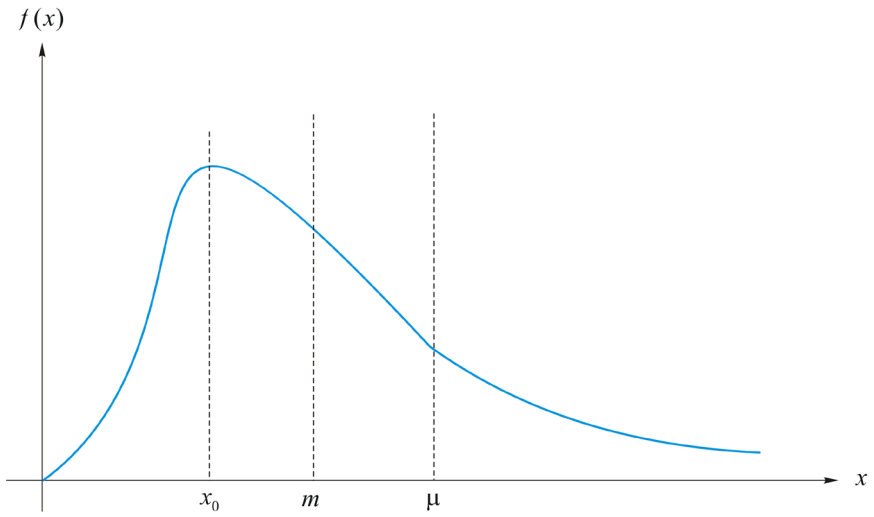
2013 年全国研究生数学建模竞赛 E 题

# 中等收入定位与人口度量模型研究

居民收入分配关系到广大民众的生活水平，分配公平程度是广泛关注的话题。其中中等收入人口比重是反映收入分配格局的重要指标，这一人口比重越大，意味着收入分配结构越合理，称之为“橄榄型”收入分配格局。在这种收入分配格局下，收入差距不大，社会消费旺盛，人民生活水平高，社会稳定。一般经济发达国家都具有这种分配格局。我国处于经济转型期，收入分配格局处于重要的调整期，“橄榄型”收入分配格局正处于形成阶段。因此，监控收入分配格局的变化是经济社会发展的重要课题，例如需要回答，与前年比较，去年的收入分配格局改善了吗？改善了多少？可见实际上需要回答三个问题：什么是“橄榄型”收入分配格局？收入分配格局怎样的变化可以称之为改善？改善了多少？直观上，中间部分人口增加，则收入分配格局向好的方向转化。于是基本问题回答什么是中间部分。

一个国家的收入分配可以用统计分布表示，图 1 是某收入分配的密度函数 $f(x)$，其中 $x \geq 0$ 表示收入(仅考虑正的收入)，$x_0$ 是众数点，$m$ 是中位数点，$\mu$ 是平均收入。收入分配经验分析说明，收入分配曲线一般是所谓正偏的，即峰值点向左偏，右端拖一个长尾巴，且通常有

$$x_0 < m < \mu \tag{1}$$




记对应的分布函数为 $F(x)$，则 $p = F(x)$ 表示收入低于或等于 $x$ 的人口比例。由于 $F(m) = 1/2$，(1)式意味着收入大于或等于平均收入的人口一定不到半数，因此是少数。

记收入低于或等于 $x$ 的人口群体拥有收入占总收入的比例为 $L(p)$，则应有

$$L(p)=\frac{1}{\mu}\int_0^x tf(t)\mathrm{d}t,\quad p=F(x) \tag{2}$$

$L(p)$ 称之为收入分配的洛伦兹曲线。显然，如果 $L_1(p)$ 与 $L_2(p)$ 是两个不同收入分配的洛伦兹曲线，若对任何 $p\in(0,1)$ 都有 $L_1(p)\geq L_2(p)$，则 $L_1(p)$ 对应的收入分配显然更优，因为在 $L_1(p)$ 中，任何低收入端人口拥有的总收入比例更大。下图中红色曲线是某收入分配的洛伦兹曲线。

