。设收入 y 有 r 个不同的要素来源（$y^1, y^2, \cdots, y^r$），满足 $y = y^1 + y^2 + \cdots + y^r$。于是

$$G = \sum_{i=1}^{n} \frac{q_i y_i}{S} \omega_i = \sum_{k=1}^{r} \sum_{i=1}^{n} \frac{q_i y_i^k}{S} \omega_i = \sum_{k=1}^{r} S(k)$$

$$= \sum_{k=1}^{r} \frac{S_k}{S} \sum_{i=1}^{n} \frac{q_i y_i^k}{S_k} (\omega_i^k + \omega_i - \omega_i^k)$$

$$= \sum_{k=1}^{r} \frac{S_k}{S} G_k + \sum_{k=1}^{r} \sum_{i=1}^{n} \frac{q_i y_i}{S} (\omega_i - \omega_i^k) \qquad (4)$$

公式中各要素的相关符号与组群分解类似。

（3）基尼系数的边际效应分析。假定第 m 个组群收入增加固定比例 e，其他的组群保持不变，可以得到基尼系数的增量表达式

$$\Delta G = G' - G = \sum_{i \in N} \frac{q_i y'_i}{S'} \omega'_i - G$$

$$= \sum_{i \in N} \frac{q_i y'_i}{S'} \omega_i - G + \sum_{i \in N} \frac{q_i y'_i}{S'} (\omega'_i - \omega_i)$$

$$= \sum_{i\in N}\frac{q_i y_i}{S+eS_m}\omega_i + \sum_{i\in N_m}\frac{eq_i y_i}{S+eS_m}\omega_i - G + \sum_{i\in N}\frac{q_i y'_i}{S'}(\omega'_i-\omega_i)$$

$$= \frac{S}{S+eS_m}G + \frac{es(m)}{S+eS_m}G - G + \sum_{i\in N}\frac{q_i y'_i}{S'}(\omega'_i-\omega_i)$$

于是得到以下公式

$$\Delta G = \frac{eG}{1+eS_m/S}\left(s(m)-\frac{S_m}{S}\right) + \sum_{i=1}^{n}\frac{q_i y'_i}{S'}(\omega'_i-\omega_i) \quad (5)$$

其中,$s(m)=S(m)/G$ 表示第 i 个组群对总收入基尼系数的贡献率,S、S'分别表示第 m 个组群增长 e 前后的全体收入总量,ω、ω'则分别表示增长前后通过排序产生的组合系数。通常 e 变化较小,式(5)右边因 ω、ω'差异产生的第二部分可以忽略不计(对于收入离散数据而言,是可以找到增长前后排序一致的 e),因此当第 m 个组群增长 e 时基尼系数增量 ΔG 的符号取决于 $s(m)$ 是否大于 S_m/S。即当第 m 个组群贡献率大于其收入份额时基尼系数变大,收入差距增大;相反则基尼系数变小,收入差距减小。上述推导对于基尼系数的要素分解也是成立的,Stark 等(1986)对个体数据的要素分解提出基尼系数相对边际效应的概念[8],现在也适用于对组数据的分析。

(4)测度税收累进性的 K 指数。Kakwani(1977)测度税收累进性的定义[9],恰好与税收关于总收入基尼系数边际效应成比例。设组数据收入模型为 $X=Y+T$,其中 T 为各组平均税收,X、Y 分别为各组总收入和可支配收入。他把税收累进性测度定义为 T 关于 X 递增排序的集中度指数与 X 基尼系数的差值,即 K 指数定义 $K=C_T-G_x$,根据式(2)和式(4)有

$$G_x = \sum_{i=1}^{n}\frac{q_i x_i}{S_x}\omega_i^x = \sum_{i=1}^{n}\frac{q_i y_i}{S_x}\omega_i^x + \sum_{i=1}^{n}\frac{q_i T_i}{S_x}\omega_i^x,$$

