利用简捷公式计算基尼系数应注意的问题

■程振源
高鸿桢

一、问题的提出

基尼系数是二十世纪初意大利经济学家基尼提出来的。其作用主要是用来定量测度社会收入在社会各集团间分配的平均程度。统计中习惯用如下一个简捷公式近似计算基尼系数:

$$G=\sum_{i=1}^{n}X_iY_i+2\sum_{i=1}^{n-1}X_i(1-V_i)-1$$

式中 X_i,Y_i 分别为分组资料中第 i 组 A_i 的人数比重和收入比重;V_i 为从第 1 组到第 i 组的累计收入比重,即 $V_i=Y_1+Y_2+\cdots+Y_i$,i=1,2,…,n,n 为分组资料中的组数。

从这个公式可以看出,基尼系数的大小受 X_i、Y_i 和 V_i 这三个因素的影响。对于某一给定的资料,各组的人数比重 X_i 和收入比重 Y_i 是确定的。但是,累计收入比重 V_i 则会随着这 n 个组在资料中排列顺序的不同而不同 只有当 $Y_1=Y_2=\cdots=Y_n$ 时,V_i 才与排列顺序无关)。因此,对应不同的排列顺序,由上述公式计算所得到的‘基尼系数’也就可能不同,甚至会出现负数。

例如,设有 A_1,A_2,A_3 和 A_4 四个组,其资料如下:

顺序	A_1	A_2	A_3	A_4	A_1	A_2	A_4	A_3
X_i(%)	20	25	20	35	20	25	35	20
Y_i(%)	15	20	30	35	15	20	35	30
V_i(%)	15	35	65	100	15	35	70	100

由上述公式求得:

对于排列顺序 $A_1A_2A_4A_3$,G=0.1375; 对于排列顺序 $A_1A_2A_4A_3$,G=-0.045。

二、问题的原因

基尼系数是在洛伦兹曲线基础上定义的,为了进一步说明问题,有必要对洛伦兹曲线进行一番考察。

(一)分配曲线

不失一般性,设 $A_1A_2\cdots A_n$ 为某 n 个组在分组资料中的任意一个排列顺序。相应的人数比重排列顺序 $X_1X_2\cdots X_n$,收入比重排列顺序为 $Y_1Y_2\cdots Y_n$;累计人数比重 $U_i=X_1+X_2+\cdots+X_i$,累计收入比重 $V_i=Y_1+Y_2+\cdots+Y_i$。

以累计人数比重 U 为横坐标,以累计收入比重 V 为纵坐标。则在直角坐标系中可得一条曲线,称为收入分配曲线。这 n 个组的一个排列顺序对应于一条分配曲线,其排列顺序共有 n! 条 有的可能会重合)。

(二)收入分配系数

定义 $K_i=Y_i/X_i$ 为 A_i 的收入分配系数。它表示该组占总人数百分之一的人数中分配到的社会收入的百分比。

若 $K_i<1$,则说明组 A_i 分配到的社会收入低于社会平均水平;若 $K_i=1$,则说明其分配到的社会收入等于社会平均水平;若 $K_i>1$,则说明其分配到的社会收入高于社会平均水平。

收入分配系数 K_i 的几何意义是线段 $\overline{L_{i-1}L_i}$ 的斜率 如图一所示)。设 $\overline{L_{i-1}L_i}$ 的倾斜角为 θ_i,则 $K_i=\mathrm{tg}\theta_i$。若 $K_i<1$,则 $\theta_i<45°$;若 $K_i=1$,则 $\theta_i=45°$;若 $K_i>1$,则 $\theta_i>45°$。我们可以用线段 $\overline{L_{i-1}L_i}$ 的倾斜角 θ_i 来近似地表示曲线段 $L_{i-1}L_i$ 的倾斜度。根据正切函数的性质可知,K_i 的值越大,倾斜角 θ_i 越大,因而曲线段 $L_{i-1}L_i$ 的倾斜度也越大。

(三)分配曲线的类型

根据各组收入分配系数的大小,我们可以将分配曲线分成以下几类:

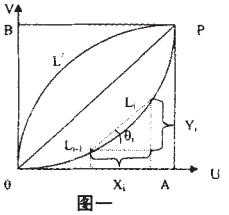
1、若 $K_1<K_2<\cdots\cdots<K_n$,则表明从左至右分配曲线的倾斜度越来越大。此情形下的分配曲线如图一中的 L 所示。

