应用概率统计　第二十八卷
第四期　2012年8月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.4 Aug. 2012

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

陈家鼎　房祥忠*　时丕旭　胡晓欧　曹　丽

(北京大学数学科学学院统计中心, 北京, 100871)

摘　要

本文研究基尼系数的估计问题. 国家统计局每年发布的统计年鉴包含城镇和农村个人收入的分组数据, 但分点不相同并且未公布. 这些情况对于估计整个社会的基尼系数带来挑战. 本文对于两个总体按照一定比例混合后的新总体, 针对来自原来两个总体的分组数据, 给出了新总体的基尼系数的下限. 并将所得的结果用于计算我国城乡合在一起时基尼系数的下限值. 这些结果容易推广到更多总体混合情形, 也可以应用到其它实际情况的基尼系数的估计, 比如国家或地区的联合体.

关键词: 基尼系数, 混合总体, 下限.

学科分类号: O212.

§1. 引　　言

如众所知, 基尼系数是衡量一个国家或地区人们贫富差距的重要指标, 也可以用于检验政府政策(特别是税收政策)在调节收入分配上的作用[1]. 在我国经济迅猛发展的今天, 人们的收入差距也在明显扩大. 如何科学地计算每年的基尼系数, 是我国政府和人民群众十分关心的问题, 也是当今人们的热门话题.

应该指出, 如果有了大量的原始调查数据, 计算基尼系数是不难的, 我们可用数学公式表明这一点.

用Y表示某国家或地区单个人的年收入, 将这个Y看成随机变量, 其分布函数是$F(x)$, 即$F(x)=\mathsf{P}(Y\le x)$ (表示随机变量Y取值不超过x的概率). 我们恒假设

$$Y\ge 0,\qquad 0<\mathsf{E}Y<\infty, \tag{1.1}$$

这里$\mathsf{E}Y$是Y的数学期望(均值), 以下用μ表示$\mathsf{E}Y$. 令

$$F^{-1}(u)=\inf\{x: F(x)>u\}\qquad (0<u<1), \tag{1.2}$$

$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(u)\mathrm{d}u\qquad (0\le p\le 1), \tag{1.3}$$

$$G=1-2\int_0^1 L(p)\mathrm{d}p, \tag{1.4}$$

*通讯作者, E-mail: xzfang@pku.edu.cn.

本文2011年5月6日收到, 2012年1月28日收到修改稿.

$F^{-1}(u)$是$F(x)$的广义反函数, 即u分位数, $L(p)$是著名的Lorenz曲线, G就是Y (或者分布函数$F(x)$)的基尼系数.

以上是Lorenz函数和基尼(Gini)系数的最一般定义[2]. 若对Y有大量的原始调查数据$y_1, y_2, \ldots, y_n$ (y_i是第i个人的年收入), 令

$$G_n = \frac{1}{2n^2\overline{y}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|, \tag{1.5}$$

其中$\overline{y}_n = (1/n) \sum_{i=1}^{n} y_i$. G_n是数组$\{y_1, \ldots, y_n\}$的"基尼系数", 它是C. Gini在1912年首先提出来的[3]. 可以证明, 只要$y_1, \ldots, y_n$是简单随机样本, 则$\mathsf{P}\big(\lim\limits_{n\to\infty} G_n = G\big) = 1$ (见[2]). 换句话说, G_n是G的强相合估计. [4]还在极广泛的条件下证明了G_n的渐近正态性, 从而给出了大样本情形下G的置信区间.

问题在于, 大多数国家和地区并不公布原始的调查数据, 而只是公布分组数据. 以中国国家统计局在《中国统计年鉴》上公布的2003~2009年的数据为例, 把每年的年收入调查数据从小到大分成了若干组(城镇为8组, 农村为5组), 只给出每组的人数(如城镇)或其占全体被调查总人数的比例(如农村)以及每组的人均年收入(参看下文). 针对这种分组数据, 如何给出基尼系数的合适估计?

这是一个从上世纪七十年代以来受到各国广泛关注的重要问题, 迄今已有多项研究成果, 尚未完全解决(参看[5]至[12]). 针对一个总体Y的分组数据的数学模型是这样的.

设

$$a_0 = 0 < a_1 < a_2 < \cdots < a_s < a_{s+1} = \infty,$$

$$p_i = \begin{cases} \mathsf{P}(Y \le a_1), & i = 1; \\ \mathsf{P}(a_{i-1} < Y \le a_i), & i = 2, \ldots, s+1, \end{cases}$$

$$\mu_i = \begin{cases} \mathsf{E}(Y | Y \le a_1), & i = 1; \\ \mathsf{E}(Y | a_{i-1} < Y \le a_i), & i = 2, \ldots, s+1. \end{cases}$$

这里$\mathsf{E}(Y|A)$表示事件A发生的条件下Y的条件期望. 第1组对应的收入区间是$[0, a_1]$, 第i组对应的收入区间是$(a_{i-1}, a_i]$ ($i \ge 2$), μ_i是第i组人均年收入.

由于被调查的人数很大, 依据频率与概率的关系以及大数定律, 我们可以认为这些p_i, μ_i是已知的, p_i就是第i组被调查人数占全体被调查人数的比例, μ_i就是第i组被调查的人均收入.

