中国基尼系数的估算研究

王祖祥 张 奎 孟 勇*

摘要: 中国的收入不平等受到了国内外的广泛关注。公开出版物上的收入分配数据都是分组形式的,这给收入不平等的测算带来困难。本文采用城乡收入分配统计分布的构造方法,利用《中国统计年鉴》(1995 - 2005)的收入分配数据估算了我国的基尼系数。结果表明,我国目前城镇与农村两部门内部的基尼系数都不大,都没有超过 0. 34,但从 2003年开始,我国的加总基尼系数已经超过了 0. 44,远远越过了警戒水平 0. 4。实际上,基尼系数的分解公式说明,影响我国收入不平等程度的关键因素是目前巨大的城乡收入差距,是这一因素决定了我国的基尼系数必然很大。

关键词: 收入分配 洛伦兹曲线 基尼系数 密度函数

中国的收入不平等程度受到了国内外的广泛关注,出现了各种各样的基尼系数估计值。我国每年在《中国统计年鉴》中都发布收入分配数据,但一般认为利用该数据难以估算基尼系数 (王学力, 2000),一是因为这种数据是分组形式的,城镇收入分配数据中只列出了从低到高若干个收入组的平均收入与人口份额,农村收入分配中只给出了各个收入区间及各个区间内的家庭百分数,二是城乡数据分列。实际上,寻求收入分配的统计分布是现代收入分配分析活跃的研究领域,洛伦兹曲线正是从收入分配的密度函数出发而定义的,又按定义,基尼系数是洛伦兹曲线与平等收入线之间面积的 2倍,可见基尼系数的估算应建立在收入分配统计分布或洛伦兹曲线的准确测算的基础上。实际工作中,在只有分组数据可用的条件下,可以先估计收入分配的密度函数,从而得到相应的洛伦兹曲线,或直接估算洛伦兹曲线,最后再估计基尼系数。国外经济理论文献中基尼系数的估算一般遵循两种途径,一是利用分户数据直接估计收入分配的密度函数从而估算基尼系数,二是利用分组数据估计洛伦兹曲线,然后再估算基尼系数。我国统计部门的城乡收入分配调查的分户数据不对外公开,因此本文考虑使用统计年鉴中的分组数据。实际上,使用统计年鉴中的数据时,城镇基尼系数的估算可以使用第二种方法,而对于农村收入分配数据,由于缺少各个收入区间内的平均收入信息使得不能利用第二种方法。王祖祥 (2006)提出了根据我国收入分配分组数据构造收入分配密度函数的方法,估算了我国中部六省的基尼系数。使用这种方法,只要相关部门提供信息量不高的分组数据,就可以计算我国任何部门、任何地域的基尼系数与其他大多数收入不平等指数,还可以利用现代收入分配分析方法对我国的收入分配进行进一步的分析。①

本文利用王祖祥 (2006)提出的方法 (同时改进了其中城镇密度函数的构造方法),估算了我国最近 10年的基尼系数,实际计算表明,我国目前城镇与农村两部门内部的基尼系数都不大,都没有超过 0. 34,但从 2003年开始,我国的加总基尼系数已经超过了 0. 44,远远越过了警戒水平 0. 4。实际上,基尼系数的分解公

* 王祖祥,武汉大学经济与管理学院,邮政编码: 430072,电子信箱: zxwang@whu. edu. cn;张奎,武汉大学经济与管理学院,邮政编码: 430072;孟勇,山西财经大学统计学院,邮政编码: 034000,电子信箱: m7025y@163. com。

本文研究得到了国家社科基金重点项目 (批准号 04AJL002)与湖北省社会科学基金重点项目“湖北省农村贫困动态评估研究”的资助。

①国内很多学者考虑了我国基尼系数的估算问题,例如李实等 (1998)、李强等 (1995)、胡祖光 (2004)、董静和李子奈 (2004)等。使用我国统计年鉴中的分组数据, Chotikapanich等 (2007)也考虑了我国的基尼系数,该文利用一种经验分布来逼近我国农村分组数据,计算得到的农村基尼系数与本文结果相差不大。

式说明,影响我国目前收入不平等的决定因素是农村与城镇之间的收入差距。从最后得到的全国洛伦兹曲线可见,2004年中占人口份额 50%的低收入群体所拥有的收入份额只有 20%左右,人口份额为 10%的高收入端拥有近 32%的总收入,这部分人口拥有的总收入是最低收入端 10%群体的近 20倍。因此,我国的收入不平等问题的动向值得关注。

一、城乡加总基尼系数的计算公式

我国城乡两部门收入分配数据分列,如何加总两部门的收入分配进而形成全国的加总收入分配一直是困扰我国经济理论界的一个问题。实际上,一旦收入分配密度函数的估算问题得到解决,这一问题将迎刃而解。这里先讨论基尼系数的一种分解公式,再说明收入分配统计分布的加总方法。