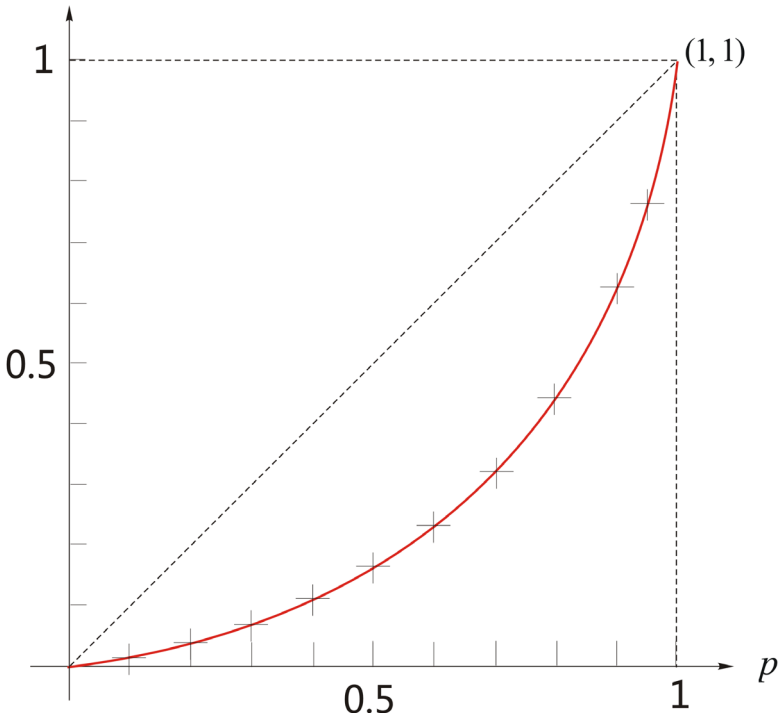



图 1

其中横轴表示人口比例，纵轴表示总收入比例。显然，图中曲线位置越高，所代表的收入分配越平等。其中 45° 线可以理解为平等收入线，这时，任何低收入端人口比例为 $p$ 的人口拥有的总收入比例也是 $p$，从而必定是完全平等的收入分配。因此定义 45° 线与 $L(p)$ 之间面积的 2 倍为基尼系数。于是基尼系数定义为

$$G=1-2\int_0^1 L(p)\mathrm{d}p \tag{3}$$

$L(p)$ 与 $f(x)$ 具有关系

$$L'(p)=\frac{x}{\mu} \tag{4}$$

$$f(x)=\frac{1}{\mu L''(p)} \tag{5}$$

其中 $p=F(x)$。记 $F(x)$ 的反函数为 $F^{-1}(p)$，则洛伦兹曲线可以表示为

$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(q)\mathrm{d}q$$

实践中通过入户调查获得家庭收入与消费等数据，如果可以得到这类数据，则可以使用例如 Kernel 法估计收入分配的统计分布。我国统计部门也进行这种调查，但数据不对外公开，而只是在统计年鉴上发布所谓的分组数据(世界上很多国家也如此)，这种数据的完整形式为

$$\left(p_i, x_i/\mu\right),\ \ i=1,2,\cdots,n \tag{6}$$

$$\left(p_i, L_i\right),\ \ i=1,2,\cdots,n \tag{7}$$

其中 $x_i$ 是收入区间点，满足 $0\le x_1<x_2<\cdots<x_n<x_{n+1}$，通常 $x_{n+1}$ 理解为充分大的正数。$n$ 通常不大，例如 $n=10$。很多国家只提供(7)式描述的数据。经济学界只能利用这种稀疏的信息进行收入分配分析。记 $p_0=0$，则 $[x_i,x_{i+1})$ 中人口比例为 $p_i-p_{i-1}$。例如图 1 中“+”中标出的点表示了形如(7)的数据点，其中 $p_i=i/10$，$i=1,2,\cdots,9$，最后的点是 $p_{10}=0.95$。如果收入分配的真实洛伦兹曲线为 $l(p)$，且若 $l'(p)$ 存在，则(6)表示的是 $l'(p)$ 曲线上的坐标点，即 $l'(p_i)=x_i/\mu$；(7)表示 $l(p)$ 曲线上的点，即 $l(p_i)=L_i$。