"分组数据"分为两大类型: 第一类是各收入分界点$a_1, a_2, \ldots, a_{s+1}$秘而不宣, 第二类是各收入分界点公开于众.

有了数据$\{p_i, \mu_i, i = 1, 2, \ldots, s+1\}$, 如何估计总体$Y$的基尼系数$G$ (见(1.4))?

这个问题的求解相当复杂. 已有的研究工作可分为两类. 一类是对总体的分布函数$F(x)$或者相应的Lorenz函数$L(p)$施加某种假定(例如VA模型, β模型等), 然后找出G的点估计; 另一类是不对$F(x)$或$L(p)$施加任何假定, 找出G的有意义的下限和上限. 这方面已有重要的结果: 最大的下限有明显的公式. 对第二类分组数据, [7]和[9]给出了优良的上限表达式; 对第一类分组数据, [8]提出了优良的上限(但是未给出上限的有效算法).

我们中国面临的问题更为复杂. 中国公布的"分组数据"是按照城镇和农村分别给出的, 这是基于城镇和农村分别进行的抽样调查而分别归纳出的. 我们要回答的问题是: 从这些分组数据出发, 如何计算我国城乡合在一起时的全国基尼系数?

这就是"分组数据"情形下混合总体的基尼系数的计算问题.

用Y_1表示城镇个人的年收入(可支配收入), 其分布函数是$F_1(x)$; 用Y_2表示农村个人的年收入(纯收入), 其分布函数是$F_2(x)$. 用Y表示我国任何个人的年收入, 其分布函数为$F(x)$. 易知

$$F(x)=\alpha_1F_1(x)+\alpha_2F_2(x), \tag{1.6}$$

其中α_1是城镇人口在全国人口中所占的比例, $0<\alpha_2=1-\alpha_1<1$. 通过人口普查或调查可以得到α_1,α_2的值或近似值.

Y_1和Y_2的分布函数$F_1(x)$和$F_2(x)$是未知的, 分别有分组数据: Y_1的数据从小到大分成了s_1+1组, 第j组数据个数占Y_1的全部数据个数的比例为$p_i^{(1)}$, 第i组数据的均值为$\mu_i^{(1)}$; Y_2的数据从小到大分成了s_2+1组, 第i组数据个数占Y_2的全部数据个数的比例为$p_i^{(2)}$, 第i组数据的均值为$\mu_i^{(2)}$.

有了$\{p_i^{(k)},\mu_i^{(k)},i=1,2,\dots,s_k+1,\ k=1,2\}$和$\alpha_1,\alpha_2$, 如何计算混合总体$Y$的基尼系数$G$?

我国对此有大量研究, 都是在对分布函数$F_1(x)$和$F_2(x)$或其Lorenz函数施加某种假定之后给出基尼系数G的估计值. 这些"假定"的合理性常受到质疑(参看[10]至[13]).

本文的任务是, 在不对$F_1(x)$和$F_2(x)$施加任何假定的条件下, 基于分组数据给出混合后总体Y的基尼系数G的下限(下界). 并针对我国公布的2003~2009年数据进行了具体计算, 给出了全国基尼系数的下限. 我们得到的下限对于了解我国贫富差距程度是很有意义的.

§2. 主要定理及其证明

沿用§1中的记号, 设Y_1和$Y2$都是非负随机变量, 其分布函数分别是$F_1(x)$和$F_2(x)$ (例如Y_1是城镇居民的个人年可支配收入, Y_2是农村居民的个人年纯收入). 易知

$$F(x)=\alpha_1F_1(x)+\alpha_2F_2(x) \tag{2.1}$$

是混合总体的分布函数, α_1是第一个总体在混合总体中的比例, $\alpha_2\,(=1-\alpha_1)$是第二个总体所占比例. 例如$F(x)$是城乡合在一起时的个人收入分布函数, α_1是城镇人口占总人口的比例, α_2是农村人口占总人口的比例.

对Y_k, $k=1,2$假设

$$a_0^{(k)}=0<a_1^{(k)}<a_2^{(k)}<\cdots<a_{s_k}^{(k)}<a_{s_k+1}^{(k)}=\infty,$$

$$p_i^{(k)}=\begin{cases}\mathsf{P}(Y_k\le a_1^{(k)}), & i=1;\\ \mathsf{P}(a_{i-1}^{(k)}<Y_k\le a_i^{(k)}), & i=2,\dots,s_k+1,\end{cases}$$

$$\mu_i^{(k)}=\begin{cases}\mathsf{E}(Y_k|Y_k\le a_1^{(k)}), & i=1;\\ \mathsf{E}(Y_k|a_{i-1}^{(k)}<Y_k\le a_i^{(k)}), & i=2,\dots,s_k+1.\end{cases}$$

(以上规定: $\mathsf{P}(A)=0$时, $\mathsf{E}(Y_i|A)=0$.)

我们的主要结果是下面的定理.

定理 2.1 设G是式(2.1)给出的分布函数$F(x)$所对应的基尼系数, 则有不等式

$$G\ge\frac{1}{2\mu}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{i=1}^{s_k+1}\sum_{j=1}^{s_l+1}\alpha_k\alpha_l p_i^{(k)}p_j^{(l)}|\mu_i^{(k)}-\mu_j^{(l)}|, \tag{2.2}$$

这里μ是分布函数$F(x)$的数学期望, 满足$\mu=\alpha_1\mu^{(1)}+\alpha_2\mu^{(2)}$, $\mu^{(k)}=\sum\limits_{i=1}^{s_k+1}p_i^{(k)}\mu_i^{(k)}$是分布函数$F_k(x)$的数学期望, $k=1,2$.