收入分配的洛伦兹曲线 $L(p)$在收入分配分析中具有重要地位,$L(p)$表示人口份额等于 p的低收入端拥有的总收入份额,因此 $L(p)$是定义于 $[0,1]$区间上的函数。按经济意义,它应满足如下条件:

(1) $L(p)$是 p的增函数,即有 $L'(p)\geq 0$。因为所考虑的低收入端人口份额 p越大,该群体拥有的总收入份额应越大。

(2) $L(p)$是凸函数,即满足 $L''(p)\geq 0$。因为 p增加到 $p+\Delta p$时,人口份额 Δp所代表的是收入更高的群体,因此 p增加时,$L(p)$应以更大比例增加。

(3) $L(p)\geq 0$,因为收入份额不能是负数。

(4) $L(0)=0, L(1)=1$。

如果对于任何 $p\in[0,1]$都有 $L(p)=p$,则此洛伦兹曲线是所谓平等收入线。对于任何洛伦兹曲线 $L(p)$,基尼系数定义为 $L(p)$与平等收入线之间面积的 2倍。

记农村与城镇两部门人口的总数为 n,记农村人口数为 n_1,城镇人口数为 n_2,记第 i个部门内的收入分配为 $Y_i=(y_{i1},y_{i2},\cdots,y_{in_i})$,即第 i个部门内第 k个成员的收入为 y_{ik},本文恒假定任何成员的收入都大于或等于零。两部门的收入分配合在一起构成全国的收入分配 Y。记第 i个部门内成员的平均收入为 μ_i,记第 i个部门内的成员占总成员的份额为 $p_i=\frac{n_i}{n}$,可见总平均收入为 $\mu=\frac{1}{n}(n_1\mu_1+n_2\mu_2)=p_1\mu_1+p_2\mu_2$,下面记第 i个部门的所有成员拥有总收入的份额为 $s_i=\frac{n_i\mu_i}{n\mu}$。

按定义,第 j个部门内的基尼系数可以表示为:

$$G_j=\frac{1}{2n_j^2\mu_j}\sum_{i=1}^{n_j}\sum_{r=1}^{n_j}|y_{ji}-y_{jr}| \tag{1}$$

又记:

$$G_{12}=\frac{1}{n_1n_2(\mu_1+\mu_2)}\left[\sum_{i=1}^{n_1}\sum_{r=1}^{n_2}|y_{1i}-y_{2r}|\right] \tag{2}$$

同样定义 G_{21},可见有 $G_{21}=G_{12}$。记全国基尼系数为 $G(Y)$, Dagum (1997)给出了基尼系数的如下分解公式:

$$\begin{aligned}G(Y)&=p_1s_1G_1+p_2s_2G_2+p_1s_2G_{12}+p_2s_1G_{21}\\&=p_1s_1G_1+p_2s_2G_2+(p_1s_2+p_2s_1)G_{12}\end{aligned} \tag{3}$$

由此即可计算我国城乡合一的基尼系数。Dagum称 G_{12}为两部门的扩展基尼系数,它反映了两部门的组间不平等程度。

(3)式是离散条件下精确的基尼系数公式,由于只能得到分组形式的收入分配数据,因此不能用它进行实际计算。如果已知农村与城镇收入分配的洛伦兹曲线,分别记为 $L_1(p)$与 $L_2(p)$,或已知两个收入分配的密度函数,例如记为 $f_1(x)$与 $f_2(x)$,则由连续条件下的基尼系数定义,农村与城镇的基尼系数 G_1与 G_2有如下公式:

$$G_i=1-2\int_0^1L_i(p)\,dp=1-\frac{2}{\mu_i}\int_0^{\infty}\left[\int_0^x yf_i(y)\,dy\right]f_i(x)\,dx \tag{4}$$

在连续分配条件下,对应于 (2)式的公式为:

$$G_{12}=\frac{1}{\mu_1+\mu_2}\int_0^{\infty}\int_0^{\infty}|x-y|f_1(x)f_2(y)\,dx\,dy \tag{5}$$

与绝大部分国外学者一样,笔者用连续分布来逼近离散的收入分配,这样,估计我国基尼系数的关键是构造收入分配的近似密度函数,只要这一问题解决了,将 (4)式与 (5)式代入 (3)式即得到我国的基尼系数,同时还得到了反映两部门之间不平等的指标 G_{12}。

由 (5)式容易看出,当 $\mu_2 > \mu_1$ 时,如果收入分配 Y_1 与 Y_2 的收入范围不重叠,则有 $G_{12} = \mu_2 - \mu_1$。又直观上可见 Y_1 与 Y_2 的重叠程度越小,则 G_{12}越大,因此 G_{12}还反映了两个收入分配的收入范围的重叠程度。这样, (3)式中最右边第三项将对我国的基尼系数有重要影响,因为我国城乡收入差距很大,即两部门收入分配的重叠程度不大,从而可以预料 G_{12}会很大。又我国农村人口比例 p_1 较大,而城镇人口拥有的总收入份额 s_2 也较大,因此 (3)式中 G_{12}的系数 $p_1 s_2 + p_2 s_1$ 将不小。这两方面因素决定我国的基尼系数不会小。注意到有:

$$|\mu_1 - \mu_2| = \left|\int_0^\infty\int_0^\infty (x - y)\, g_1(x)\, g_2(y)\, dxdy\right| \le \int_0^\infty\int_0^\infty |x - y|\, g_1(x)\, g_2(y)\, dxdy$$

