

**数理统计大作业（二）**

**聚类分析和判别分析**

|  |  |
| --- | --- |
| 学 院： | 计算机学院 |
| 学 号： | ZY2006109 |
| 姓 名： | 姬轶 |
| 班 号： | 23班 |
| 序 号： | 172 |

北京航空航天大学

**2020 年12月**

摘 要

本文利用聚类分析的方法，借助 SPSS 软件，使用2019年地区生产总值与第一、第二、第三产业的增加值作为标准，将我国其中 31 个省份分为 5 类。并利用判别分析的方法，判断某些省份是否存在于相应的类别中，从而判断聚类分析和判别分析的准确性。

关键词：地区生产总值，聚类分析，判别分析，SPSS

目录

[一、 引言 1](#_Toc59823819)

[1.1. 研究背景 1](#_Toc59823820)

[1.2. 研究内容 1](#_Toc59823821)

[二、 聚类分析 1](#_Toc59823822)

[2.1. 系统聚类 1](#_Toc59823823)

[2.2. 离差平方和法（Ward法） 2](#_Toc59823824)

[2.3. 变量与数据集 2](#_Toc59823825)

[2.4. 操作步骤 3](#_Toc59823826)

[2.5. 聚类结果 4](#_Toc59823827)

[三、 判别分析 5](#_Toc59823828)

[3.1. Fisher判别法 5](#_Toc59823829)

[3.2. 样本数据 5](#_Toc59823830)

[3.3. 判别结果 5](#_Toc59823831)

[四、 结论与展望 6](#_Toc59823832)

[参考文献 8](#_Toc59823833)

# 引言

## 研究背景

地区生产总值（地区GDP）是指本地区所有[常住单位](https://baike.baidu.com/item/%E5%B8%B8%E4%BD%8F%E5%8D%95%E4%BD%8D/6782096)在一定时期内[生产活动](https://baike.baidu.com/item/%E7%94%9F%E4%BA%A7%E6%B4%BB%E5%8A%A8/9750432)的[最终](https://baike.baidu.com/item/%E6%9C%80%E7%BB%88/11000661)[成果](https://baike.baidu.com/item/%E6%88%90%E6%9E%9C/3820163)。地区生产总值等于各[产业增加值](https://baike.baidu.com/item/%E4%BA%A7%E4%B8%9A%E5%A2%9E%E5%8A%A0%E5%80%BC/8617409)之和。1993年中国将GDP正式纳为[国民经济核算](https://baike.baidu.com/item/%E5%9B%BD%E6%B0%91%E7%BB%8F%E6%B5%8E%E6%A0%B8%E7%AE%97/915453)的核心指标。2019年，中国将实施地区GDP统一核算，与国内GDP数据基本衔接。因此，对地区生产总值的分析具有重大的经济与社会意义[1]。

## 研究内容

地区生产总值大致可以由第一、第二、第三产业的增加值构成，其中，第一产业是指农、林、牧、渔业（不含农、林、牧、渔专业及辅助性活动）。第二产业是指采矿业（不含开采专业及辅助性活动），制造业（不含金属制品、机械和设备修理业），电力、热力、燃气及水生产和供应业，建筑业。第三产业即服务业，是指除第一产业、第二产业以外的其他行业。

本文的研究内容是对各省份从几个产业增长值方面进行分类。进而总结各省份的财政状况。分类的一个重要原则就是保证类与类之间的成员的差距尽量的大，类内部成员的差距尽量的小，因此本文使用离差平方和法（Ward 法）进行系统聚类。在判别分析上，研究时使用了 Fisher 判别方法，因为数据总体的分布类型未知，而 Fisher 判别正好对对总体的分布类型没有要求，能够在该题目下的判别分析上有较好的表现[2]。

# 聚类分析

## 系统聚类

聚类分析是研究如何将对象按照多个方面的特征进行综合分类的一种统计 方法，按照分类对象的不同，可以分为 Q 型聚类和 R 型聚类。Q 型聚类是对样本进行分类，而 R 型聚类是对变量进行分类。本文分类的对象是我国各省份，因此属于 Q 型聚类。Q 型聚类最常用的聚类方法是系统聚类法。分类有许多种方法，最常用的一种方法是在样品距离的基础上定义类与类之间的距离。首先将n个样品分成n类，每个样品自成一类，然后每次将具有最小距离的两类合并，合并后重新计算类与类之间的距离，这个过程一直持续到将所有的样品归为一类为止，并把这个过程画成一张聚类图，参照聚类图可方便地进行分类。因为聚类图很像一张系统图，所以这种方法就叫系统聚类法，又称为分层聚类法。

## 离差平方和法（Ward法）

Ward法是由瓦尔德提出的离差平方和法[3]，该方法的基本思想是同类样本间的离差平方和较小，异类样本间的离差平方和大。其通过离差平方和来定义类与类之间的相似性测度。