注记 1 若记$s=s_1+s_2+2$,

$$p_i=\begin{cases}\alpha_1 p_i^{(1)}, & i=1,2,\dots,s_1+1;\\ \alpha_2 p_{i-(s_1+1)}^{(2)}, & i=s_1+2,s_1+3,\dots,s,\end{cases} \tag{2.3}$$

$$\mu_i=\begin{cases}\mu_i^{(1)}, & i=1,2,\dots,s_1+1;\\ \mu_{i-(s_1+1)}^{(2)}, & i=s_1+2,s_1+3,\dots,s,\end{cases} \tag{2.4}$$

则不等式(2.2)可以简写为

$$G\ge\frac{1}{2\mu}\sum_{i=1}^{s}\sum_{j=1}^{s}|\mu_i-\mu_j|p_ip_j, \tag{2.5}$$

其中$\mu=\sum\limits_{i=1}^{s}p_i\mu_i$.

不等式(2.2)(或(2.5))给出了混合总体的基尼系数的下限. 在实际应用中, 由于被调查的人数众多, 概率$p_i^{(k)}$ $(i=1,2,\dots,s_k+1,\ k=1,2)$均可用该组被调查的人数所占的比例作为近似值, 条件期望$\mu_i^{(k)}$ $(i=1,2,\dots,s_k+1,\ k=1,2)$均可用该组个人收入的均值作为近似值(参看下文的§3).

为了证明定理2.1, 需要两个引理.

引理 2.1 设$F(x)$是任何分布函数, 满足

$$F(0-)=0, \qquad 0<\mu=\int_0^\infty x\mathrm{d}F(x)<\infty,$$

则$F(x)$对应的基尼系数G有下列公式

$$G=1-\frac{1}{\mu}\int_0^\infty (1-F(x))^2\mathrm{d}x. \tag{2.6}$$

证明: 令$F^{-1}(u)=\inf\{x:F(x)>u\}$ $(0<u<1)$, 则$F(x)$对应的Lorenz函数为

$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(u)\mathrm{d}u, \qquad 0\le p\le 1,$$

从而

$$\begin{aligned}
G &= 1-2\int_0^1 L(p)\mathrm{d}p \qquad (\text{见 (1.4)})\\
&= 1-\frac{2}{\mu}\int_0^1\int_0^p F^{-1}(u)\mathrm{d}u\mathrm{d}p\\
&= 1-\frac{2}{\mu}\iint_{0\le u\le p\le 1} F^{-1}(u)\mathrm{d}u\mathrm{d}p\\
&= 1-\frac{2}{\mu}\int_0^1\Big(\int_u^1 F^{-1}(u)\mathrm{d}p\Big)\mathrm{d}u\\
&= 1-\frac{2}{\mu}\int_0^1 F^{-1}(u)(1-u)\mathrm{d}u.
\end{aligned} \tag{2.7}$$

另一方面,

$$\begin{aligned}
\frac{1}{2}\int_0^\infty (1-F(x))^2\mathrm{d}x &= \int_0^\infty\Big(\int_0^{1-F(x)} u\mathrm{d}u\Big)\mathrm{d}x\\
&= \int_0^1 u\Big(\int_{\substack{x\ge 0,\\ 1-F(x)\ge u}} 1\mathrm{d}x\Big)\mathrm{d}u\\
&= \int_0^1 u\Big(\int_{\substack{x:F(x)\le 1-u\\ x\ge 0}} 1\mathrm{d}x\Big)\mathrm{d}u.
\end{aligned}$$

令$A=\{x:x\ge 0, F(x)\le 1-u\}$, 则

$$[0,F^{-1}(1-u))\subset A\subset[0,F^{-1}(1-u)].$$

由此知

$$\frac{1}{2}\int_0^\infty (1-F(x))^2\mathrm{d}x=\int_0^1 uF^{-1}(1-u)\mathrm{d}u=\int_0^1(1-u)F^{-1}(u)\mathrm{d}u.$$

利用公式(2.7)知

$$G=1-\frac{1}{\mu}\int_0^\infty (1-F(x))^2\mathrm{d}x. \qquad \square$$

注记 2 对于$F(x)$有分布密度$p(x)$且有$M>0$, 使得$x>M$时$p(x)=0$的情形, 参考文献[3]中证明了公式(2.6)成立. 我们在最一般情形下证明了(2.6)式成立.

引理 2.2 设$y_1,y_2,\ldots,y_{n_1}$是$F_1(x)$的简单随机样本, $y_{n_1+1},\ldots,y_{n_1+n_2}$是$F_2(x)$的简单随机样本, 且$\{y_1,\ldots,y_{n_1}\}$与$\{y_{n_1+1},\ldots,y_{n_1+n_2}\}$相互独立. $n=n_1+n_2$, G_n是数组$\{y_1,\ldots,y_n\}$的基尼系数, 即

$$G_n=\frac{1}{2n^2\overline{y}_n}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i-y_j|, \tag{2.8}$$

则$n_1\to\infty$, $n_2\to\infty$且$n_1/(n_1+n_2)\to\alpha_1$时, 有

$$\mathsf{P}\Big(\lim_{n\to\infty}G_n=G\Big)=1, \tag{2.9}$$

这里$\overline{y}_n=(1/n)\sum\limits_{i=1}^{n}y_i$, G是混合分布$F(x)=\alpha_1F_1(x)+\alpha_2F_2(x)$对应的基尼系数.