因此得到 G_{12}的下界估计:

$$G_{12} \ge \frac{|\mu_1 - \mu_2|}{\mu_1 + \mu_2}$$

例如 2003年我国农村人均纯收入为 $\mu_1 = 2\ 622\ 24$,城镇平均可支配收入为 $\mu_2 = 8\ 472\ 20$,考虑这两个分配构成全国收入分配 Y时就有:

$$G_{12} \ge 0.\ 5272875$$

注意到将上述 G_{12}的下界代入 (3)式可以得到基尼系数的一个下界。归纳起来得到:

定理 1:中国收入分配的基尼系数可以表示为 (3)式,基尼系数的下界为:

$$p_1 s_1 G_1 + p_2 s_2 G_2 + (p_1 s_2 + p_2 s_1)\frac{|\mu_1 - \mu_2|}{\mu_1 + \mu_2}$$

其中 p_i、s_i、μ_i与 G_i分别是部门 i内的人口份额、收入份额、平均收入与基尼系数, G_{12}是两部门的扩展基尼系数。

显然,如果估计得到了城镇与农村收入分配的密度函数,分别记为 $f_1(x)$与 $f_2(x)$,记相应分布函数分别为 $F_1(x)$与 $F_2(x)$, $F_1(x)$表示城镇人口中收入不高于 x的人口占城镇总人口的份额, $F_2(x)$表示农村人口中收入不高于 x的人口占农村总人口的份额,于是全国人口中收入不高于 x的人口占全国总人口的份额应为: $F(x) = p_1 F_1(x) + p_2 F_2(x)$, $F(x)$即全国收入分配的统计描述。① 于是全国收入分配的密度函数应为: $f(x) = p_1 f_1(x) + p_2 f_2(x)$。按洛伦兹曲线与基尼系数的定义,如果估计得到了两部门的密度函数 $f_1(x)$与 $f_2(x)$,就可以得到整个国家的洛伦兹曲线 $L(p)$的另一公式:

$$L(p) = \frac{1}{\mu}\int_0^x y f(y)\, dy$$

其中, $p = F(x)$,而全国基尼系数公式为:

$$G(y) = 1 - \frac{2}{\mu}\int_0^\infty f(x)\int_0^x y f(y)\, dydx$$

其中, μ是全国平均收入。因此,如果得到了两部门的密度函数,一是可以利用 (5)式估计两部门之间的收入不平等,二是可以通过定理 1得到基尼系数的下界估计,三是可以通过 (4)式与 (3)式或上述积分计算两部门或全国的基尼系数。同时,利用这些密度函数还可以进行收入分配的其他分析。可见,获得收入分配密度函数的方法本身具有重要意义。

二、城镇密度函数的构造方法

一般各国统计部门都是通过抽样调查对收入分配进行估计,又由于保密等原因,一般将抽样数据化成分组形式予以发布,理论界只能在这种数据的基础上对收入分配进行分析。根据可能得到的数据形式,可以直接估算收入分配的密度函数,第三部分构造农村收入分配的密度函数时将采用这一方法。也可以从估计洛伦兹曲线入手获得密度函数,这里采将用这一方法获得城镇收入分配的密度函数,将按经济意义与数学性质

①王祖祥 (2006)使用了这种加总公式, Chotikapanich等 (2007)也是使用这种加总分布计算我国基尼系数的。

选择适当的函数作为洛伦兹曲线的经验公式,再利用分组数据估计其中的参数,从而得到近似洛伦兹曲线,最后利用洛伦兹曲线与密度函数的关系而得到后者。

设收入分配的密度函数为 $r(x)$,相应分布函数记为 $R(x)$,记 $p=R(x)$,则由于洛伦兹曲线定义为:

$$L(p)=\frac{1}{\mu}\int_0^x r(t)\,dt$$

其中 μ是相应的平均收入,则可见有:

$$\frac{dL(p)}{dp}=\frac{dL(p)}{dx}\frac{dx}{dp}=\frac{x}{\mu} \tag{6}$$

$$\frac{d^2L(p)}{dp^2}=\frac{1}{\mu r(x)} \tag{7}$$

因此对于任何 x,从 (6)式解出 p,即得到 x处的分布函数值 $p=R(x)$。又对于任何 $p=R(x)$,计算二阶导数值 $L''(p)$,由 (7)式即得 x处的密度函数值 $r(x)$。