基于方差分析的思想，如果分类正确，同类样品之间的离差平方和应当较小，类与类之间的离差平方和应当较大。具体步骤如下所示：

1）将每个样品各自成一组。

2）每次通过合并减少一组。

3）此时离差平方和出现并逐步增大，选择使离差平方和增加最小的两组合并，知道所有的样品归为一类为止。

将个区域样本分成 类，，用表示中的第个样本，表示中的样本个数，是的重心（即该类样本的均值），则中的样本离差平方和为：

整体类内的离差平方和为：

## 变量与数据集

一个地区的发展，可以由该地的地区生产总值、第一、第二、第三产业的增加值来看出，因此，本文选择了这四个数据作为指标，，可以为变量定义符号，如下表所示：

表2-1 各变量说明

|  |  |
| --- | --- |
| 变量名 | 符号 |
| 地区生产总值 |  |
| 第一产业的增加值 |  |
| 第二产业的增加值 |  |
| 第三产业的增加值 |  |

因为需要后续进行判别分析，选取了 28 个省份的数据（数据来自2020年统计年鉴），数据如下所示：

表2-2 我国各省份地区经济情况数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 地区 | 地区生产总值 | 第一产业增加值 | 第二产业增加值 | 第三产业增加值 |
| 天 津 | 14104.28 | 185.23 | 4969.18 | 8949.87 |
| 河 北 | 35104.52 | 3518.44 | 13597.26 | 17988.82 |
| 山 西 | 17026.68 | 824.72 | 7453.09 | 8748.87 |
| 内蒙古 | 17212.53 | 1863.19 | 6818.88 | 8530.46 |
| 辽 宁 | 24909.45 | 2177.77 | 9531.24 | 13200.44 |
| 黑龙江 | 13612.68 | 3182.45 | 3615.21 | 6815.03 |
| 上 海 | 38155.32 | 103.88 | 10299.16 | 27752.28 |
| 江 苏 | 99631.52 | 4296.28 | 44270.51 | 51064.73 |
| 浙 江 | 62351.74 | 2097.38 | 26566.60 | 33687.76 |
| 安 徽 | 37113.98 | 2915.70 | 15337.90 | 18860.38 |
| 福 建 | 42395.00 | 2596.23 | 20581.74 | 19217.03 |
| 江 西 | 24757.50 | 2057.56 | 10939.83 | 11760.11 |
| 山 东 | 71067.53 | 5116.44 | 28310.92 | 37640.17 |
| 河 南 | 54259.20 | 4635.40 | 23605.79 | 26018.01 |
| 湖 北 | 45828.31 | 3809.09 | 19098.62 | 22920.60 |
| 湖 南 | 39752.12 | 3646.95 | 14946.98 | 21158.19 |
| 广 东 | 107671.07 | 4351.26 | 43546.43 | 59773.38 |
| 广 西 | 21237.14 | 3387.74 | 7077.43 | 10771.97 |
| 海 南 | 5308.93 | 1080.36 | 1099.03 | 3129.54 |
| 重 庆 | 23605.77 | 1551.42 | 9496.84 | 12557.51 |
| 四 川 | 46615.82 | 4807.24 | 17365.33 | 24443.25 |
| 贵 州 | 16769.34 | 2280.56 | 6058.45 | 8430.33 |
| 云 南 | 23223.75 | 3037.62 | 7961.58 | 12224.55 |
| 西 藏 | 1697.82 | 138.19 | 635.62 | 924.01 |
| 甘 肃 | 8718.30 | 1050.48 | 2862.42 | 4805.40 |
| 青 海 | 2965.95 | 301.90 | 1159.75 | 1504.30 |
| 宁 夏 | 3748.48 | 279.93 | 1584.72 | 1883.83 |
| 新 疆 | 13597.11 | 1781.75 | 4795.50 | 7019.86 |

## 操作步骤

本次聚类分析使用 SPSS 软件的系统聚类功能，方法选择 Ward 法，测量距离使用平方欧氏距离（本次使用的变量无单位问题，故可以使用平方欧氏距离），变量使用 2.4 节所声明的变量。