证明: 令

$$\widehat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{(y_i\le x)},$$

这里$I_{(y_i\le x)}$是示性函数

$$I_{(y_i\le x)}=\begin{cases}1 & y_i\le x;\\ 0 & y_i>x.\end{cases}$$

易知

$$\widehat{F}_n(x)=\frac{n_1}{n}F_{n_1}^{(1)}(x)+\frac{n_2}{n}F_{n_2}^{(2)}(x),$$

这里

$$F_{n_1}^{(1)}(x)=\frac{1}{n_1}\sum_{i=1}^{n_1}I_{(y_i\le x)},\qquad F_{n_2}^{(2)}(x)=\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}I_{(y_i\le x)}.$$

于是对$F(x)=\alpha_1F_1(x)+\alpha_2F_2(x)$有

$$\begin{aligned}\sup_x|\widehat{F}_n(x)-F(x)| \;\le\; &\Big|\frac{n_1}{n}-\alpha_1\Big|+\sup_x|F_{n_1}^{(1)}(x)-F_1(x)|\\ &+\Big|\frac{n_2}{n}-\alpha_2\Big|+\sup_x|F_{n_2}^{(2)}(x)-F_2(x)|.\end{aligned}$$

利用Glevenko-Cantelli定理知

$$\sup_x|\widehat{F}_n(x)-F(x)|\to0,\qquad \text{a.s.}\ \Big(n_1\to\infty,\ n_2\to\infty\ \text{且}\ \frac{n_1}{n_1+n_2}\to\alpha_1\Big). \tag{2.10}$$

我们进一步指出

$$G_n=1-\frac{1}{\overline{y}_n}\int_0^{\infty}(1-\widehat{F}_n(x))^2\mathrm{d}x. \tag{2.11}$$

实际上, 从(2.8)式知

$$G_n=\frac{n+1}{n}-\frac{2}{\overline{y}_n}\sum_{i=1}^{n}(n-i+1)y_{(i)}, \tag{2.12}$$

这里$y_{(1)} \le y_{(i)} \le \cdots \le y_{(n)}$是$y_1, \ldots, y_n$的从小到大排列, 记$y_{(0)} = 0$.

另外一方面,

$$\begin{aligned}
\int_0^\infty (1-\widehat{F}_n(x))^2 \mathrm{d}x &= \sum_{i=1}^n \int_{y_{(i-1)}}^{y_{(i)}} \Big(1-\frac{i-1}{n}\Big)^2 \mathrm{d}x \\
&= \sum_{i=1}^n \Big(1-\frac{i-1}{n}\Big)^2 (y_{(i)} - y_{(i-1)}) \\
&= y_{(n)}\Big(1-\frac{n-1}{n}\Big)^2 + \sum_{i=1}^{n-1} y_{(i)}\Big(2-\frac{2i-1}{n}\Big)\frac{1}{n} \\
&= \frac{y_{(n)}}{n^2} + \frac{1}{n^2}\sum_{i=1}^{n-1} y_{(i)}(2n-2i+1) \\
&= \frac{y_{(n)}}{n^2} + \frac{2}{n^2}\sum_{i=1}^{n-1} y_{(i)}(n-i+1) - \frac{1}{n^2}\sum_{i=1}^{n-1} y_{(i)} \\
&= \frac{2}{n^2}\sum_{i=1}^{n} y_{(i)}(n-i+1) - \frac{1}{n^2}\sum_{i=1}^{n} y_{(i)},
\end{aligned}$$

于是

$$1 - \frac{1}{\overline{y}}\int_0^\infty (1-\widehat{F}_n(x))^2 \mathrm{d}x = 1 + \frac{1}{n} - \frac{2}{n^2\overline{y}}\sum_{i=1}^{n}(n-i+1)y_{(i)} = G_n.$$

这就证明(2.11)式成立. 又由

$$\begin{aligned}
\Big|\int_0^\infty (1-\widehat{F}_n(x))^2 \mathrm{d}x - \int_0^\infty (1-F(x))^2 \mathrm{d}x\Big| &\le \int_0^\infty |(1-\widehat{F}_n)^2 - (1-F)^2| \mathrm{d}x \\
&\le 2\int_0^\infty |\widehat{F}_n - F| \mathrm{d}x,
\end{aligned}$$

易知

$$\begin{aligned}
&\int_0^\infty (1-\widehat{F}_n(x))\mathrm{d}x = \frac{1}{n}\sum_{i=1}^n y_i = \frac{n_1}{n}\frac{\sum_{i=1}^{n_1} y_i}{n_1} + \frac{n_2}{n}\frac{\sum_{i=n_1+1}^{n_1+n_2} y_i}{n_2} \\
\to\quad & \alpha_1\mu^{(1)} + \alpha_2\mu^{(2)} = \int_0^\infty (1-F(x))\mathrm{d}x < \infty \qquad \text{a.s.},
\end{aligned}$$