在《中国统计年鉴》的城镇收入分配分组数据中,依收入从低到高给出了 $m+1(=8)$个收入组的家庭数、每组家庭的平均人口、年人均可支配收入,另给出了平均可支配收入 μ,但没有收入区间信息。依每组家庭的平均人口可以算出每组的人口份额,记第 i个收入组的人口份额为 f_i,平均收入为 μ_i。于是前 i个低收入组的人口份额为 $p_i=\sum_{j=0}^{i}f_j$,这部分成员拥有的收入份额为 $L(p_i)=\frac{1}{\mu}\sum_{j=0}^{i}f_j\mu_j$。因此由分组数据能够得到洛伦兹曲线上的点列 $(p_i,L(p_i))_{i=1}^{m+1}$。对于满足洛伦兹曲线条件的适当函数 $L(p,\tau)$,其中 τ是参数向量,通过解所谓非线性最小二乘问题:

$$\min\sum_{i=1}^{m+1}(L_i(p_i)-L(p_i,\tau))^2$$

确定 τ,其中目标函数的最优值是所谓残差平方和。显然,该残差平方和越小,相应经验公式 $L(p,\tau)$越好。对于给定的函数形式 $L(p,\tau)$,上式是一个非线性规划问题,可以使用一般非线性规划方法求解,笔者使用 Levenberg - Marquardt算法 (何光渝, 1993)编程求解上述问题,该算法是一种可靠的非线性最小二乘参数估计方法。下面讨论中为简化记号而约去参数向量 τ。

寻找合适的经验公式 $L(p,\tau)$的研究工作是活跃的研究领域,有关参考文献很多,例如 Chotikapanich (1993), Kakwani等 (1973, 1976), Rasche (1980), Ortega (1991), Schader (1994), Ogwang (1996)等。比较著名的是 Kakwani (1986)给出的如下经验公式:

$$L(p)=p-ap^{\alpha}(1-p)^{\beta} \tag{8}$$

其中约去了参数向量 $\tau=(a,\alpha,\beta)$,可见此公式中含有三个参数,它们应满足 $a>0,\alpha>0,\beta>0$。用 (8)式对很多国家收入分配的洛伦兹曲线进行拟合时,残差平方和往往很小。但此式的缺点是 $p\to0+$时,$L''(p)\to-\infty$,这导致 $L(p)$在 $p=0$的 ε - 邻域内有一负的极小值。Ortega (1991)提出了只含两个参数的著名改进公式以克服这一问题:

$$L(p)=p^{\alpha}(1-(1-p)^{\beta}) \tag{9}$$

并证明当 $\beta\in(0,1]$时,对于任何 $\alpha>0$, $L(p)$满足洛伦兹曲线的条件。我们进一步提出如下三个参数的改进公式:

$$L(p)=p^{\alpha}(1-\omega(1-p)^{\beta}) \tag{10}$$

后面的计算结果显示,对我国的城镇数据,用 (10)式能够得到比 (9)式更理想的拟合结果。计算结果中,城镇收入分配都使用 (10)式进行拟合。由此得到:

定理 2 当 $\beta\in(0,1]$、$\omega\in(0,1)$、$\alpha\geq1$时, (10)式定义的 $L(p)$满足洛伦兹曲线的条件。若 $\omega=1$,则对于任何 $\beta\in(0,1]$及任何 $\alpha>0$, (10)式中 $L(p)$满足洛伦兹曲线的条件 (证明见附录 1)。

定理 3 当使用洛伦兹曲线经验公式 (9)或 (10)时, $r(x)$满足密度函数的条件,即有 $r(x)\geq0$,且 $\int_0^\infty r(x)\,dx=1$, $\int_0^\infty xr(x)\,dx=\mu$ (证明见附录 2)。

三、农村收入分配密度函数的构造

农村收入分配密度函数的构造相对更困难。《中国统计年鉴》中给出了农村纯收入分配的分组数据,其

中含从低到高的收入区间 $(x_0, x_1], [x_1, x_2], \cdots, [x_m, x_{m+1})$ 与各个收入区间上的家庭份额,另给出了总人均纯收入 μ 与家庭平均人口数。显然对这种形式的数据,不能使用第二部分中的方法来构造密度函数。实际计算说明,用下面所述的二次样条函数来逼近密度函数时可以得到比较理想的效果,王祖祥等(2006)利用这一方法对农村贫困问题进行了评估,结果发现这种方法是非常有效的。为完整起见,下面对这一方法进行简单描述。记农村密度函数为 $g(x)$,第 i 个区间 $[x_{i-1}, x_i]$ 内成员份额为 f_i,设 $[x_{i-1}, x_i]$ 上 $g(x)$ 的表达式为:

$$g_i(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2$$

要求 $g_i(x)$ 满足连续性条件 $g_i(x_i) = g_{i+1}(x_i)$ 与导数连续性条件 $g_i'(x_i) = g_{i+1}'(x_i)$,并要求满足分布函数条件 $\int_{x_{i-1}}^{x_i} g_i(x)\,dx = f_i$,以此来确定其中的参数 a_i、b_i、c_i,从而得到 $g(x)$ 在区间 $[x_{i-1}, x_i]$ 上的表达式。对密度函数进行分段逼近的方法并不是新思想,Kakwani(1980)甚至采用密度函数的线性逼近,但线性方法产生的逼近有可能不满足密度函数的条件,且可能振荡得很厉害。