$$C_T = \sum_{i=1}^{n}\frac{q_i T_i}{S_T}\omega_i^x = \frac{S_x}{S_T}s(T)\,G_x$$

$$K = C_T - G_x = \frac{S_x}{S_T}s(T)\,G_x - G_x$$

$$= \frac{S_x}{S_T}G_x\left(s(T)-\frac{S_T}{S_X}\right) \quad (6)$$

因此,T 关于总收入基尼系数的边际效应与 K 指数具有相同的符号。如果将 T 的边际效应定义为税收的累进性测度,也是一个不错的选择,容易证明这样的测度对不同税种的累进性满足可加性。

(5)收入份额法的适用性。基尼系数对应于洛伦兹曲线,用 F_i、L_i 分别表示累计人口份额和累计收入份额,洛伦兹曲线就是由点(F_i, L_i) $(i=12,\cdots,n)$ 构成的过单位正方形两个顶点(0,0)和(1,1)的曲线。基尼系数被广泛应用于不平等测度,但横向坐标的收入分布性质虽然赋予了基尼系数的非负性特点,也导致了内在识别能力的弱化。如对以下三种收入分布无法区分:甲(40,20,40)、乙(20,40,40)和丙(40,40,20)三种收入分布按递增排序,收入分配的基尼系数都相同,等于 0.1333。

按其他的社会经济属性排序,如就甲乙丙依次排序,同样利用累计人口份额、累计收入份额构造曲线(它类似洛伦兹曲线,只是没有按收入递增排序),用相同的方法计算面积,那么三种收入分配的不平等指数就变成了 0,0.1333 和 -0.1333。这种不按自身属性排序计算的不平等指数称为集中度指数,记为 CI。集中度指数是基尼系数的第一种变化,K 指数就是税收集中度指数与总收入基尼系数的差。CI 对应的曲线其横坐标仍为累计人口份额,但已不是按自身的均值大小排序的累计人口份额即不是自己的分布函数。由于集中度指数 CI 的几何算法与基尼系数 G 的几何算法一致,因此收入份额法也适用于集中度指数。

基尼系数的第二种变化是洛伦兹曲线横坐标中的累计人口份额被其他变量的累计份额所代替,如税收累进性测度的 S 指数。S 指数是由 Suits(1977)提出的[10],它对应的曲线是在按收入递增排序后以累计收入份额为横坐标、累计税收份额为纵坐标。当税收与收入成比例时,累计税收份额等于累计收入份额,曲线与对角线重合;如果累计税收份额低于累计收入份额,曲线在对角线下方,那么这种税收不平等有利于低收入者,税收是累进的;若曲线出现在对角线上方,税收就是累退的。这类指数称为广义集中度指数[11]。假定曲线坐标为(L_i,P_i) $(i=1,2,\cdots,n)$,其中 L_i 为某一社会经济指标 y 的累计份额,P_i 为另一社会经济指标 x 的累计份额,例如对于 S 指数,x 和 y 分别表示税收和收入。那么不平等指数 H 对应的收入份额法公式为

$$H = \sum_{i=1}^{n}\frac{q_i x_i}{S_x}(L_i + L_{i-1} - 1)$$

$$S_x = q_1x_1 + \cdots + q_nx_n$$

$$P_i = \frac{q_1x_1 + \cdots + q_ix_i}{q_1x_1 + \cdots + q_nx_n} \quad (7)$$

这里 $P_i(i=1,2,\cdots,n)$ 为曲线的纵坐标，根据面积公式我们还有

$$H = 1 - \sum_{i=1}^{n} \frac{q_iy_i}{S_y}(P_i + P_{i-1})$$

$$= \sum_{i=1}^{n} \frac{q_iy_i}{S_y}(1 - P_i - P_{i-1})$$

$$S_y = q_1y_1 + \cdots + q_ny_n \quad (8)$$

式(7)和式(8)的组合系数都是单调有界的$(-1,1)$，但前者递增、后者是递减的。因此用收入份额法计算集中度指数、广义集中度指数都具有很强的适用性，通常将以上两种指数都称为集中度指数。