2、若 $K_1>K_2>\cdots\cdots>K_n$,则表明从左至右分配曲线的倾斜度越来越小。此情形下的分配曲线如图一中的 L′所示。

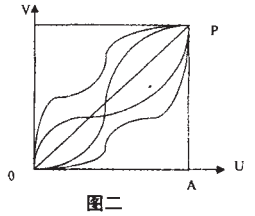
3、若 $K_1=K_2=\cdots\cdots K_n=1$,则可推出 $X_1=Y_1$,$X_2=Y_2$,……,$X_n=Y_n$。即各组的人数比重等于其收入比重,这表示社会收入在各组间的分配绝对平均。此情形下的分配曲线如图一中的线段 $\overline{OP}$ 所示。我们称之为绝对平均线。

4、若 $K_1=K_2=\cdots\cdots=K_{n-1}=0$,$K_n\neq0$ 或 $K_1\neq0$,$K_2=K_3=\cdots\cdots K_n=0$),则可推出 $Y_1=Y_2=\cdots\cdots=Y_{n-1}=0$,$Y_n=100\%$ 或 $Y_1=100\%$,$Y_2=Y_3=\cdots\cdots=Y_n=0$),这表示社会收入被组 A_n 或 A_1)独吞,分配绝对不平均。此情形下的分配曲线如图一中的折线 OAP 或 OBP)所示,我们称之为绝对不平均线。

5、其它类型。包括除以上四种类型以外的其它所有分配曲线。此类型分配曲线的特点是倾斜度有时大有时小,呈单调上升的阶梯状。具体又可分为三类:一类是分配曲线处于绝对平均线的上方;一类是分配曲线处于绝对平均线的下方;还有一类是分配曲线在绝对平均线的上下方摆动 如图二所示)。



图一



图二

综上所述,对于某一给定的资料,各组的排列顺序不同,其对应的分配曲线就有可能不同。根据基尼系数的定义,基尼系数等于洛伦兹曲线与绝对平均线所围面积的两倍。如果将不同排列顺序所对应的不同分配曲线都当成洛伦兹曲线,则由上述简捷公式计算出的‘基尼系数’自然就可能不同,且有正有负。分配曲线若处于绝对平均线的下方,则计算结果为正;若处于绝对平均线的上方,则计算结果为负;若一部分处于绝对平均线的下

产业发展水平与科技创新能力的相关性

■罗发友 刘伶俐 刘友金

产业是国民经济中具有同一性质、承担一定社会经济功能的生产或其它经济社会活动单元构成的、具有相当规模和社会影响的组织体系。除了受政治、经济、文化等因素影响外,产业发展水平更主要地由科技创新能力因素决定。本文基于跨国截面数据,通过建立产业发展水平的典型相关模型,定量判别产业发展水平与科技创新能力因素的相关程度,从而为制定恰当的区域产业发展与科技创新政策提供客观依据。

一、典型相关分析的数学描述

设随机向量 $x=(x_1,x_2,\Lambda,x_p)'$,$y=(y_1,y_2,\Lambda,y_q)'$,x,y 的协方差阵为:

$$\mathrm{cov}\begin{pmatrix} x \\ y \end{pmatrix}=\sum=1\begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix} \quad (1)$$

为了研究两组变量 x 与 y 之间的典型相关关系,考察其线性组合

$$\begin{cases} u=a'x=a_1x_1+a_2x_2+\Lambda a_px_p \\ v=b'y=b_1y_1+b_2y_2+\Lambda b_qy_q \end{cases} \quad (2)$$

在 x,y 及 $\sum$ 给定条件下,即是求 a,b,使 u 与 v 之间的相关系数

$$r=\mathrm{cov}(u,v)/\sqrt{\mathrm{var}(u)\mathrm{var}(v)} \quad (3)$$

达到最大。

标准化随机变量 u 及 v,则有

$$\begin{cases} \mathrm{var}(u)=\mathrm{var}(a'x)=a'\sum_{11}a=1 \\ \mathrm{var}(v)=\mathrm{var}(b'y)=b'\sum_{22}b=1 \end{cases} \quad (4)$$

故 $r=\mathrm{cov}(u,v)=a'\mathrm{cov}(x,y)b=a'\sum_{12}b$ (5)