对于任何区间 $[x_{i-1}, x_i]$,若记 $h_i = x_i - x_{i-1}$,按所给出的3个条件,可以得到 $g_i(x)$ 应该满足的三个方程:

$$a_i + b_i h_i + c_i h_i^2 - a_{i+1} = 0$$

$$b_i + 2c_i h_i - b_{i+1} = 0$$

$$a_i h_i + 2b_i h_i^2 + 3c_i h_i^3 = f_i$$

综合 $m+1$ 个区间得到 $3(m+1) \times 3(m+1)$ 阶的线性方程组,解之即得到密度函数的分段二次多项式逼近。显然限制条件:

$$g_1(x_0) = 0,\ g_{m+1}(x_{m+1}) = 0$$

是合理的,加上这两个条件后,方程组可以减少两阶。本文稍后结论都是加上这一限制时得到的。实际数据中并没有给出最后收入区间的右端点 x_{m+1},笔者将 x_{m+1} 取成 Kx_m,例如 $K=100$。计算实践说明取 $K=100$ 与取 $K=1\ 000$ 得出的基尼系数几乎没有差别。

样条拟合方法是逼近论中的著名方法,De Boor(1978)曾对有关理论问题进行过深入广泛的讨论。上述方法容易出现的问题是 $g_1(x)$ 与 $g_{m+1}(x)$ 可能不满足非负性条件,若出现这种问题,则可以将这两段曲线换为其他逼近,例如将它们分别换为帕累托分布:

$$g_1(x) = a(x/x_1)^{\alpha},\ g_{m+1}(x) = (bx+c)(x_m/x)^{\beta}$$

其中,要求 $g_1(x)$ 在 x_1 处满足连续性条件 $g_1(x_1) = g_2(x_1)$ 与密度条件 $\int_{x_0}^{x_1} g_1(x)\,dx = f_1$,而要求 $g_{m+1}(x)$ 在 x_m 处满足连续性条件 $g_{m+1}(x_m) = g_m(x_m)$、密度条件与平均收入条件:$\int_{x_m}^{\infty} g_{m+1}(x)\,dx = f_{m+1}$,$\int_{x_m}^{\infty} x g_{m+1}(x)\,dx = f_{m+1}\mu_{m+1}$,其中 μ_{m+1} 是 $[x_m, \infty)$ 上人口的平均收入。注意到,估计得到 $g_1(x), \cdots, g_m(x)$ 后,即能得到前面各个区间上的平均收入 $\mu_1, \mu_2, \cdots, \mu_m$ 等,由于总平均收入 μ 应满足条件:

$$\mu = f_{m+1}\mu_{m+1} + \sum_{i=1}^{m} f_i \mu_i$$

从中即可确定 μ_{m+1}。这样既保证了产生的近似密度函数满足平均收入条件,也保护了平均收入 μ 这一数据信息。这样处理时区间 $[x_m, \infty)$ 上的密度曲线可能隆起一个小的峰,但观察洛伦兹曲线的图形可以预料,对最高收入区间上的微小摄动对估算基尼系数的影响不大。帕累托分布是国外经济学界在收入分配分析中经常使用的分布,有关参考文献见 Kakwani(1986)等。

四、基尼系数的估算结果

国家统计局每年城乡调查规模甚大,例如2003年两部门调查总户数达到了116 218户,这不能算是小样本调查,因此计算结果应该具有一定的可信度。观察《中国统计年鉴》中的数据可以发现,农村收入分配与城镇收入分配呈现较大的收入差距,例如2003年农村人均纯收入2 500元以下者占总调查户的55.12%,这些人的收入属于城镇困难户的水平,也就是说,城乡两个收入分布的重叠部分相对较小,因此两个收入分配合并形成的收入分配的基尼系数不会很小。

笔者利用1995-2004年的数据,考虑农村纯收入的分配与城镇可支配收入分配时,得到如表1的计算

结果。①

表 1 中国农村、城镇及加总收入分配的基尼系数

年份	农村基尼系数	城镇基尼系数	扩展基尼系数下界	扩展基尼系数	全国基尼系数下界	全国基尼系数
2004	0.314051	0.334857	0.524778	0.540162	0.432313	0.440505
2003	0.322798	0.328452	0.527288	0.543596	0.433887	0.442622
2002	0.312860	0.317530	0.513554	0.530913	0.421706	0.431016
2001	0.310270	0.272488	0.487014	0.502022	0.395540	0.403550
2000	0.303018	0.253133	0.471859	0.485980	0.381615	0.389115
1999	0.290662	0.241247	0.451825	0.465793	0.365374	0.372721
1998	0.281547	0.233619	0.430085	0.445563	0.349721	0.357752
1997	0.288455	0.226117	0.423448	0.439753	0.347059	0.355427
1996	0.305703	0.215235	0.388862	0.421492	0.333491	0.349795
1995	0.305440	0.215035	0.423202	0.441751	0.351250	0.360543