这里

$$\mu^{(k)} = \int_0^\infty x \mathrm{d}F_k(x), \qquad k = 1, 2.$$

由于$1-\widehat{F}_n(x) \ge 0$以及(2.10)式知道

$$1-\widehat{F}_n(x) \xrightarrow{L_1} 1-F(x),$$

即有

$$\int_0^\infty |\widehat{F}_n(x) - F(x)| \mathrm{d}x \to 0, \qquad n \to \infty, \text{ a.s..}$$

由此知当$n_1 \to \infty$, $n_2 \to \infty$且$n_1/(n_1+n_2) \to \alpha_1$时,

$$\int_0^\infty (1-\widehat{F}_n(x))^2 \mathrm{d}x \to \int_0^\infty (1-F(x))^2 \mathrm{d}x \qquad \text{a.s..}$$

而

$$\overline{y}_n = \frac{\sum\limits_{i=1}^n y_i}{n} \to \mu = \int_0^\infty (1-F(x))\mathrm{d}x \qquad \text{a.s.,}$$

从(2.11)式知

$$\lim_{n\to\infty} G_n = 1 - \frac{1}{\mu}\int_0^\infty (1-F(x))^2 \mathrm{d}x \qquad \text{a.s..}$$

再利用引理2.1知(2.9)式成立, 引理2.2证毕. □

定理2.1的证明: 设数据$y_1, y_2, \ldots, y_{n_1}$是来自$F_1(x)$的简单随机样本, $y_{n_1+1}, \ldots, y_{n_1+n_2}$是来自$F_2(x)$的简单随机样本.

将$y_1, \ldots, y_{n_1}$按照分界点$\{a_i^{(1)}\}$划分为s_1+1个组, 属于$[0, a_1^{(1)}]$的记为$y_{11}, y_{12}, \ldots, y_{1m_1}$, 属于$(a_{i-1}^{(1)}, a_i^{(1)}]$的记为$y_{i1}, y_{i2}, \ldots, y_{im_i}$, $i = 2, 3, \ldots, s_1+1$;

将$y_{n_1+1}, \ldots, y_{n_1+n_2}$按照分界点$\{a_i^{(2)}\}$划分为$s_2+1$个组, 属于$[0, a_1^{(2)}]$的记为$y_{s_1+2,1}$, $y_{s_1+2,2}, \ldots, y_{s_1+2,m_{s_1+2}}$, 属于$(a_{i-1}^{(2)}, a_i^{(2)}]$的记为$y_{s_1+1+i,1}, y_{s_1+1+i,2}, \ldots, y_{s_1+1+i,m_{s_1+1+i}}$, $i = 2, 3, \ldots, s_2+1$.

记$n = n_1+n_2$, $s = s_1+s_2+2$, 这样的n个数据$y_1, y_2, \ldots, y_n$划分为s个组, 第i组的数据是$y_{i1}, \ldots, y_{im_i}$, $i = 1, 2, \ldots, s$, 由公式(2.8)知, 数组$y_1, y_2, \ldots, y_n$的基尼系数G_n为

$$G_n = \frac{1}{2n^2\overline{y}} \sum_{1\le i,u\le s} \sum_{\substack{1\le j\le m_i \\ 1\le v\le m_u}} |y_{ij} - y_{uv}|, \tag{2.13}$$

这里$\overline{y} = 1/n\sum\limits_{i=1}^n y_i$. 令

$$E_{iu} = \sum_{j=1}^{m_i}\sum_{v=1}^{m_u} |y_{ij}-y_{uv}|, \qquad \overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij},$$

由于$|x| = 2x^+ - x$, (这里$x^+ = \max(x, 0)$)

$$\begin{aligned} E_{iu} &= 2\sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv})^+ - \sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv}) \\ &= 2\sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv})^+ - m_i m_u(\overline{y}_i - \overline{y}_u). \end{aligned} \tag{2.14}$$

又知

$$\sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv})^+ \ge \sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv}) = m_i m_u(\overline{y}_i - \overline{y}_u),$$

于是

$$\sum_{j=1}^{m_i}\sum_{v=1}^{m_u}(y_{ij}-y_{uv})^+ \geq m_i m_u(\overline{y}_i-\overline{y}_u)^+.$$

利用(2.14)的结果得到

$$E_{iu} \geq m_i m_u[2(\overline{y}_i-\overline{y}_u)^+ - (\overline{y}_i-\overline{y}_u)] = m_i m_u|\overline{y}_i-\overline{y}_u|.$$

再依据(2.13)式有

$$G_n \geq \frac{1}{2n^2\overline{y}}\sum_{1\leq i,u\leq s} m_i m_u|\overline{y}_i-\overline{y}_u|. \tag{2.15}$$

既然假设$n_1\to\infty$, $n_2\to\infty$且$n_1/(n_1+n_2)\to\alpha_1$, 根据约定(2.3)和(2.4)

$$\lim_{n\to\infty}\frac{m_i}{n}=p_i \qquad \text{a.s.}\ \ i=1,\ldots,s,$$
$$\lim_{n\to\infty}\overline{y}_i=\mu_i \qquad \text{a.s.}\ \ i=1,2,\ldots,s,$$
$$\lim_{n\to\infty}\overline{y}=\mu=\sum_{i=1}^{s}p_i\mu_i \qquad \text{a.s.}.$$

