文章编号:1672 -058X(2012)08 -0018 -04

基于距离判别法的中国经济社会发展水平研究*

李 晋 燕
(重庆师范大学 数学学院,重庆 401331)

摘 要:运用多元统计中的距离判别法,对中国部分省市经济社会发展水平进行了研究,根据 2011 年中国人类发展指数报告的统计数据,先对 10 个省市进行距离判别法判别,得出了判别函数,然后选取 6 个省市进行判别分析,判别归类与原统计资料完全吻合,有利于对其他省份发展水平鉴定工作的展开。

关键词:距离判别法;F 检验;发展水平
中图分类号:O212.4 **文献标志码**:A

人文发展指数是经济社会发展水平的一项综合指标,对于衡量地区发展状况具有重要意义。随着市场经济的不断深化与发展,我国各省市的经济社会发展水平在不同程度上发生了变化,运用多元统计的距离判别法对经济社会发展水平进行评价。经济社会水平由 3 个指标构成:预期寿命、成人识字率和人均 GDP 的对数,这 3 个指标分别反映了人的长寿水平、知识水平和生活水平。

1 数据的处理

从 2011 年中国人类发展指数报告[1]中选取高发展水平 5 个、中等发展水平 5 个作为两组样品,另选 6 个省份作为待判样品作距离判别分析。

表 1 中国人类发展指数统计报告

类别	序号	省份名称	寿命指数 x_1	教育指数 x_2	收入指数 x_3
第一类(高发展水平省份)	1	天津	0.832	0.962	0.833
	2	浙江	0.828	0.907	0.787
	3	江苏	0.815	0.921	0.776
	4	广东	0.805	0.96	0.768
	5	辽宁	0.806	0.964	0.737
第二类(中等发展水平省份)	6	河北	0.792	0.951	0.687
	7	福建	0.793	0.898	0.731
	8	河南	0.776	0.927	0.659
	9	湖北	0.768	0.923	0.661
	10	湖南	0.761	0.942	0.64
待判样品	11	山东	0.815	0.921	0.746
	12	重庆	0.779	0.924	0.645
	13	陕西	0.751	0.919	0.647
	14	四川	0.770	0.899	0.618
	15	江西	0.733	0.936	0.612
	16	安徽	0.781	0.860	0.608

收稿日期:2011 -11 -01;**修回日期**:2011 -12 -28.
* **基金项目**:重庆自然科学基金(CSTC2009BB2056).
作者简介:李晋燕(1986-),女,山西长治人,硕士研究生,从事金融系统分析.

变量个数 $p=3$,两类总体分别有 5 个样品,另有 6 个待判样品,假定两总体协差阵相等。

2　距离判别法

距离判别法是多元统计中判别分析[2]的一种统计方法。首先根据已知分类的数据,分别计算各类的重心,判别准则对任给的一次观测通过它与第几类的重心距离最近来认定它属哪类。

(1) 计算两类样本均值。

$$\overline{X}^{(1)}=\begin{bmatrix}0.817\ 2\\0.942\ 8\\0.780\ 2\end{bmatrix},\overline{X}^{(2)}=\begin{bmatrix}0.778\ 0\\0.928\ 2\\0.675\ 6\end{bmatrix}$$

(2) 计算样本协差阵。

从而得

$$S_1=\sum_{\alpha=1}^{5}(X_\alpha^{(1)}-\overline{X}^{(1)})(X_\alpha^{(1)}-\overline{X}^{(1)})'=\begin{pmatrix}0.000\ 614\ 8 & -0.000\ 501\ 8 & 0.001\ 496\ 8\\-0.000\ 501\ 8 & 0.002\ 870\ 8 & -0.000\ 264\ 0\\0.001\ 496\ 8 & -0.000\ 263\ 8 & 0.004\ 866\ 8\end{pmatrix}\tag{1}$$

类似地

$$S_2=\sum_{\alpha=1}^{5}(X_\alpha^{(2)}-\overline{X}^{(2)})(X_\alpha^{(2)}-\overline{X}^{(2)})'=\begin{pmatrix}0.000\ 814 & -0.000\ 314 & 0.001\ 775\\-0.000\ 314 & 0.001\ 650\ 8 & -0.001\ 808\ 6\\0.001\ 775 & -0.001\ 808\ 6 & 0.004\ 955\ 2\end{pmatrix}$$

经计算

$$S=S_1+S_2=\begin{pmatrix}0.001\ 428\ 8 & -0.000\ 815\ 8 & 0.003\ 271\ 8\\-0.000\ 815\ 8 & 0.004\ 521\ 6 & -0.002\ 072\ 4\\0.003\ 271\ 8 & -0.002\ 072\ 4 & 0.009\ 822\ 0\end{pmatrix}$$

$$\hat{\sum}=\frac{1}{n_1+n_2-2}(X_1+S_2)=\begin{pmatrix}0.000\ 178\ 6 & -0.000\ 101\ 98 & 0.000\ 408\ 98\\-0.000\ 101\ 98 & 0.000\ 565\ 2 & -0.000\ 259\ 05\\0.000\ 408\ 98 & -0.000\ 259\ 05 & 0.001\ 227\ 75\end{pmatrix}$$

$$\hat{\sum}^{(-1)}=\begin{pmatrix}238\ 877.019\ 6 & 733.\ 648\ 527\ 8 & -779\ 8.958\ 92\\733.648\ 527\ 8 & 1\ 981.246\ 593 & 173.646\ 389\ 6\\-7\ 798.958\ 92 & 173.646\ 389\ 6 & 344\ 9.074\ 581\end{pmatrix}\tag{2}$$