数据来源:农村取纯收入的收入分配,城镇取可支配收入的收入分配。收入分配数据见《中国统计年鉴》(1995 - 2005)。

可见我国城镇人口内部的基尼系数从比农村基尼系数低 9个百分点左右,上升到高于农村基尼系数 2个百分点的水平,10年间增加了近 12个百分点,同时也说明 1995 - 2004年城镇人口之间的收入分配格局发生了巨大变化。与此相反,过去 10年中我国农村人口内部的基尼系数却变化不大,几乎始终在 0.30左右徘徊。一般经济快速发展过程中,收入分配格局往往随之发生变化,改革开放二十多年来,相对于农村地区,我国的经济增长主要发生于城镇部门,可见两部门内部的基尼系数的变化恰好反映了这一点。我国城镇部分的基尼系数虽然增长速度可观,但目前仍处于可以接受的范围内,都没有超过 0.35。观察表 1中最后一列,可见两部门加总基尼系数由 0.35左右逐渐变化到 0.44左右,10年中增加了近 9个百分点,这一增速非常快。目前我国加总基尼系数的值远远超过了警戒水平 0.4。值得注意的是扩展基尼系数的变化,10年间它增加了 10个百分点,且目前达到了 0.54这一比较高的水平,这反映了城镇与农村两部门的收入不平等快速增加,因为扩展基尼系数最大不会超过 1。由 (3)式可见,由于我国目前两部门的基尼系数都不算大,但加总基尼系数却增加甚速,原因之一是城镇内部的基尼系数快速增加,原因之二是城乡两部门之间收入差距的迅速扩大。但由于我国农村人口比重远大于城镇,由 (3)式可见后一原因对基尼系数快速增加的贡献更大。表 1还说明,目前我国的加总基尼系数不会低于 0.43,城镇农村两部门之间的收入不平等不会低于 0.52。

有人怀疑我国农村内部或城镇内部的基尼系数可能很高了,达到了 0.4以上。这种估计可能过高,因为以 2003年为例,农村人口拥有总收入的份额为 $s_1 = 0.31$,城镇人口拥有的份额为 $s_2 = 0.69$,农村与城镇的人口份额分别为 $p_1 = 0.59$与 $p_2 = 0.41$,两部门的平均收入分别为 $\mu_1 = 2\,622$与 $\mu_2 = 8\,472$,按定理 1,当农村与城镇内部的基尼系数都取 0.4时,该年基尼系数不能低于:

$$(p_1 s_1 + p_2 s_2) \times 0.4 + (p_1 s_2 + p_2 s_1)\frac{8472 - 2622}{8472 + 2622}$$

$$= 0.4658 \times 0.4 + 0.5342 \times 0.527 \approx 0.4678$$

这样 2003年的基尼系数应在 0.47以上。由上式可见,正是巨大的城乡收入差距决定了我国的基尼系数必然很大,因为即使在上式中把农村与城镇的基尼系数都换为 0.3,由于城乡收入差距的作用,2003年我国的整体基尼系数也会达到 0.42左右。

表 2中给出了利用 (9)式与 (10)式对城镇数据进行拟合时计算得到的参数,感兴趣的读者可以用《中国统计年鉴》上的数据对表 1中城镇基尼系数进行验证,更重要的是可以将这些参数代入 (9)式或 (10)式而得到我国城镇收入分配的洛伦兹曲线,从而可能对城镇收入分配展开进一步的分析。

首先注意到,表中公式 (9)的参数 α、β都属于区间 (0, 1),因此代入 (9)式后产生的曲线都满足洛伦兹曲线的条件。由定理 2可见,表中公式 (10)的参数使该式也满足洛伦兹曲线的条件。为节约空间,表中残差平方和使用了所谓科学计数法表示,例如 1.4897E - 5表示 1.4897×10^{-5}。可见用 (10)式进行拟合时,残差平方和大约是 (9)式的 1/10, (10)式残差的数量级达到了 10^{-6},这种误差基本上可以忽略。因此,文中基尼系数的准确程度是比较高的。

---

①最近 Chotikapanich等 (2007)给出了利用《中国统计年鉴》上的数据估计我国基尼系数的方法,本文农村与城镇的估算结果与该文差别不大,但本文两部分加总的结果稍大于他们的结果,由公式 (3)可见这种差异应该是人口或平均收入信息的原因,另外,我们这里没有考虑价格指数对收入分配的影响。

