# 北京航空航天大学研究生课程考核记录

## 2018 - 2019 学年 第一学期

学号 ZF1821308　　姓名 夏涛　　成绩

课程名称：《数理统计》

论文题目：河北省各城市综合经济实力分析

任课教师评语：

任课教师签字：　　考核日期：　　年　　月　　日

# 摘　　要

城市综合经济实力是城市所拥有的全部实力、潜力及其在国内外经济社会中的地位和影响力，是城市综合竞争力的重要基础，评价城市综合经济实力应偏重与经济总量有关的数据。本文对河北省的 11 个地级市 2017 年 12 项经济指标进行因子分析,数据来源于 2017 年河北省经济年鉴各城市的主要经济指标，反映河北省各地区综合经济实力现状,为今后各市的发展提供了理论依据。运用统计学、多元统计分析等相关知识，结合中国综合经济实力的环境和河北各城市综合经济实力的实际情况和特点，采用因子分析对河北省各市的部分指标进行研究分析，找出反应城市经济实力的主要指标，为河北省各城市的经济实力进行量化打分。并且，进一步对河北省各城市进行聚类分析，将河北省城市进行分类，针对于每一类城市分析其发展的特点。最终，我们对河北省各市在经济发展方面所面临问题提出相应对策及建议

**关键词：**经济实力，因子分析，K-means，spss

# 目　录

# 第一章 数据采集和数据预处理

## 1.1 数据的采集和整理

本文的数据主要来自河北省统计局官方网站发布的河北省《2017 年经济年鉴》，我们主要收集了河北省 11 个地级市(包括石家庄、承德市、张家口市、秦皇岛市、唐山市、廊坊市、保定市、沧州市、衡水市、邢台市、邯郸市)的主要经济指标。分别是：x 1 是各市人口总数，各个地区的人口数量的多少常常和经济生活连成一体，一个城市经济发展越好年底的常住人口数量越大； x 2 是生产总值 GDP(亿元)，反映出一个地区的财富与国力，也表现出一个地区的经济能力； x 3 是全社会固定资产投资总额(亿元)，反映的是固定资产投资规模速度比例关系和使用方向； x 4 是地方一般预算收入(亿元)，是财政收入的来源之一，反映了城市的经济发展前景； x 5 是财政支出(亿元)，是财政部门按照预算向有关部门进行支付的活动；x 6 是城镇非私营单位就业人员工资总额(亿元)，反映了单位在报告期直接支付给单位人员的劳动报酬，侧面反映了城市的综合经济情况； x 7 是城镇居民人均可支配收入(元)，表示城镇家庭日常生活的那部分的收入；x 8 是农村居民人均纯收入(元)，表示农村居民家庭全年收入； x 9 是农林牧渔业总产值(亿元)，表示农林牧渔业全部产品总量； x 10 是规模以上工业总产值(亿元)，表示工业企业在报告期内生产的工业产品总量，反映了企业的效率，关系着经济的发展； x 11 是进出口总额(亿元)，它可以观察国家或地区对外贸易方面的总规模； x 12 是社会消费品零售总额(亿元)，反映各行业通过商品流通向居民和社会供应的生活消费品总量。将收集的数据进行整理如图 1 所示篇幅有限只展示了部分数据。

| 城市 | 人口总数（万人） | 生产总值（亿元） | 全社会固定资产投资总额（亿元） | 一般预算收入(亿元) | 财政支出（亿元） | 城镇非私 |
|---|---|---|---|---|---|---|
| 石家庄 | 1078.46 | 5927.7293 | 3066.2 | 314.97 | 443.6 | |
| 承德市 | 353.18 | 1438.5741 | 276.8 | 38.21 | 104.5 | |
| 张家口市 | 442.51 | 1465.9911 | 692.6 | 80.76 | 179.57 | |
| 秦皇岛市 | 309.46 | 1349.3526 | 588.08 | 99.11 | 174.85 | |
| 唐山市 | 784.36 | 6354.8675 | 2679.41 | 250.58 | 404.47 | |
| 廊坊市 | 461.5 | 2720.458 | 438.64 | 50.6 | 61.7 | |
| 保定市 | 1163.45 | 3477.1269 | 883.26 | 121.13 | 204.88 | |
| 沧州市 | 750.55 | 3544.68 | 1106.79 | 99.06 | 156.02 | |
| 衡水市 | 445.31 | 1420.1825 | 402.86 | 25.8 | 120.4 | |
| 邢台市 | 731.99 | 1975.746 | 343.41 | 11.66 | 58.63 | |
| 邯郸市 | 949.28 | 3361.1123 | 1602.22 | 111.49 | 239.34 | |