经济学界采用所谓的洛伦兹曲线模型 $L(p,\tau)$ 拟合上述数据(7)，其中 $\tau$ 是一组参数，使用非线性最小二乘法求解

$$\min\sum_{i=1}^{n}\left(L(p_i,\tau)-L_i\right)^2 \tag{8}$$

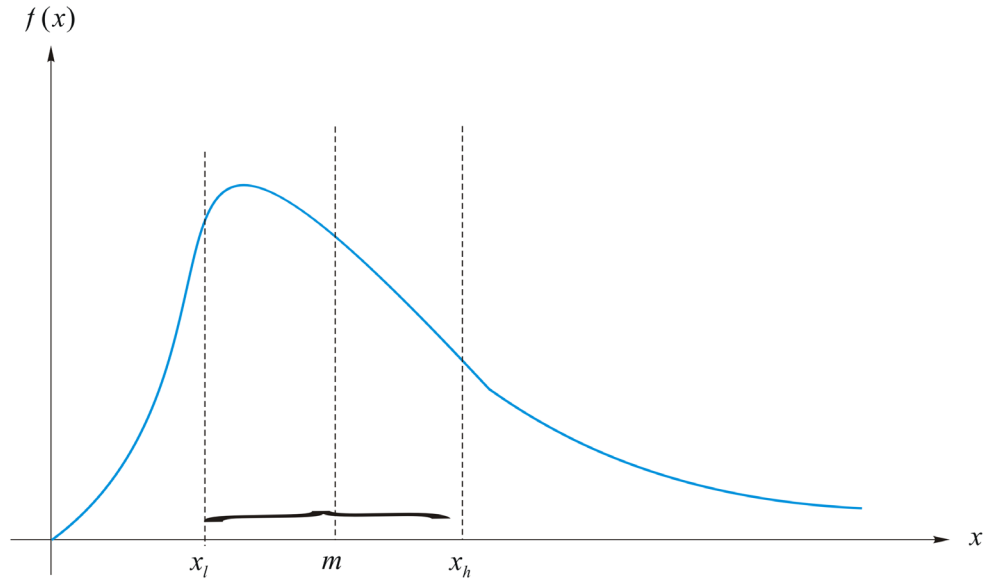
确定其中参数向量 $\tau$ 的估计值 $\hat{\tau}$，然后用 $L(p,\hat{\tau})=\hat{L}(p)$ 作为近似的洛伦兹曲线来进行收入分配分析，显然，这时就能通过(4)、(5)式确定相应的统计密度与分布的估计。$L(p,\tau)$ 是定义在 $[0,1]$ 区间上、取值于 $[0,1]$ 区间的函数，满足

$$L(0,\tau)=0,\ \ L(1,\tau)=1,\ \ L'(p,\tau)\ge 0,\ \ L''(p,\tau)\ge 0 \tag{9}$$

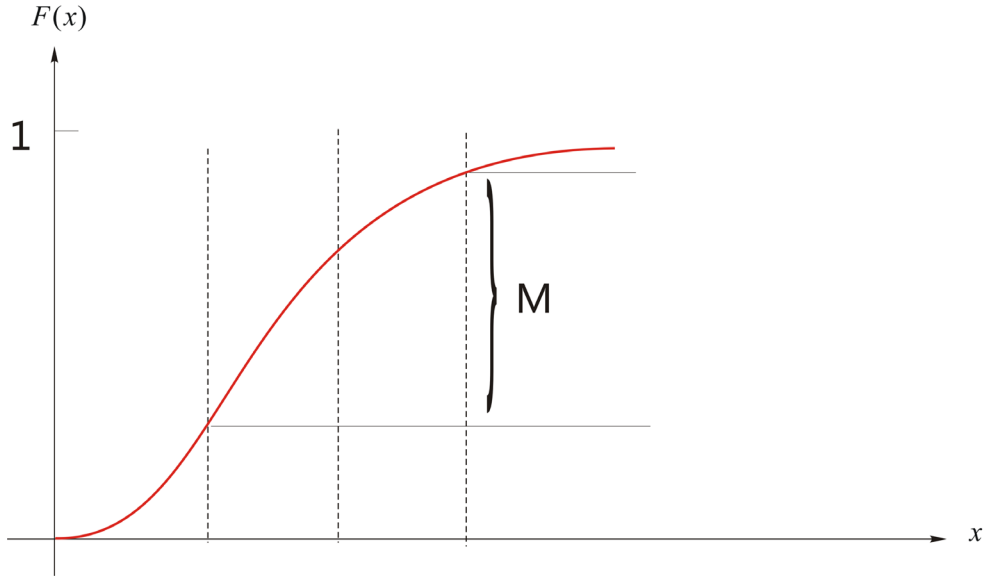
即 $L(p,\tau)$ 在 $[0,1]$ 上是凸增函数。文献中常常略去参数 $\tau$ 以求表述简练。