三、协方差法

利用个体数据的协方差计算基尼系数较早已得到解决(Anand，1983)[12]，后来人们发现在连续收入分布的情况下也有同样结果(Lerman 和 Yitzhaki，1984)[13]。但检索已有文献，这一算法至今仍然还无法在组数据中实现。由于收入份额法的出现，现在可以给出基尼系数协方差算法的离散形式。

1. 组数据基尼系数的协方差公式。

由式(2)我们有

$$G = \sum_{i=1}^{n} \frac{q_iy_i}{q\bar{y}}\omega_i = \sum_{i=1}^{n} \frac{q_i(y_i - \bar{y})}{q\bar{y}}(F_i + F_{i-1} - 1)$$

$$= \frac{2}{\bar{y}}\sum_{i=1}^{n} f_i(y_i - \bar{y})\left(\sum_{k=1}^{i-1} f_k + \frac{f_i}{2} - \frac{1}{2}\right)$$

$$= \frac{2}{\bar{y}}\sum_{i=1}^{n} f_i(y_i - \bar{y})\left(R_i - \frac{1}{2}\right) = \frac{2}{\bar{y}}\mathrm{Cov}(y_i, R_i)$$

$$R_i = \sum_{k=1}^{i-1} f_k + \frac{f_i}{2} = F_i - \frac{f_i}{2} \quad (9)$$

式(9)等号右边利用了组合系数 ω_i 的性质，R_i 表示累计到第 i 组中点的人口份额。R_i 的样本均值等于 1/2 $(i=1,2,\cdots,n)$，是因为

$$\omega_i = 2R_i - 1 \Rightarrow f_i\omega_i = 2f_iR_i - f_i$$

$$\Rightarrow \sum_{i=1}^{n} f_i\omega_i = 2\sum_{i=1}^{n} f_iR_i - \sum_{i=1}^{n} f_i$$

$$\Rightarrow 2\sum_{i=1}^{n} f_iR_i - 1 = 0 \Rightarrow \sum_{i=1}^{n} f_iR_i = \frac{1}{2}$$

当组数据退化为个体数据即每组只有 1 人时，式(9)可以简化为

$$G = \frac{2}{\bar{y}}\mathrm{Cov}(y_i, R_i) = \frac{2}{\bar{y}}\mathrm{Cov}\left(y_i, \frac{i}{n} - \frac{1}{2n}\right)$$

$$= \frac{2}{n\bar{y}}\mathrm{Cov}(y_i, i) \quad (10)$$

式(10)为个体数据基尼系数的协方差公式(Anand，1983)。下面证明式(9)就是连续收入分布基尼系数协方差的离散形式。假定收入区间无限细分，就有 R_i 收敛于 F_i，于是 R_i 就收敛于 y 的分布函数 $F(y)$，即

$$G = \frac{2}{\bar{y}}\mathrm{Cov}(y_i, R_i) \xrightarrow[R\to F(y)]{\bar{y}\to\mu} G = \frac{2}{\mu}\mathrm{Cov}(y, F(y)) \quad (11)$$

而式(11)是连续收入分布基尼系数的协方差形式(Lerman 和 Yitzhaki，1984；Lambert，1989)[14]。因此 R_i 就相当于 y 分布函数 $F(y)$ 的离散化，式(9)就是对应于式(11)连续分布协方差的离散化形式。只是 R_i 与 $F_i(i=1,2,\cdots,n)$ 略有不同，这就是为什么不能从连续收入分布基尼系数的协方差形式导出离散化结果的原因。