表 2　　用公式 (9)与 (10)拟合中国城镇可支配收入分配数据时的参数

	经验公式 (9)			经验公式 (10)			
	α	β	残差平方和	α	ω	β	残差平方和
2004	0.37199	0.636379	1.4897E - 5	1.251904	0.493361	0.502925	4.0589E - 6
2003	0.376215	0.648728	1.6709E - 5	1.258804	0.480384	0.512448	3.5453E - 6
2002	0.368341	0.663516	1.9625E - 5	1.245643	0.473186	0.530251	2.0994E - 6
2001	0.317372	0.716113	7.3157E - 5	1.133969	0.491755	0.626031	1.7760E - 6
2000	0.292581	0.737949	4.0562E - 5	1.169443	0.414664	0.613026	1.7241E - 6
1999	0.267604	0.745446	3.8292E - 5	1.154333	0.398434	0.614574	1.0724E - 6
1998	0.255769	0.752410	2.7179E - 5	1.160203	0.372631	0.605870	1.1798E - 6
1997	0.246467	0.760525	2.3550E - 5	1.159517	0.355854	0.606481	1.8448E - 6
1996	0.225431	0.767584	2.3564E - 5	1.142164	0.345702	0.611357	1.1450E - 6
1995	0.225157	0.767871	1.6191E - 5	1.153334	0.331653	0.597587	1.2967E - 6

注:中国城镇可支配收入分配数据见《中国统计年鉴》(1995 - 2005)。

五、结语

为说明我国目前的收入不平等程度及变化,当农村收入分配取为纯收入分配、城镇收入分配取为可支配收入分配时,计算全国加总洛伦兹曲线上各个十分位点处的值,得到表 3。例如表 3中第 2列是各年中人口份额占 10%的低收入群体所拥有的总收入的份额,其他各列类推。可见人口份额 5%的高收入层所拥有的总收入的份额由 1995年的 14.52%上升到 2004年的 20.37%,人口份额为 10%的高收入层所拥有的总收入的份额由 1995年的 24.78%上升到 2004年的 32.12%。在 2004年,人口份额为 20%的高收入阶层拥有近 50%的总收入。而这些年中低收入阶层所拥有的收入份额则持续下降,人口份额 10%的低收入层所拥有的收入份额由 1995年的 2.34%下降到 2004年的 1.68%,人口份额为 20%的低收入层所拥有的收入份额由 1995年的 6.13%下降到 2004年的 4.66%,2004年中人口份额为 50%的低收入阶层只拥有大约 20%的总收入。而且,除 1996年外的任何年份中,人口份额为 50%的低收入层所拥有的收入都没有超过 25%。在 1995年,高收入端 10%的人口所拥有的收入是低收入端 10%群体的 10.59倍,到 2004年时,这一数字差不多翻了一番,达到了 19.12倍。2004年与 2003年比较时情况有微小改变,2004年低收入端的收入分配要优于 2003年,这可能与 2004年农村税收政策的调整有关,因为据 2005年《中国农村住户年鉴》介绍,2004年农村税收调整使农民收入得到了一定的提高。

表 3　　2004年中国农村城镇加总收入分配的洛伦兹曲线值

	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
2004	1.68	4.66	8.65	13.73	20.13	28.21	38.12	50.58	67.88	79.63
2003	1.63	4.53	8.45	13.46	19.81	27.90	38.13	50.68	68.11	79.95
2002	1.77	4.83	8.91	14.08	20.59	28.78	39.21	51.60	68.88	80.60
2001	1.87	5.14	9.51	15.07	22.02	30.70	41.62	54.39	71.59	82.98
2000	1.97	5.38	9.92	15.66	22.80	31.71	42.86	55.76	72.79	83.87
1999	2.23	5.89	10.65	16.57	23.86	32.85	43.99	57.07	73.71	84.48
1998	2.42	6.27	11.25	17.38	24.88	34.02	45.19	58.25	74.50	84.95
1997	2.35	6.22	11.23	17.41	24.95	34.10	45.28	58.71	74.94	85.29
1996	2.52	6.54	11.69	18.03	25.69	34.72	45.50	58.64	75.21	85.50
1995	2.34	6.13	10.99	16.97	24.31	33.38	44.61	58.56	75.22	85.48

可见,我国 2004年的基尼系数已经达到了 0.44左右,且从 1997年开始,基尼系数以很快的速度增长。① 目前我国正致力于建立和谐社会,而收入不平等的扩大显然与这一目标背道而驰。可见为解决这一问题,控制城乡两部门内的收入不平等固然重要,但更重要的是提高农民的收入,缩小城乡差距。我国正大力开展新农村建设,这将使城乡收入差距得到根本改善,笔者希望十一五规划结束时,重新计算的基尼系数会大大减少。

①但这里需要指出,基尼系数的估算结果与所基于的数据有关,若对于同一个群体进行两次收入分配抽样调查,可能得到相差很远的数据,因而得到完全不同的收入不平等估算结果,因此,过分关注某一种数据来源的估算结果是不必要的。

附录 1:定理 2的证明

证明:显然对于任何 $p\in[0,1]$有 $L(p)\geq 0$,且满足 $L(0)=0$与 $L(1)=1$,还有:

$$L'(p)=\alpha p^{\alpha-1}(1-\omega(1-p)^{\beta})+\omega\beta p^{\alpha}(1-p)^{\beta-1}\geq 0$$

$$L''(p)=\alpha(\alpha-1)p^{\alpha-2}(1-\omega(1-p)^{\beta})+2\alpha\beta\omega p^{\alpha-1}(1-p)^{\beta-1}+\omega\beta(1-\beta)p^{\alpha}(1-p)^{\beta-2}\geq 0$$