也可以使用其他方法(例如多项式、样条函数逼近)来确定洛伦兹曲线，但实践证明使用洛伦兹曲线模型是比较理想的方法之一，有关洛伦兹曲线模型的最近

文献见参考文献[3]。经济理论中提出的另一种方法是使用经验分布拟合分组数据而直接形成收入分配的近似分布，有关参考文献见[1]。




(A)




(B)

图 2

经济理论界考虑取收入落在中位收入 $m$ 的一个范围内的人口为中等收入人口，可以视这种方法为“收入空间法”。例如图 2(A)，取其中收入属于 $(x_l, x_h)$ 中的人口为中等收入人口，这时中等收入人口比例 $M$ 显然等于 $F(x_h)-F(x_l)$，见图 2(B)。显然，这种方法中 $x_l$ 与 $x_h$ 的取法具有任意性，由于经济进步，通货膨胀等因素的影响，收入的区间是变化的，更多的情形是所有人口的收入都提高了，

即全社会的收入区间右移，可见 $x_l$ 与 $x_h$ 的任意性使纵向比较各年的中等收入人口时出现困难。

另一种方法可以视为“人口空间法”，即选择 $F(m)=1/2$ 邻近的一个范围为中等收入人口，例如取范围 $p_1=20\%$到 $p_2=80\%$，当然，按定义，中等收入人口比例已经取定为 60%。再用此 60%的人口所拥有的收入占总收入的比例来描述中等收入人口的状态，此时中等收入人口的收入范围 $[x_l,x_h]$ 当然容易算得。例如当范围取为 20%到 80%时，中等收入人口的状态即定义为

$$S=L(0.8)-L(0.2)=\frac{1}{\mu}\int_{0.2}^{0.8}F^{-1}(p)\mathrm{d}p$$

注意到平均收入为

$$\mu=\int_0^1 F^{-1}(p)\mathrm{d}p$$

即图 3 中 $F(x)$ 左侧区域的面积，而 $S$ 是图中淡蓝色区域的面积。

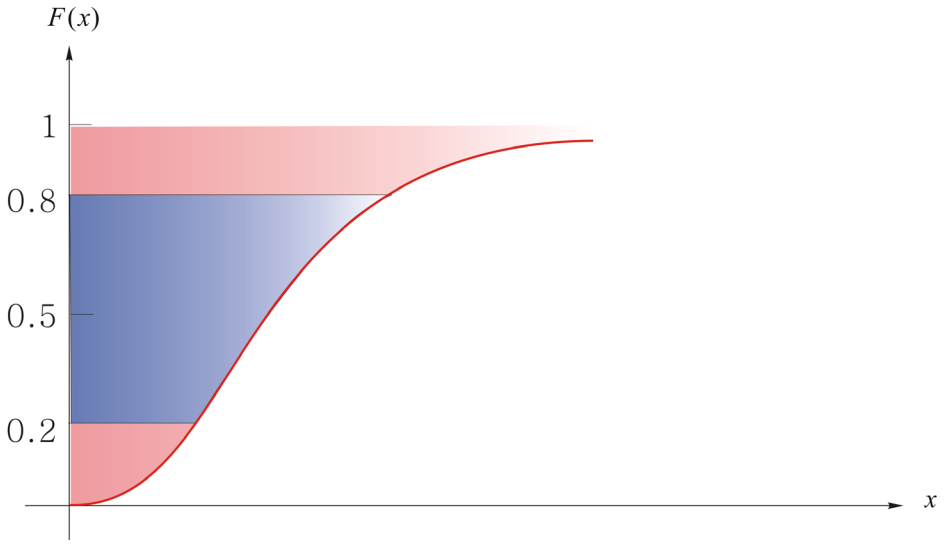



图 3

[2]讨论了两种方法的缺陷。第一种方法是前面提到的任意性，再考虑第二种方法。这种方法似乎有道理，例如经济发展、收入增加导致所有人口的收入都右移时，总是取中间的 60%进行纵向比较似乎总是可行的。设收入分配是 $[10000,30000]$ 上的均匀分布，这时中位收入是 $m=20000$。此时，中间 60%人口拥有总收入的 60%，收入范围为 14000 到 26000。考虑收入分配发生了变化，变成了 $[0,40000]$ 上的均匀分布，这时收入范围拉大了，低端人口收入下降了，高