2. 组数据基于协方差的基尼系数要素分解。

设收入 y 有 r 个不同的要素$(y^1, y^2, \cdots, y^r)$，满足 $y = y^1 + y^2 + \cdots + y^r$。于是根据协方差的性质

$$G = \frac{2}{\bar{y}}\mathrm{Cov}(y_i, R_i) = \frac{2}{\bar{y}}\mathrm{Cov}\left(\sum_{k=1}^{r} y_i^k, R_i\right)$$

$$= \frac{2}{\bar{y}}\sum_{k=1}^{r}\mathrm{Cov}(y_i^k, R_i) = \sum_{k=1}^{r}\alpha_kG_k, \alpha_k$$

$$= \frac{\bar{y}_r}{\bar{y}}\frac{\mathrm{Cov}(y_i^k, R_i)}{\mathrm{Cov}(y_i^k, R_i^r)}, k = 1,2,\cdots,r \quad (12)$$

其中，R^k 表示按第 k 个要素均值递增排序由人口份额产生的近似累积分布(末位组仅频率的一半)，G_k 表示第 k 个要素的基尼系数$(k=1,2,\cdots,r)$。总收入基尼系数 G 也可以直接分解为各要素的集中度指数之和，集中度指数算法与基尼系数相似，只是公式中累计人口份额的排序依据不是按自身要素递增，仍沿用相应总收入均值递增。显然，组数据基尼系数的协方差公式(9)也适用于集中度指数，Kakwani 等(1997)曾给出了集中度指数的组数据协方差定义，但他们既没有给出证明也没有进行必要的说明[15]。

3. 基于协方差要素分解的边际效应分析。

利用基尼系数的协方差形式进行要素分解在本文之前的研究中是最为常见的做法。由式(12)基尼系数的要素分解可以发现，当第 m 个要素发生细微变化，如增长一个固定比例 d，而其他要素不变

时,在不改变收入分布的假定下(对于离散数据只要 d 小到一定程度,这样的假定是可以成立的),可以计算基尼系数的增量

$$G = \frac{2}{\bar{y}}\mathrm{Cov}(y_i, R_i) = \frac{2q}{S}\mathrm{Cov}\left(\sum_{k=1}^{r} y_i^k, R_i\right)$$

$$G' = \frac{2q}{S'}\mathrm{Cov}\left(\sum_{k=1}^{r} y_i^k + dy_i^m, R_i\right)$$

$$= \frac{2q}{S'}\mathrm{Cov}\left(\sum_{k=1}^{r} y_i^k, R_i\right) + \frac{dq}{S'}\mathrm{Cov}(y_i^m, R_i)$$

$$\Rightarrow G' - G = \frac{S}{S'}G + ds(m)\,G - G$$

$$= ds(m)\,G - \frac{dS_m}{S'}G \approx dG\left(s(m) - \frac{S_m}{S}\right)$$

等式中的符号与前面一致。由于 d 的取值很小,S' 与 S 相差无几,因此基尼系数的变化符号就由等式右边括号内符号决定,$s(m) - S_m/S$ 即为相对边际效应。然而,要进行基尼系数组群分解的边际效应分析就显得无能为力了。

四、回归系数法

连续收入分布基尼系数的协方差算式表明,基尼系数可以表为收入变量与收入分布变量的相关系数。因此由式(9)连续收入分布的离散化处理,我们可以尝试将收入 y_i 关于 R_i 进行线性回归($i = 1, 2, \cdots, n$),但该式也表明直接使用普通最小二乘法显然无法达到目的。

1. 基尼系数的回归方程模型。

下面利用待定系数法结合加权最小二乘法建立回归方程模型

$$ky_i t_i = \alpha t_i + \beta R_i t_i + \varepsilon_i, (i = 1, 2, \cdots, n)$$

其中 k 为待定系数,这里的 ε_i、t_i 分别为误差项和加权函数(取 $t_i^2 = f_i, i = 1, 2, \cdots, n$)。

让误差平方和最小,通过极值求解可以得到截距、斜率参数满足的方程

$$\hat{\alpha} = k\bar{y} - \hat{\beta}\bar{R},$$

$$\hat{\beta} = \frac{k\mathrm{Cov}(y_i, R_i)}{\sigma_R^2} = \frac{k\bar{y}G}{2\sigma_R^2} \Rightarrow k = \frac{2\sigma_R^2}{\bar{y}},$$