对(2.15)式取极限, 利用引理2.2即知道

$$G\geq\frac{1}{2\mu}\sum_{i=1}^{s}\sum_{j=1}^{s}|\mu_i-\mu_j|p_ip_j.$$

这说明(2.5)式成立, 定理2.1证毕. □

值得注意的是, 不等式(2.2)的右端并不含有分界点$a_1^{(k)},a_2^{(k)},\ldots,a_{s_k}^{(k)}$ $(k=1,2)$. 换句话说, 只要分界点存在而$\{p_i^{(k)},\mu_i^{(k)}\}$ $(i=1,2,\ldots,s_k+1,\ k=1,2)$以及$\alpha_k$ $(k=1,2)$已知, 就可以计算混合总体的基尼系数的下限.

下面指出, 这个下限是最大下限, 因而是一个最优下限. 实际上, 假设总体Y_1和Y_2分别有分组数据$\{p_i^{(k)},\mu_i^{(k)}\}$ $(i=1,2,\ldots,s_k+1,\ k=1,2)$, 又混合比例$\alpha_1,\alpha_2$已知. 我们可以找到相互独立的随机变量$Y_1^*$和$Y_2^*$使得$Y_k^*$和$Y_k$有相同的分组数据$(k=1,2)$, 但是分界点可能不同, 使得$Y_1^*$和$Y_2^*$混合后的随机变量$Y^*$的基尼系数恰好等于不等式(2.2)的右端.

更加确切的叙述如下. 沿用上面的记号, 对于$k=1,2$, 令

$$Y_k^*=\begin{cases}\mu_1^{(k)}, & \text{当 } Y_k\leq a_1^{(k)};\\ \mu_i^{(k)}, & \text{当 } a_{i-1}^{(k)}<Y_k\leq a_i^{(k)},\ i=2,\ldots,s_k+1,\end{cases}$$

则不难证明

$$\mathsf{P}(Y_k^*\leq\mu_1^{(k)})=\mathsf{P}(Y_k\leq a_1^{(k)})=p_1^{(k)},$$
$$\mathsf{P}(\mu_{i-1}^{(k)}<Y_k^*\leq\mu_i^{(k)})=\mathsf{P}(a_{i-1}^{(k)}<Y_k\leq a_i^{(k)})=p_i^{(k)}, \qquad i=2,\ldots,s_k+1,$$
$$\mathsf{E}(Y_k^*|Y_k^*\leq\mu_1^{(k)})=\mu_1^{(k)},$$
$$\mathsf{E}(Y_k^*|\mu_{i-1}^{(k)}<Y_k^*\leq\mu_i^{(k)})=\mu_i^{(k)}, \qquad i=2,\ldots,s_k+1.$$

总之, Y_k^*和Y_k有相同的分组数据, 但是分界点可能不同.

设Δ是与(Y_1, Y_2)相互独立的随机变量, 且

$$\mathsf{P}(\Delta = 1) = \alpha_1 = 1 - \mathsf{P}(\Delta = 2) \qquad (0 < \alpha_1 < 1).$$

令$Y^* = Y_\Delta^*$, 则我们有下面的定理.

定理 2.2 设混合总体Y^*的基尼系数是G^*, 则G^*等于不等式(2.2)的右端.

证明: 设Y_k^*的分布函数是$F_k^*(x)$, $k = 1, 2$, 则Y^*的分布函数$F^*(x) = \alpha_1 F_1^*(x) + \alpha_2 F_2^*(x)$, 于是$Y^*$的基尼系数为

$$G^* = 1 - \frac{1}{\mu^*}\int_0^\infty (1 - F^*(x))^2 \mathrm{d}x \qquad (\text{依据引理2.1}),$$

其中$\mu^* = \mathsf{E}Y^* = \alpha_1\mu^{(1)} + \alpha_2\mu^{(2)}$, 这里$\mu^{(k)} = \mathsf{E}Y_k^* = \sum\limits_{i=1}^{s_k+1} p_i^{(k)}\mu_i^{(k)}$, 易知这个$\mu^*$与(2.2)中的$\mu$相等.

同时易知

$$\begin{aligned}\int_0^\infty (1 - F^*(x))^2 \mathrm{d}x &= \alpha_1^2\int_0^\infty (1 - F_1^*(x))^2 \mathrm{d}x + \alpha_2^2\int_0^\infty (1 - F_2^*(x))^2 \mathrm{d}x \\ &\quad + 2\alpha_1\alpha_2\int_0^\infty (1 - F_1^*(x))(1 - F_2^*(x))\mathrm{d}x,\end{aligned}$$