端收入人口收入增加了，直观上两极分化扩大了，也即中等收入人口应该是下降了，但按第二种方法，中间 60%的人口拥有的总收入比例仍是 60%。这与经济直观不符。

中等收入人口的多少与两极分化(polarization)的程度有关，所谓两极分化，用密度函数表示时，例如严重右偏且厚尾，也即中间部分空洞化。两极分化与收入不平等(inequality)是不同的概念，文献[2]对这两个概念进行了准确阐述。 [2]建立了一种指数，这种指数说明两极分化的大小或严重程度，该指数扩大意味着两极分化严重了，这时表示中等收入人口缩小了。反之若该指数缩小了，则意味着中等收入人口扩大了。但该文献并没有给出测算中等收入人口比例大小的方法。

为此，需要研究中等收入定位与人口度量问题，请你根据表一中给出的分组数据，用数学模型研究给出的问题。

表一：收入分配分组数据

| $x_j$ | $x_{j+1}$ | $f_j$ | $p_j$ | $L_j$ |
|---|---|---|---|---|
| 0.00 | 999.00 | 0.0780 | 0.0780 | 0.0059 |
| 1000.00 | 1499.00 | 0.0560 | 0.1340 | 0.0165 |
| 1500.00 | 1999.00 | 0.0420 | 0.1760 | 0.0276 |
| 2000.00 | 2499.00 | 0.0470 | 0.2230 | 0.0436 |
| 2500.00 | 2999.00 | 0.0420 | 0.2650 | 0.0611 |
| 3000.00 | 3499.00 | 0.0440 | 0.3090 | 0.0828 |
| 3500.00 | 3999.00 | 0.0410 | 0.3500 | 0.1061 |
| 4000.00 | 4999.00 | 0.0860 | 0.4360 | 0.1647 |
| 5000.00 | 5999.00 | 0.0920 | 0.5280 | 0.2413 |
| 6000.00 | 6999.00 | 0.0880 | 0.6160 | 0.3279 |
| 7000.00 | 7999.00 | 0.0800 | 0.6960 | 0.4188 |
| 8000.00 | 8999.00 | 0.0650 | 0.7610 | 0.5024 |
| 9000.00 | 9999.00 | 0.0520 | 0.8130 | 0.5772 |
| 10000.00 | 11999.00 | 0.0780 | 0.8910 | 0.7071 |
| 12000.00 | 14999.00 | 0.0560 | 0.9470 | 0.8216 |
| 15000.00 | 24999.00 | 0.0430 | 0.9900 | 0.9453 |
|  | 25000.00 | 0.0100 | 1.0000 | 1.0000 |

表中$[x_j,x_{j+1}]$是收入区间，单位为元，$f_j$是该区间内的人口比例，$p_j$是$[0,x_{j+1}]$中人口比例，$L_j$是$[0,x_{j+1}]$中人口拥有的总收入比例，因此$(p_j,L_j)$是洛伦兹曲线上的点，其中 25000 以上人口比例为 1%。总平均收入$\mu=6603$元。

请研究如下问题：

一．构造满足(9)式的新模型 $L(p,\tau)$，使得能很好的拟合上述分组数据、反映经济规律。

例如文献[3]证明

$$L(p)=\left(1-(1-p)^{\beta_1}\right)^{\alpha}\left(1-(1-p)^{\beta_2}\right)^{\beta}$$

$$\beta_1,\beta_2\in(0,1]，\ \alpha\geq 0，\ \beta\geq 0，\ \alpha+\beta\geq 1 \tag{10}$$

满足条件(9)。该文中还提出了其他一些模型，并说明利用这些模型时，产生的估计结果优于密度函数的 Kernel 估计法。请在现有参考文献中(文献[4]的参考文献部分列出了大部分有关的文献)找出至少 10 种模型，与你们提出的模型进行比较。通过比较，说明你们的模型不差。