$$\hat{\beta} = G, \hat{\alpha} = 2\sigma_R^2 - \bar{R}G$$

因此设定 k 对 ky_i 关于 R_i 的回归方程使用加权最小二乘法(WLS)估计参数 β,也可以得到基尼系数的计算值。使用的回归方程模型为

$$\frac{2\sigma_R^2}{\bar{y}}y_i\sqrt{f_i} = \alpha\sqrt{f_i} + \beta R_i\sqrt{f_i} + \varepsilon_i, (i = 1, 2, \cdots, n) \quad (13)$$

由于回归方程的引入使得我们可以进行参数估计的统计检验,并进一步对基尼系数进行区间估计。

2. 基尼系数的区间估计。

(1)加权回归法。由于在数据处理过程中收入 y 按递增排序,R 也是递增变化的,这样收入 y 对变量 R 直接进行回归可能导致误差项的异方差现象。采用加权处理后消除了异方差,在一般情况下统计软件在估计参数 β 的同时还会给出估计值的标准差和 t 检验值,于是可以估计基尼系数的方差,得出基尼系数在不同置信度下的区间估计。

(2)抽样分布法。对于总体抽样形成的组数据,直接利用式(2)计算基尼系数可能低估总体的基尼系数。因为组平均数据产生的洛伦兹折线必然位于个体收入数据洛伦兹折线的上方,根据它与对角线围成面积计算的基尼系数自然变小,因此给出基尼系数的区间估计显得十分必要。由收入份额法计算公式以及组数据的相关信息,可以对方差进行估算。设收入总体人口数为 q,对应于第 i 个组数据的人口数为 $q_i (i = 1, 2, \cdots, n)$;记 y_i、f_i 分别为总体第 i 组的收入水平和频率,设它们的样本估计值满足

$$\hat{y}_i = y_i + \eta_i, \hat{f}_i = f_i + \delta_i (i = 1, 2, \cdots, n)$$

其中 η_i 服从正态分布 $N(0, \sigma_i^2)$ $(i = 1, 2, \cdots, n)$,它们可能存在异方差且相互独立;对于 δ_i 的分布可以给出数字特征。第 i 组的样本频率可以看作 q 个概率等于 f_i 的两点分布 $x_1^i, \cdots, x_q^i$ 的平均,即

$$\hat{f}_i = \frac{x_1^i + \cdots + x_q^i}{q} \Rightarrow E(\hat{f}_i) = f_i,$$

$$\mathrm{Var}(\hat{f}_i) = \frac{f_i(1 - f_i)}{q} \quad (i = 1, 2, \cdots, n)$$

$$\mathrm{Cov}(\hat{f}_i, \hat{f}_j) = E\left(\frac{x_1^i + \cdots + x_q^i}{q} \times \frac{x_1^j + \cdots + x_q^j}{q}\right) - f_i f_j$$

$$= -\frac{f_i f_j}{q} \quad (i \neq j)$$

记 y_i 和 f_i、η_i 和 δ_i 对应的随机向量为 y 和 f、η 和 δ,那么 δ 的方差矩阵可以表示为 $q\mathrm{Var}(\delta) = \mathrm{diag}(f) - ff'$,这里用 $\mathrm{diag}(f)$ 表示由 f 分量构成的对角矩阵;而 η 的方差矩阵是一个以 $\sigma_i^2 (i = 1, 2, \cdots, n)$ 为元素的对角矩阵。用样本对基尼系数进行估计,由式(2)有

$$G = \sum_{i=1}^{n} \frac{q_i y_i}{S}\omega_i = \frac{f'\mathrm{diag}(\omega)\, y}{f'y}$$

$$= \frac{f'(Af - l)y}{f'y} \Rightarrow \hat{G} = \frac{\hat{f}'(A\hat{f} - l)\hat{y}}{\hat{f}'\hat{y}}$$