不难看出

$$F_k^*(x) = \sum_{i=1}^{s_k+1} I(\mu_i^{(k)} \le x)p_i^{(k)},$$

这里$I(A)$表示集合A的示性函数,

$$1 - F_k^*(x) = \sum_{i=1}^{s_k+1} I(\mu_i^{(k)} > x)p_i^{(k)},$$

于是

$$\begin{aligned}\int_0^\infty (1 - F_k^*(x))^2 \mathrm{d}x &= \sum_{i=1}^{s_k+1}\sum_{j=1}^{s_k+1}\int_0^\infty I(\mu_i^{(k)} > x)I(\mu_j^{(k)} > x)p_i^{(k)}p_j^{(k)}\mathrm{d}x \\ &= \sum_{i=1}^{s_k+1}\sum_{j=1}^{s_k+1}\min\{\mu_i^{(k)}, \mu_j^{(k)}\}p_i^{(k)}p_j^{(k)}.\end{aligned}$$

利用$\min\{x, y\} = (x + y - |x - y|)/2$易知

$$\begin{aligned}\int_0^\infty (1 - F_k^*(x))^2 \mathrm{d}x &= \frac{1}{2}\sum_{i=1}^{s_k+1}\sum_{j=1}^{s_k+1}(\mu_i^{(k)} + \mu_j^{(k)})p_i^{(k)}p_j^{(k)} \\ &\quad - \frac{1}{2}\sum_{i=1}^{s_k+1}\sum_{j=1}^{s_k+1}|\mu_i^{(k)} - \mu_j^{(k)}|p_i^{(k)}p_j^{(k)} \\ &= \mu^{(k)} - \sum_{1\le i<j\le s_k+1}|\mu_i^{(k)} - \mu_j^{(k)}|p_i^{(k)}p_j^{(k)},\end{aligned}$$

$$\begin{aligned}\int_0^\infty (1-F_1^*(x))(1-F_2^*(x))\mathrm{d}x &= \sum_{i=1}^{s_1+1}\sum_{j=1}^{s_2+1}\min\{\mu_i^{(1)},\mu_j^{(2)}\}p_i^{(1)}p_j^{(2)}\\ &= \frac{1}{2}\sum_{i=1}^{s_1+1}\sum_{j=1}^{s_2+1}(\mu_i^{(1)}+\mu_j^{(2)})p_i^{(1)}p_j^{(2)}\\ &\quad -\frac{1}{2}\sum_{i=1}^{s_1+1}\sum_{j=1}^{s_2+1}|\mu_i^{(1)}-\mu_j^{(2)}|p_i^{(1)}p_j^{(2)},\end{aligned}$$

于是

$$\begin{aligned}G^* &= 1-\frac{1}{\mu^*}\Big\{\alpha_1^2\mu^{(1)}-\alpha_1^2\sum_{1\le i<j\le s_1+1}|\mu_i^{(1)}-\mu_j^{(1)}|p_i^{(1)}p_j^{(1)}\\ &\quad +\alpha_2^2\mu^{(2)}-\alpha_2^2\sum_{1\le i<j\le s_2+1}|\mu_i^{(2)}-\mu_j^{(2)}|p_i^{(2)}p_j^{(2)}\\ &\quad +\alpha_1\alpha_2(\mu^{(1)}+\mu^{(2)})-\alpha_1\alpha_2\sum_{i=1}^{s_1+1}\sum_{j=1}^{s_2+1}|\mu_i^{(1)}-\mu_j^{(2)}|p_i^{(1)}p_j^{(2)}\Big\}\\ &= 1-\frac{\alpha_1\mu^{(1)}+\alpha_2\mu^{(2)}}{\mu^*}+\frac{\alpha_1^2}{\mu^*}\sum_{1\le i<j\le s_1+1}|\mu_i^{(1)}-\mu_j^{(1)}|p_i^{(1)}p_j^{(1)}\\ &\quad +\frac{\alpha_2^2}{\mu^*}\sum_{1\le i<j\le s_2+1}|\mu_i^{(2)}-\mu_j^{(2)}|p_i^{(2)}p_j^{(2)}+\frac{\alpha_1\alpha_2}{\mu^*}\sum_{i=1}^{s_1+1}\sum_{j=1}^{s_2+1}|\mu_i^{(1)}-\mu_j^{(2)}|p_i^{(1)}p_j^{(2)}.\end{aligned}$$

这表明G^*等于(2.2)式的右端. 定理2.2证毕. □

§3. 我国2003~2009年个人收入基尼系数的下限

本节是利用定理2.1中的不等式(2.2)对《中国统计年鉴》上公布的2003至2009年我国个人年收入的分组数据进行计算, 给出各年基尼系数的下限值. 该分组数据是就城镇和农村分别列出的. 以2008年的为例介绍如下.

表1 2008年城镇居民收入的分组数据

项目	困难户	很低收入户	低收入户	中等偏下户	中等收入户	中等偏上户	高收入户	最高收入户
调查户数(户)	3137	3216	6485	12983	12993	12998	6445	6418
调查户比重(%)	4.85	4.97	10.03	20.07	20.09	20.10	9.96	9.92
平均每户家庭人口(人)	3.34	3.33	3.22	3.06	2.89	2.74	2.62	2.51
平均每人可支配收入(元)	3734.35	5753.37	7363.28	10195.56	13984.23	19254.08	26250.1	43613.75

注: 我们将原表中的最低收入户分为困难户和很低收入户两类.