(3) 求线性判别函数 $W(X)$[4]。

解线性方程组

$$\hat{\sum}a=(\overline{X}^{(1)}-\overline{X}^{(2)})\tag{3}$$

$$a = \hat{\sum}^{(-1)}(\overline{X}^{(1)} - \overline{X}^{(2)}) = (130.919\ 332\ 8 \quad 75.848\ 634\ 9 \quad 57.589\ 248\ 75)'$$

$$W(x) = a'(X - \overline{X}) = a'\left[X - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)})\right] =$$

$$130.919\ 332\ 8x_1 + 75.848\ 634\ 9x_2 + 57.589\ 248\ 75x_3 - 217.296\ 900\ 2 \tag{4}$$

(4) 判别函数的检验。

①对已知类别的样品判别分类。对已知类别的样品(通常称为训练样品)用线性判别函数进行判别归类,结果如表2,全部判对。

表2 原类和判别类比较

样品	判别函数 $W(X)$ 的值	原类号	判别类别
1	12.566 247 95	1	1
2	5.221 784 75	1	1
3	3.948 235 875	1	1
4	5.136 425 008	1	1
5	3.785 470 968	1	1
6	-1.912 896 32	2	2
7	-3.268 025 99	2	2
8	-7.440 472 93	2	2
9	-8.676 043 56	2	2
10	-9.360 729 86	2	2

②对判别效果作检验[5]。所谓判别效果的检验就是检验两个正态总体的均值向量是否相等,如果不存在显著差异,则判别意义不大。

检验统计量[6]为:

$$F = \frac{(n_1 + n_2 - 2) - p + 1}{(n_1 + n_2 - 2)p}T^2 \sim F(p, n_1 + n_2 - p - 1) \tag{5}$$

其中

$$T^2 = (n_1 + n_2 - 2)\left[\sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{X}^{(1)} - \overline{X}^{(2)})'S^{-1} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{X}^{(1)} - \overline{X}^{(2)})\right] \tag{6}$$

将上边计算结果代入统计量后可得:

$$F = 7.664\ 526 > F_{0.05}(3,6) = 4.76$$

故在检验水平下,两个总体间差异显著,判别函数有效。

(5) 对判别样品判别归类[7],结果如表3。

表3 判别样品归类表

样品号	省份名称	判别函数 $W(X)$ 的值	判别类别
11	山东	2.220 557 25	1
12	重庆	-8.081 510 86	2
13	陕西	-12.011 341 85	2
14	四川	-12.710 935 44	2
15	江西	-15.094 086 76	2
16	安徽	-14.804 812 03	2

待判结果表明:山东为高发展水平省份即第一类,重庆、陕西、四川、江西、安徽为中等发展水平省份即第二类,这与统计资料相符。

3　结　论

多元统计分析方法已经越来越多地为人们广泛应用,而对各种多元统计分析方法的适用性及应用效果的检验重视不够。此处基于 10 个省市的指标,运用多元统计中的距离判别法得到了判别 2011 年我国各省市经济社会发展水平的判别函数,并经过判别函数有效性的检验,判别了 3 个省市的经济社会发展水平,同样将其他各省的指数分别代入判别函数可判别其经济社会发展水平,该方法科学有效且简单易行。但还存在不足之处:只选取了 10 个省市的指标得到的判别函数,并且只对中国其他城市中选取了 6 个省市进行判别归类,至于其他省市判别归类是否与统计资料相符还有待研究考证。

参考文献:

[1] 国家统计局. 中国人类发展指数统计报告[DB]. 国家统计局,2010
[2] 李小亮,刘新平. 基于多元统计分析的旅游决策研究[J]. 重庆工商大学学报:自然科学版,2006,23(4):354-356
[3] 于秀林,任雪松. 多元统计分析[M]. 北京:中国统计出版社,2003
[4] 郭志刚. 社会统计分析方法 SPSS 软件应用[M]. 北京:中国人民大学出版社,1999
[5] 聂海燕. 多样本距离判别检验方法应用研究[J]. 商业时代,2011:124-125
[6] 盛骤,谢式千,潘承毅. 概率论与数理统计[M]. 北京:高等教育出版社,2005
[7] 余锦华,杨维权. 多元统计分析与应用[M]. 广州:中山大学出版社,2005

Study on China's Economic and Social Development Level Based on Distance Discriminating Method

LI Jin-yan

(School of Mathematics, Chongqing Normal University, Chongqing 401331, China)

Abstract: Distance discriminating method in multivariate statistics is used to study economic and social development level of part of provinces and municipalities of China. According to statistic data in 2011 China Human Development Index Report, ten provinces and municipality are discriminated by distance discriminating method, discriminating function is obtained, then six provinces and municipality are chosen to make discriminating analysis, and discriminating classification completely fits for statistic materials, which is helpful to evaluate the development level of other provinces.

Key words: distance discriminating method; F test; development level

责任编辑:李翠薇