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 2 & \cdots & 1 \end{pmatrix}, \quad l = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

将基尼系数的估计式按 y 和 f 一阶 Taylor 展开，我们有

$$\hat{G} = \frac{\hat{f}'(A\hat{f} - l)\hat{y}}{\hat{f}'\hat{y}} \approx G + P'\eta + Q'\delta + o\left(\frac{1}{q}\right)$$

$$\Rightarrow \mathrm{Var}(\hat{G}) \approx P'\mathrm{Var}(\eta)P + Q'\mathrm{Var}(\delta)Q$$

这样就可以得到参数 β 的方差估计值。其中随机误差向量 η 和 δ 相互独立，P 和 Q 为系数向量，有等式

$$P = \frac{\mathrm{diag}(\omega - G)}{\bar{y}} f = \left(\frac{\omega_i - G}{\bar{y}} f_i\right)_{n\times 1}$$

$$Q = \frac{\mathrm{diag}(\omega - G) + A'\mathrm{diag}(f)}{\bar{y}} y$$

$$= \left(\frac{\omega_i - G}{\bar{y}} y_i + 2 - L_i - L_{i-1}\right)_{n\times 1}$$

其中 $\mathrm{diag}(\omega)$ 表示由 $\omega_1, \cdots, \omega_n$ 构成的对角矩阵。估计方差时只要把 G、f、y 和 ω 用样本对应的计算值代入即可。而 η 方差对角矩阵中 σ_i^2（$i = 1, 2, \cdots, n$）为样本各组收入数据的方差。在公式推导过程中仅使用了组数据的人口份额和收入份额，当组数据按收入递增排序时结果对应于基尼系数，当组数据按其他指标排序时结果就对应于集中度指数。

记系数向量 Q 为 $(Q_1, Q_2, \cdots, Q_n)'$，基尼系数的方差估计可以进一步简化为

$$\mathrm{Var}(\hat{G}) \approx \sum_{i=1}^{n} (\omega_i - G)^2 f_i^2 \frac{\sigma_i^2}{\bar{y}^2} + \frac{1}{q}\left[\sum_{i=1}^{n} Q_i^2 f_i - \left(\sum_{i=1}^{n} Q_i f_i\right)^2\right] \quad (14)$$

等式右边的第二部分可以看作是系数向量 Q 关于 f 分布列的方差。因此，基尼系数估计量方差的确定关键在两个系数向量，从式（14）可以看出它们的计算并不复杂。

五、应用实例

实证分析采用了中国健康与营养调查（CHNS）1989 – 2009 年的 8 次调查的汇总数据，该数据库由美国北卡罗纳大学与中国预防医学会等单位联合调查建立。从原数据库中剔除数据缺失、个体收入小于等于 0 的记录后，余下 51828 个记录供本文分析。通过在数据库软件 VFP6.0 中简单编程，对相关记录进行了初步处理。其中 1989 年的数据包含了 8 个省份 6146 个记录，表 1 给出了 8 个省份的样本平均收入、调查人数、省内收入基尼系数，以及计算省间基尼系数和区间估计所需要的数据：样本均值的方差（σ^2/n）、样本频率（f）和近似收入分布的 R 值。

表 1 1989 年调查省份居民收入水平及相关数据 （元）

省份	平均收入	样本数	基尼系数	均值方差	样本频率	近似分布
湖北	1150.36	767	0.3768	1359.206	0.1248	0.0624
辽宁	1232.28	836	0.4068	2246.101	0.1360	0.1928
河南	1274.38	703	0.4914	3177.156	0.1144	0.3180
四川	1316.23	750	0.5792	73771.290	0.1220	0.4362
广西	1525.39	956	0.5578	10831.389	0.1555	0.5750
江苏	1543.20	728	0.3786	2371.858	0.1185	0.7120
山东	1570.58	650	0.4271	5902.175	0.1058	0.8241
湖南	1902.64	756	0.4902	8899.409	0.1230	0.9385
省间基尼系数 0.0861			回归标准差 0.0116		Taylor 标准差 0.0100	