提示：可以搜集到现成的无约束非线性最小二乘计算程序，利用参数变换对类似(10)的条件进行变换，将约束非线性最小二乘问题化为无约束的。如果 $L(p,\tau)$ 是你们找到的模型，分组数据是 $\{(p_i,L_i)\}_{i=1}^{n}$，$\hat{\tau}$ 是你们求得的 $\tau$ 的估计，拟合精度的好坏可以采用以下三种标准进行比较。

均方误差(MSE, mean squared error )：

$$\frac{1}{n}\sum_{i=1}^{n}\left[L(p_i,\hat{\tau})-L_i\right]^2$$

平均绝对误差(MAE, mean absolute error)：

$$\frac{1}{n}\sum_{i=1}^{n}\left|L(p_i,\hat{\tau})-L_i\right|$$

最大绝对误差(MAS, maximum absolute error)

$$\max_{1\leq i\leq n}\left|L(p_i,\hat{\tau})-L_i\right|$$

注意，本题中最好能构造新模型，而不是通过简单处理(例如加权)文献中的已有模型而得到的模型。

二．研究可否改进上述提到的收入空间法，这时需要研究确定中等收入的范围、中等收入人口的范围的科学方法，以克服中等收入区间取法的任意性；研究可否改进上述提到的人口空间法，例如研究在各年中 $p_1$ 与 $p_2$ 取不同的值时，纵向比较各年中等收入人口与收入的变动的方法。

提示：目前经济理论界将中等收入人口定义为中位收入附近的人口，于是

若中间部分比前一年隆起得更高，则认为中等收入人口扩大了；若两边人口扩大了，则中等收入人口下降了。所提出的原理与模型应与这一直观相符。其他有关价值取向方面的示例性提示见问题四。

三．利用最后表二~表五所附 A, B 两个地区前后两个不同年份的收入分配分组数据，请研究：(1) 对各地区、各年份的中等收入的数量(或范围)、中等收入人口的数量或范围进行定量描述，说明中等收入人口的变化趋势；(2)比较两个地区的中等收入人口、收入等变化情况。

四．除二题中所述方法外，提出中等收入人口的定义、原理及经济学意义，并提出与之相应的中等收入人口的测算方法、模型或指数，说明其经济学意义。

提示：所提出的方法应满足普遍的价值判断或价值取向，也应反映经济规律。例如 Sen(见参考文献[5])在构造贫困指数 $p(z)$ 时采用的方法，$p(z)$ 是一数量，贫困越严重 $p(z)$ 越大。这一指数之所以有用，正是它反映经济规律，满足普遍的价值判断。**这种贫困指数的构造方法与本题没有关系，但请参考其中的思想。**

设 $z$ 是贫困线，Sen 先规定 $p(z)$ 应满足以下两个公理(axiom)，这两种公理实际上是经济规律方面的要求：

单调性：贫困线以下人口增加时，$p(z)$ 增加；

转移性：从贫困线以下任何人处转移收入给比他富有的人时，$p(z)$ 增加；

记 $g_i = z - y_i$，$y_i$ 是第 $i$ 个人的收入。$g_i$ 是所谓的贫困缺口，贫困线以下人口的缺口为正，否则为负。设整个社会的收入分配为 $Y$，记 $S(x)$ 是收入低于 $x$ 的人口集合，取贫困指数为以下加权和

$$Q(x) = A(z,Y)\sum_{i\in S(x)} g_i v_i(z,Y)$$

其中 $v_i(z,Y)$ 是非负权数，$A(z,Y)$ 是非负规范化因子。$Q(x)$ 是一种加权贫困缺口，定义在 Z 给定下全社会最大化贫困缺口为贫困指数，即取 $p(z)=Q(z)$ 为贫困指数。要求权数分配满足所谓的相对公平条件：记第 $i$ 个成员的福利水平为 $W_i(Y)$，则 $W_i(Y)<W_j(Y)$ 时，取 $v_i(z,Y)>v_j(z,Y)$。可见这是价值取向方面的条件。

另外还加上几种技术性的公理，Sen 最后推导出一种目前广泛使用的贫困指数(见参考文献[5])。

表二：收入分配分组数据(地区 A，年份之一)

| $x_j$ | $x_{j+1}$ | $p_j$ | $L_j$ |
|---|---|---|---|
| 0.00 | 2228.28 | 0.10 | 0.0250 |
| 2228.28 | 3066.03 | 0.20 | 0.0673 |
| 3066.03 | 3790.18 | 0.30 | 0.1221 |
| 3790.18 | 4519.24 | 0.40 | 0.1882 |
| 4519.24 | 5254.75 | 0.50 | 0.2663 |
| 5254.75 | 6166.38 | 0.60 | 0.3569 |
| 6166.38 | 7273.48 | 0.70 | 0.4631 |
| 7273.48 | 8813.52 | 0.80 | 0.5901 |
| 8813.52 | 11424.93 | 0.90 | 0.7485 |
| 11424.93 | 14171.91 | 0.95 | 0.8493 |