表2 2008年农村居民收入的分组数据

项目	低收入户	中低收入户	中等收入户	中高收入户	高收入户
调查户比重(%)	20	20	20	20	20
平均每户人口数	4.54	4.32	4.07	3.76	3.37
平均每人年纯收入(元)	1499.81	2934.99	4203.12	5928.6	11290.2

从表1知, 城镇居民按收入高低分为8 ($s_1 = 7$)组. 被调查的总人数是187924, 人均年收入是16049元, 各组人数所占比例依次是0.0558, 0.0568, 0.1112, 0.2114, 0.1998, 0.1896, 0.0898, 0.0857, 这些看作(2.2)式中$p_1^{(1)}, \ldots, p_8^{(1)}$的值, 表1中的最后一行可以看作$\mu_1^{(1)}, \ldots, \mu_8^{(1)}$的值.

从表2知, 农村居民按收入从低到高分为5 ($s_2 = 4$)组, 各组人数所占比例依次是0.2263, 0.2154, 0.2029, 0.1824, 0.1680, 这些可以看作(2.2)式中的$p_1^{(2)}, \ldots, p_5^{(2)}$的值, 表2的最后一行可以看作$\mu_1^{(2)}, \ldots, \mu_5^{(2)}$的值, 不难推知农村的人均纯收入是4832元.

有了$p_i^{(k)}, \mu_i^{(k)}$ ($i = 1, 2, \ldots, s_k + 1$, $k = 1, 2$)的值以及城镇人口所占比例$\alpha_1 = 45.68\%$ (根据全国人口普查得到), 代入(2.2)式可得全国基尼系数的下限值为0.4542.

类似地, 我们也对2003至2007年以及2009年的数据进行了计算, 各年的基尼系数的下限值见表3.

应该指出的是, 北京大学基尼系数课题组[13]对基尼系数进行了系统的研究, 在对我国城镇和农村居民收入的分布函数$F_1(x)$和$F_2(x)$加了两种比较合理的假定(设相应的Lorenz函数为$L_1(p)$和$L_2(p)$分别适合VA模型和β模型)后, 给出了城乡合在一起时基尼系数G的估计值, 也列出在表3中.

表3 基尼系数的估计值与下限比较

年份	基于VA模型	基于β模型	下限
2003	0.4563	0.4585	0.4417
2004	0.4561	0.4566	0.4441
2005	0.4621	0.4642	0.4507
2006	0.4619	0.4636	0.4503
2007	0.4621	0.4684	0.4494
2008	0.4633	0.4678	0.4542
2009	0.4664	0.4683	0.4540

从表3中可以看到, 我们得到的下限值与点估计值相差不太大. 应该说, 对于分布函数的任何特殊假定(假设适合某种模型)都可能引发争议, 而我们在无任何特殊假定下得到的下限是无可争议的. 这样的下限对于了解我国的贫富差距程度有重要意义.

参 考 文 献

[1] 厉以宁, 秦宛顺, 现代西方经济学概论(第二版), 北京大学出版社, 1992.

[2] 陈奇志, 陈家鼎, 关于洛伦兹曲线和基尼系数的一点注记, 北京大学学报(自然科学版), **42(5)**(2006), 613–618.

[3] 徐宽, 基尼系数的研究文献在过去八十年是如何拓展的, 经济学(季刊), **2(4)**(2003), 757–778.

[4] 陈奇志, 陈家鼎, 关于洛伦兹曲线和基尼系数的统计推断, 北京大学技术报告, 2008.

[5] Schader, M. and Schmid, F., Fitting parametric Lorenz curves to grouped income distributions – a critical note, *Empirical Economics*, **19**(1994), 361–370.

[6] Cheong, K.S., An empirical comparison of alternative functional forms for the Lorenz curve, *Applied Economics Letters*, **9**(2002), 171–176.

[7] Gastwirth, J.L., The estimation of the Lorenz curve and Gini index, *The Review of Economics and Statistics*, **54**(1972), 306–316.

[8] Mehran, F., Bounds on the Gini index based on observed points of the Lorenz curve, *Journal of the American Statistical Association*, **70**(1975), 64–66.

[9] Ogwang, T., Bounds of the Gini index using sparse information on mean incomes, *Review of Income and Wealth*, **49**(2003), 415–423.

[10] 周文兴, 中国总体基尼系数测定问题 — 兼评"陈宗胜–李实论战"并与陈宗胜教授商榷, 南开经济研究, **3**(2003), 37–41.

[11] 程永宏, 二元经济中城乡混合基尼系数的计算与分解, 经济研究, **41(1)**(2006), 109–120.

[12] 罗青, 中国改革开放时期的基尼系数的研究, 北京大学博士学位论文, 2005.

[13] 北京大学基尼系数课题组, 洛伦兹曲线和基尼系数的统计推断与科学计算, 北京大学技术报告, 2010.

Lower Bound of Gini Index for Mixed Distribution — with Results for China with Urban-Rural Dualistic Structure

Chen Jiading Fang Xiangzhong Shi Pixu Hu Xiaoou Cao Li

(*School of Mathematical Sciences, Peking University, Beijing, 100871*)

This paper studies the lower bound of the Gini index for the mixed distribution of two distributions with certain weights by grouped data. These results are used to calculate the Gini index for China with urban-rural dualistic structure. They could be expanded to mixed distributions with more than two sub-distributions and could be used to other real situations, such as union of countries or areas.

Keywords: Gini index, mixed distribution, lower bound.

AMS Subject Classification: 62P20, 62F30.