利用表 1 中的数据通过软件 Eviews6.0 选择加权最小二乘法对方程式（13）的参数进行估计，获得了 1989 年 CHNS 调查的省间居民收入基尼系数，它与根据式（2）计算的结果完全一致，同时还得到了基尼系数估计的标准差（简称回归标准差）；再计算系数向量 Q，由式（14）可以算出按一阶 Taylor 展开的基尼系数标准差（简称 Taylor 标准差）。基尼系数的估计量，及其两种算法的标准差列入了表 1 的最后一行，可以发现两者是相当接近的。

表 2 2009 年调查省份居民收入水平及相关数据 （元）

省份	平均收入	样本数	基尼系数	均值方差	样本频率	近似分布
广西	12719.88	951	0.5128	547196.81	0.1426	0.0713
河南	13369.67	672	0.5383	484896.77	0.1008	0.1930
辽宁	16751.34	814	0.4371	348400.83	0.1221	0.3045
湖北	16887.64	706	0.4763	994259.33	0.1059	0.4185
黑龙江	16907.97	767	0.4456	614629.85	0.1150	0.5289
四川	17799.81	609	0.5452	1576401.73	0.0913	0.6321
山东	18114.72	714	0.4990	1519796.27	0.1071	0.7314
江苏	20920.99	797	0.4254	775147.08	0.1195	0.8447
湖南	21437.37	637	0.5203	2334835.21	0.0955	0.9522
省间基尼系数 0.0904			回归标准差 0.0101		Taylor 标准差 0.0110	

为了便于分析省内居民收入基尼系数的变化规律，表 2 给出了由 2009 年 CHNS 调查数据计算的居民平均收入以及相关指标的估算结果。通过对比分析，可以发现经过 20 年的时间，各调查省份居民收入水平有了很大的提高，尽管省间居民收入的基尼系数没有出现十分明显的变化，但省内基尼系数超过 0.5 的省份由 1989 年的四川、广西两省变化为

表 3　　CHNS 调查的省间居民收入基尼系数及其 95% 置信区间　　(单位:人、元)

年份	基尼系数	标准差 σ_1	标准差 σ_2	样本数量	样本收入均值	回归标准差 σ_1		Taylor 标准差 σ_2	
						左端	右端	左端	右端
1989	0.0861	0.0116	0.0101	6146	1437.78	0.0634	0.1089	0.0664	0.1059
1991	0.0795	0.0093	0.0066	6893	1460.31	0.0613	0.0977	0.0666	0.0924
1993	0.1223	0.0092	0.0073	6416	2129.49	0.1044	0.1403	0.1079	0.1367
1997	0.1165	0.0106	0.0067	6511	4547.31	0.0959	0.1372	0.1035	0.1296
2000	0.0974	0.0132	0.0080	6748	5629.19	0.0716	0.1231	0.0817	0.1130
2004	0.1329	0.0126	0.0080	6399	7507.69	0.1082	0.1575	0.1171	0.1486
2006	0.0953	0.0086	0.0097	6048	10427.59	0.0784	0.1121	0.0763	0.1143
2009	0.0904	0.0101	0.0110	6667	17055.85	0.0707	0.1102	0.0688	0.1121

注:相对于 1989 年 CHNS 调查的省份,黑龙江于 1997 年替代了辽宁省,在随后若干年中两个省份并存。

2009 年的四川、河南、湖南、广西四省;1989 年省内基尼系数低于 0.4 的湖北、江苏也于 2009 年超出了 0.4,而辽宁、山东的省内基尼系数也有所上升。也就是说,虽然省内基尼系数最高的四川、广西有所回调,但其余各省基尼系数都有不同程度的上升,后面新增的 CHNS 调查省份黑龙江 2009 年的基尼系数也接近 0.45,因此总体上可以认为居民的收入分配差距扩大,且处于较不公平状态。再从基尼系数区间估计的标准差来看,两种对标准差的估计方法并不存在一种估计方法的结果大于另一种的现象。