平均：6281.34 元

表三：收入分配分组数据(地区 A，年份之二)

| $x_j$ | $x_{j+1}$ | $p_j$ | $L_j$ |
|---|---|---|---|
| 0.00 | 3081.27 | 0.10 | 0.0241 |
| 3081.27 | 4199.72 | 0.20 | 0.0651 |
| 4199.72 | 5272.06 | 0.30 | 0.1187 |
| 5272.06 | 6383.72 | 0.40 | 0.1843 |
| 6383.72 | 7461.83 | 0.50 | 0.2623 |
| 7461.83 | 8751.34 | 0.60 | 0.3532 |
| 8751.34 | 10294.02 | 0.70 | 0.4601 |
| 10294.02 | 12500.51 | 0.80 | 0.5865 |
| 12500.51 | 16362.67 | 0.90 | 0.7468 |
| 16362.67 | 20288.83 | 0.95 | 0.8488 |

平均：8890.21 元

表四：收入分配分组数据(地区 B，年份之一)

| $x_j$ | $x_{j+1}$ | $p_j$ | $L_j$ |
|---|---|---|---|
| 0.00 | 8465.55 | 0.10 | 0.0427 |
| 8465.55 | 10293.33 | 0.20 | 0.0978 |
| 10293.33 | 11770.00 | 0.30 | 0.1630 |
| 11770.00 | 13173.47 | 0.40 | 0.2367 |
| 13173.47 | 14422.27 | 0.50 | 0.3180 |
| 14422.27 | 16246.88 | 0.60 | 0.4084 |
| 16246.88 | 18510.11 | 0.70 | 0.5108 |
| 18510.11 | 21794.50 | 0.80 | 0.6290 |
| 21794.50 | 26918.59 | 0.90 | 0.7713 |
| 26918.59 | 34375.61 | 0.95 | 0.8596 |

平均：16938.46 元

表五：收入分配分组数据(地区 B，年份之二)

| $x_j$ | $x_{j+1}$ | $p_j$ | $L_j$ |
|---|---|---|---|
| 0.00 | 11062.50 | 0.10 | 0.0411 |
| 11062.50 | 13531.18 | 0.20 | 0.0970 |
| 13531.18 | 15472.69 | 0.30 | 0.1622 |
| 15472.69 | 17599.77 | 0.40 | 0.2369 |
| 17599.77 | 19814.62 | 0.50 | 0.3210 |
| 19814.62 | 22681.13 | 0.60 | 0.4163 |
| 22681.13 | 25818.75 | 0.70 | 0.5249 |
| 25818.75 | 29848.37 | 0.80 | 0.6499 |
| 29848.37 | 35288.50 | 0.90 | 0.7948 |
| 35288.50 | 42150.00 | 0.95 | 0.8804 |

平均：22228.53 元

## 参考文献


[1] Chotikapanich, D., D. S.P. Rao, and K.K. Tang, 2007. Estimating income inequality in China using grouped data and the generalized Beta distribution. The Review of Income and Wealth 53, 127-47.

[2] Foster, J.E. and M.C. Wolfson, 2009. Polarization and the decline of the middle class: Canada and the U.S. Journal of Economic Inequality 8, 247-273.

[3] Wang, Z.X., Y-K Ng, and R. Smyth, 2011. A general method for creating Lorenz curves. The Review of Income and Wealth 57, 561-582.

注：本文中有关定理的完整证明可通过 EcoPapers 下载(在谷歌中键入 EcoPapers 与文章名称即可搜索到)，文章名为：A general method to create Lorenz curves.

[4] Wang, Z.X. and R. Smyth, 2013. A hybrid method for creating Lorenz curves with an application to measuring world income inequality.

注：本文可以通过 EcoPapers 下载。

[5] Sen, A., 1976. Poverty: An ordinal approach to measurement. Econometrica 44, 219-232.