定理的第二个结论的证明见 Ortega(1991)。

定理中第一个结论不能有 $\alpha<1$,因为当 $\omega<1$时,对 (10)式有 $\lim\limits_{p\to 0^+}L''(p)=\lim\limits_{p\to 0^+}\alpha(\alpha-1)(1-\omega)p^{\alpha-2}=-\infty$,因此在 $p=0^+$附近 (10)式中 $L(p)$满足 $L''(p)<0$。(9)式的优势是拟合结果一般能满足洛伦兹曲线的条件,而利用 (10)式时,有时可能出现 $\omega>1$,这时在 $p=0$附近有 $L(p)<0$,此时就需要使用其他函数形式,例如可以转而使用 (9)式。笔者对我国 1995 - 2004年的城镇数据进行拟合时,没有出现 $\omega>1$的情形。估计得到了 $L(p)$后,由 (7)式即可得到任何收入数量 x处的密度函数值:对于任何给定的 x,由 (6)式解出相应的 p,代入 (7)式即得密度函数值:

$$r(x)=(\mu L''(p))^{-1}$$

定理 3保证 $r(x)$是满足要求的密度函数。定理 3的证明见附录 2。

附录 2:定理 3的证明

证明:显然 $r(x)\geq 0$成立。在积分中作变量替换 $x=\mu L'(p)$,由 $r(x)$的定义即有:

$$\int_0^{\infty} r(x)\,dx=\int_0^1 dp=1$$

$$\int_0^{\infty} r(x)\,dx=\mu\int_0^1 L'(p)\,dp=\mu(L(p)-L(0))=\mu$$

参考文献:

1. 董静、李子奈:《修正城乡加权法及其应用》,载《数量经济技术经济研究》, 2004(5)。
2. 何光渝:《FORTRAN77算法手册》,北京,科学出版社, 1993。
3. 李实、赵人伟、张平:《中国经济转型与收入分配变动》,载《经济研究》, 1998(4)。
4. 李强、洪大用、宋时歌:《我国社会各阶层收入差距分析》,载《科技导报》, 1995(11)。
5. 胡祖光:《基尼系数理论最佳值及其简易计算公式研究》,载《经济研究》, 2004(9)。
6. 王学力:《个人收入差距的现状、问题和对策》,载《改革》, 2000(6)。
7. 王祖祥:《中部六省基尼系数的估算研究》,载《中国社会科学》, 2006(4)。
8. 王祖祥、范佳强、何耀:《中国农村贫困评估研究》,载《管理世界》, 2006(3)。
9. Chotikapanich, D., 1993. "A Comparison of Alternative Functional Forms for the Lorenz Curve." Economics Letters, Vol 3, pp. 187 - 192.
10. Chotikapanich, D.; Prasada, D. S and Tang, K K, 2007. "Estimating Income Inequality in China Using Grouped Data and the Generalized Beta Distribution." Review of Income and Wealth, Vol 53, pp. 127 - 147.
11. Dagum, C., 1997. "A New Approach to the Decomposition of the Gini Income Inequality Ratio." Empirical Economics, Vol 22, pp. 515 - 531.
12. De Boor, C., 1978. A Practical Guide to Splines. New York: Springer - Verlag.
13. Kakwani, N. C., 1980. Income Inequality and Poverty - Methods of Estimation and Policy Applications. Oxford: Oxford University Press.
14. Kakwani, N. C., 1986. Analyzing Redistribution Policies. London: Cambridge University Press.
15. Kakwani, N. C. and Podder, N., 1973. "On the Estimation of Lorenz Curves from Grouped Observations." International Economic Review, Vol 14, pp. 278 - 292.
16. Kakwani, N. C. and Podder, N., 1976. "Efficient Estimation of Lorenz Curve and Associated Inequality Measures from Grouped Observations." Econometrica, Vol 44, pp. 137 - 148.
17. Ogwang, T. and Rao, U. L. G., 1996. "A New Functional Form for Approximating the Lorenz Curve." Economics Letters, Vol 52, pp. 21 - 29.
18. Ortega, P.; Martín, A.; Fernández, A.; Ladoux, M. and García, A., 1991. "A New Functional Form for Estimating Lorenz Curves." Review of Income and Wealth, Vol 37, pp. 447 - 452.
19. Rasche, R. H.; GaVney, J.; Koo, A. Y. C. and Obst, N., 1980. "Functional Forms for Estimating the Lorenz Curve." Econometrica, Vol 48, pp. 1061 - 1062.
20. Schader, M. and Schmid, F., 1994. "Fitting Parametric Lorenz Curves to Grouped Income Distributions - a Critical Note." Empirical Economics, Vol 19, pp. 361 - 370.

(责任编辑:陈永清)