于是,问题是在(4)式约束下,求 $a\in R^p,b\in R^q$,使得(5)式达到最大。构造 lagrange 函数

$$L=a'\sum_{12}b-\frac{\lambda}{2}(a'\sum_{11}a-1)-\frac{\mu}{2}(b'\sum_{22}b-1) \quad (6)$$

求 L 的一阶偏导数。并令其为 0,得方程组

$$\begin{cases} \partial L/\partial a=\sum_{12}b-\lambda\sum_{11}a=0 \\ \partial L/\partial b=\sum_{21}a-\mu\sum_{22}b=0 \end{cases} \quad (7)$$

利用(4)式,有 $\lambda=a'\sum_{12}b=b'\sum_{21}a=\mu$,与 u、v 之间的相关系数 r 恰相等。代入上式后可求得 a_{jk} 和 b_{jk},并写出各典型变量 uk 和 vk。

对所求得的典型变量,还需对其显著性予以检验,只有通过检验的典型变量才用来进行经济分析。检验统计量

$$Q_j=-[n-j-\frac{1}{2}(p+q+1)\ln[\prod_{i=j}^{k}(1-\lambda_i)^2],\quad k=\min\{p,q\} \quad (8)$$

服从 χ^2 分布,自由度为 f=(p-j+1)(q-j+1)。

二、指标选择与数据整理

(一)指标选择

在国际层次上,'克拉克分类法'把产业划分为三个层次,即农业、工业和服务业。产业发展水平主要体现在生产率上,根据上述产业划分基准,选取如下三个指标作为因变量组,简称'生产率组':①农业生产率(美元/人),以每个农业劳动力创造的相应 GDP 表示,记为 y_1;②工业生产率(美元/人),以每个工人雇员创造的相应 GDP 表示,记为 y_2;③服务业生产率(美元/人),以每个服务业雇员创造的相应 GDP 表示,记为 y_3。考虑到可比性,以上三个指标均采用购买力平价估计值。

影响产业发展水平的科技创新能力表现在多个方面,结合各指标的实际经济意义,确定'科技创新能力组'指标包括:①人均研究与开发支出经费(百万美元/人),按现行价格和汇率计算的人均额,记为 x_1;②信息技术熟练工人的可获得性指数,指合格的信息技术雇员在劳

方,另一部分处于绝对平均线的上方,当下方面积大于上方面积时,则计算结果为正,反之为负。

三、各组排列顺序的选择

对于同一资料,基尼系数应该是唯一的。因此,各组在资料中的排列顺序不能随意排列。笔者从一些公开发表的文章中看到,不少学者是将各组按收入比重 Y_i 从小到大排列的。但笔者认为,正确的排列顺序应该是将各组按收入分配系数 K_i 从小到大排列。即要求 $K_1<K_2<\cdots<K_n$。理由是:(1)基尼系数是定义在洛伦兹曲线基础之上的,而洛伦兹曲线是一条特殊的分配曲线。只有这种排列顺序所对应的分配曲线才具有我们常见的洛伦兹曲线的形状;(2)这种排列顺序所对应的分配曲线(即洛伦兹曲线)便于用抛物线或指数曲线来拟合,且拟合误差较小;(3)根据这种排列顺序,由上述公式计算出的结果不会出现负数。而如果将各组按 $Y_1<Y_2<\cdots<Y_n$ 的顺序排列,则不能保证 $K_1<K_2<\cdots<K_n$。只有当 $X_1=X_2=\cdots=X_n$ 时,按收入比重从小到大排列与按分配系数从小到大排列才没有区别。

各组在资料中的顺序除了按 $K_1<K_2<\cdots<K_n$ 排列之外,按 $K_1>K_{21}>\cdots_1>K_n$ 的顺序排列也是可以的(其对应的分配曲线如图一中的 L′所示)。不过此时计算基尼系数的简捷公式应为:

$$G=\sum_{i=1}^{n}X_iY_i+2\sum_{i=2}^{n}X_iV_{i-1}-1$$

对于同一资料,由上述两个简捷公式计算出的结果是相等的,即不难证明

$$G=\sum_{i=1}^{n}X_i\cdot Y_i+2\sum_{i=1}^{n-1}X_i(-V_i)-1$$

$$=\sum_{j=1}^{n}X_j'\cdot Y_j'+2\sum_{j=2}^{n}X_j'\cdot Y_{j-1}'-1$$

(其中 j=n-(i-1))

(作者单位/厦门大学计统系)

(责任编辑/刘智伟)