## 聚类结果

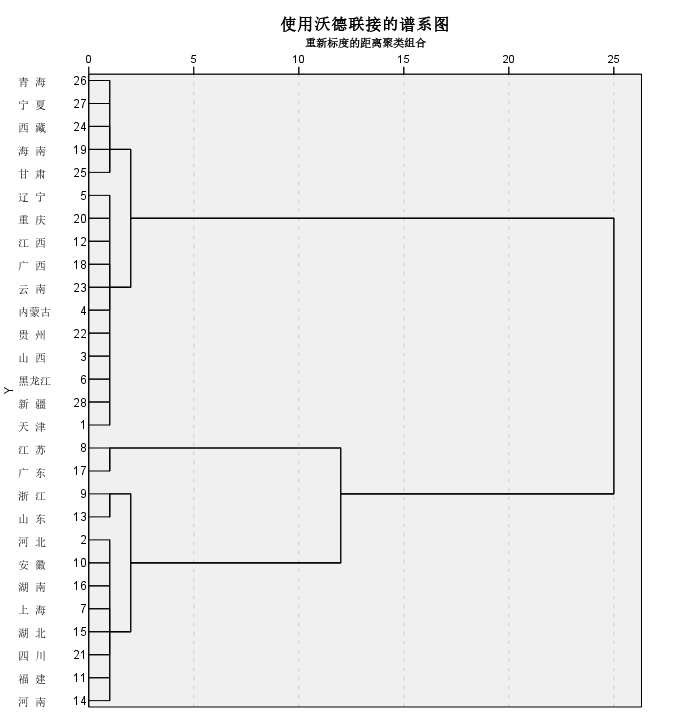


图2-1 聚类分析谱系图

根据图2-1可以很明显的看出，28个省份可以大致被分为五类，青海、宁夏、西藏、海南、甘肃为一类，该类中的特点为该省份的地区生产总值主要分布在第三产业，且生产总值较小，辽宁、重庆、江西、广西、云南、内蒙古、贵州、山西、黑龙江、新疆、天津为一类，该类中的特点为该省份的地区生产总值主要分布在第三产业，且生产总值相比于第一类较大，江苏、广东为一类，该类中的特点是第一、第二、第三产业的增加值比较均衡，浙江、山东为一类，该类中的特点为该省份的地区生产总值主要分布在第二、第三产业，且生产总值较大，河北、安徽、湖南、上海、湖北、四川、福建、河南为一类，该类中的特点为该省份的地区生产总值仍是以分布在第二、第三产业为主，但生产总值没有第四类那么大。可以看出每个类内部的省份都有一定的共同特点，而类间的省份区别较为明显，故认为聚类分析准确。

# 判别分析

## Fisher判别法

判别分析是一种根据样本的观察值，判定该样本归属何种总体的一种统计 方法。判别分析是机器学习、大数据、模式识别等领域的重要理论基础，其主 要有距离判别、Bayes 判别、Fisher 判别等多种判别方法。本文使用的是 Fisher 判别方法。

Fisher 判别是由 R·A·Fisher 提出的一种降维处理方法，该方法可以对一般分布的总体导出线性判别函数。 Fisher 判别法因为其对总体分布类型无特殊要求，仅仅要求总体的协方差矩阵存在，故选取作为在该实验中分布情况未知的判别方法。

## 样本数据

判别分析中选取从总体中提前摘出的三个省份（背景、吉林、陕西）作为样本数据进行Fisher 判别分析，数据如表 3-1 所示。

表3-1 样本数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 地区 | 地区生产总值 | 第一产业增加值 | 第二产业增加值 | 第三产业增加值 |
| 北 京 | 35371.28 | 113.69 | 5715.06 | 29542.53 |
| 吉 林 | 11726.82 | 1287.32 | 4134.82 | 6304.68 |
| 陕 西 | 25793.17 | 1990.93 | 11980.75 | 11821.49 |

## 判别结果

在SPSS中设定Fisher 判别分析如图 3-1 所示：

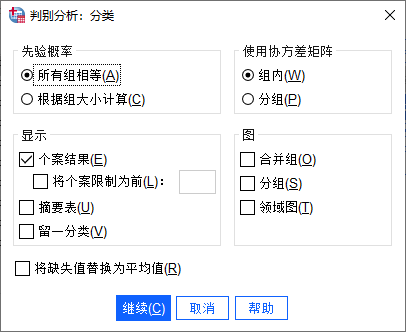
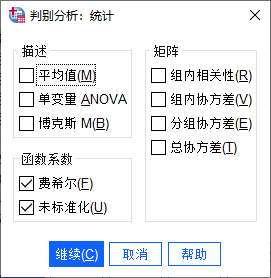


图3-1 判别分析设定



图 3-2 判别函数特征值



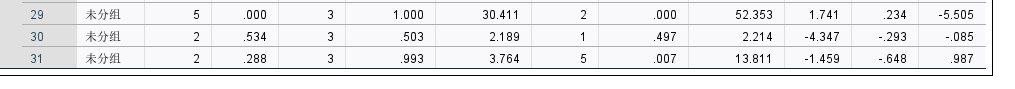


图 3-3 个案统计情况和判别结果

由个案统计情况和判别结果所示，北京被分为第5类，陕西和吉林被分为第二类，北京和上海经济模式较为相同，故在同一类中可以认为是准确的。陕西、吉林位于我国的西北、东北部，东北部是典型的黑吉辽经济模式，故也可认为分类准确。可以判断聚类分析效果很好。

# 结论与展望

通过对全国28个省份城市的地区生产总值进行聚类分析，可以看出某些相邻近的区域有着相同的经济运转模式，二不同大区之间，经济模式有有较大的差别。本文的研究内容是对各省份从几个产业增长值方面进行分类。进而总结各省份的财政状况。在判别分析上，研究时使用了 Fisher 判别方法，因为数据总体的分布类型未知，而 Fisher 判别正好对对总体的分布类型没有要求，能够在该题目下的判别分析上有较好的表现，适用性较广。

可将该判别方法推广到全球各国家，以更加准确地探讨各种工业模式下的经济发展状况。

# 参考文献

[1] 《中国统计年鉴》，2020

[2] 孙海燕，周梦，李卫国，冯伟. 数理统计[M]. 北京：北京航空航天大学应用数学与系统 科学学院，2015: P160-P171

[3] 王斌会编著,多元统计分析及R语言建模 第4版,暨南大学出版社,2016.03,第138页