为了进一步了解 CHNS 调查省份间基尼系数的动态变化,表 3 还给出了其他 6 个年度的基尼系数的计算结果,同时利用本文提出的两种区间估计方法估算了标准差及基尼系数 95% 的置信区间。从表 3 可以发现,自 1989 年以来的 20 年间居民收入水平总体上处于上升通道,省间基尼系数先是上升,2004 年达到最高值,随后出现了明显的下降。对比两种基尼系数区间估计方法对标准差估算的结果,可以发现随着居民收入水平的提高,回归标准差由前面若干年的略高于 Taylor 标准差,变化为近年来的略低于 Taylor 标准差。

应用实例表明,本文提出的基尼系数区间估计的加权回归法和抽样分布法能够有效地简化标准差的估算问题。前者可以借助软件 Eviews6.0 完成,后者则只须进行并不复杂的代数运算,从而克服了传统算法中对标准差的繁杂计算。

参考文献

[1] 胡志军. 基于分组数据的基尼系数估计与社会福利:1985 – 2009 [J]. 数量经济技术经济研究,2012(9):111 – 121.

[2] 陈家鼎,陈奇志. 关于洛伦兹曲线和基尼系数的统计推断 [J]. 应用数学学报,2011(3):385 – 399.

[3] 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.

[4] 徐宽. 基尼系数的研究文献在过去八十年是如何拓展的 [J]. 经济学(季刊),2003(4):757 – 778.

[5] 陈希孺. 基尼系数及其估计 [J]. 统计研究,2004(8):58 – 60.

[6] 王春雷,黄素心. 基尼系数与样本信息含量 [J]. 数量经济技术经济研究,2007(2):136 – 144.

[7] 陈家鼎,房祥忠,时丕旭,等. 混合总体基尼系数的下限——兼论我国城乡合在一起时基尼系数的计算 [J]. 应用概率统计,2012(4):367 – 389.

[8] Stark O, Taylor J E, Yitzaki S. Remittances and inequality [J]. The Economic Journal, 1986,383:722 – 740.

[9] Kakwani N C. Measurement of Tax Progressivity: An International Comparison [J]. The Economic Journal, 1977, 87(345):71 – 80.

[10] Suits D B. Measurement of Tax Progressivity [J]. The American Economic Review, 1977, 67(4):747 – 752.

[11] Wagsraff A, Paci P, Doorslaer E V. On the Measurement of Inequalities in Health [J]. Social Science and Medicine, 1991,33(5):545 – 557.

[12] Anand, Sudhir. Inequality and Poverty in Malaysia: Measurement and Decomposition [M]. New York: Oxford University Press, 1983.

[13] Lerman R I, Yitzhaki S. A Note on the Calculation and Interpretation of the Gini Index [J]. Economics Letters, 1984,15:363 – 368.

[14] Lambert P J. The Distribution and Redistribution of Income: A Mathematical Analysis [M]. Cambridge, Massachusetts: Basil Blackwell Inc, 1989.

[15] Kakwani N, Wagstaff A, van Doorslaer E. Socioeconomic Inequalities in Health: Measurement, Computation, and Statistical Inference [J]. Journal of Econometrics, 1997,77:87 – 103.

作者简介

戴平生,男,广东兴宁人,2004 年毕业于厦门大学统计系,获经济学博士学位,现为厦门大学经济学院副教授、硕士生导师,教育部计量经济学重点研究室(厦门大学)、福建省统计科学重点实验室(厦门大学)兼职研究员。研究方向为数量经济学、经济统计。

(责任编辑:程　晞)