**图 1 河北省各市主要经济指标**

## 1.2 数据的初步处理和分析

通过观察上述数据我们可以发现在上述 12 项经济指标中，财政收入这一项发现某些市的财政收入远远低于财政支出(地方财政支出会有一部分来自中央的补贴)，因此若直接使用此项指标对于衡量地方经济发展水平会对分析结果带来干扰，因此我们先将财政支出这一项经济指标剔除用剩余的 11 项经济指标对各市的经济状况进行分析。从原始数据上我们可以初步得出结论石家庄、唐山的经济发展水平明显高于河北省其他城市，而承德和张家口的经济发展水平较其他地区稍显不足。为了进一步更精确的分析各个城市的经济发展水平我们对这 11 项经济指标进行因子分析找出它们背后的公共因子从而对各市的经济状况进行评价。

# 第二章 因子分析

## 2.1 因子分析简介

因子分析是主成分分析的推广，PCA 是因子分析省略了特殊因子的一个特例，因此因子分析比主成分分析更为精确。因子分析模型主要是利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，将一些错综复杂关系的变量通过归纳和总结的方法将其归结为少数因子的一种多变量统计分析方法。从以上可以看出，因子分析的出发点是原始变量的相关矩阵。

依据相关性大小把原始变量分组因子是本次分析的基本思想，本次研究使不同组之间的变量的相关性较低，而同组内的变量之间则相关性较高。通常来说，每组变量就是代表着一个基本结构，同时，采用一个不可观测的综合变量对其表示，这个基本结构就成为公共因子。对于所研究的某一具体问题，原始变量可以分解成两部分之和的形式，一部分是与公共因子无关的特殊因子，另一部分是少数几个不可测得的公共因子的线性函数。

一般因子分析模型如图 2：

$$\begin{cases} X_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1 \\ X_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\varepsilon_2 \\ \cdots\cdots \\ X_p=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\varepsilon_p \end{cases}$$

**图 2 一般因子分析模型**

其矩阵形式为： X =AF+ ε 其中

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

为因子载荷矩阵， X 是一个 p 维向量， F 是一个 m 维向量。在进行因子分析之前我们一般首先需要对数据进行标准化。使得 E(X)=0，D(X)=1

## 2.2 Spss 进行因子分析

### 2.2.1 相关性分析

如上文所述，因子分析的出发点是变量的相关系数矩阵，因此进行因子分析的第一步就是进行相关性检验。我们将数据导入 spss 可以得到变量的相关系数矩阵。如表 1 所示。

**相關性矩陣 [a,b]**

| | | 人口总数（万人） | 生产总值（亿元） | 全社会固定资产投资总额(亿元） | 一般预算收入(亿元) |
|---|---|---|---|---|---|
| 相關 | 人口总数（万人） | 1.000 | .708 | .629 | .584 |
| | 生产总值（亿元） | .708 | 1.000 | **.932** | .895 |
| | 全社会固定资产投资总额（亿元） | .629 | **.932** | 1.000 | **.964** |
| | 一般预算收入(亿元) | .584 | .895 | **.964** | 1.000 |
| | 城镇非私营单位就业人员工资总额（万元） | .852 | .897 | .809 | .832 |
| | 城镇居民人均可支配收入（元） | .006 | .565 | .479 | .531 |
| | 农村居民可支配收入（元） | .337 | .720 | .581 | .563 |
| | 农林牧副渔总产值（万元） | .853 | **.908** | .876 | .821 |
| | 规模以上工业总产值（亿元） | .640 | .872 | **.950** | **.963** |
| | 社会消费品零售总额（万元） | .703 | **.910** | .860 | .873 |
| | 进出口总额（千美元） | .461 | .877 | .864 | .898 |

**表 1 相关性矩阵**

由于篇幅有限我们只截取部分表从上表中我们可以看出多个变量之间的相关性高于 0.5 红色部分相关变量之间的相关性超过 0.9，说明 11 个经济指标之间具有很强的相关性。因此可以进行因子分析找出公共因子。

### 2.2.2 共同度分析

**Communalities**

| | 起始 | 擷取 |
|---|---|---|
| 人口总数（万人） | 1.000 | .838 |
| 生产总值（亿元） | 1.000 | .952 |
| 全社会固定资产投资总额（亿元） | 1.000 | .901 |
| 一般预算收入(亿元) | 1.000 | .892 |
| 城镇非私营单位就业人员工资总额（万元） | 1.000 | .892 |
| 城镇居民人均可支配收入（元） | 1.000 | .964 |
| 农村居民可支配收入（元） | 1.000 | .837 |
| 农林牧副渔总产值（万元） | 1.000 | .929 |
| 规模以上工业总产值（亿元） | 1.000 | .894 |
| 社会消费品零售总额（万元） | 1.000 | .875 |
| 进出口总额（千美元） | 1.000 | .916 |

**表 2 共同度矩阵**

共同度表示的是公共因子 F 的线性组合对于每一个变量的贡献度，从表中可以看出公共因子对于每个变量的贡献度均大于 0.5 表明公共因子可以较好的反映原始的各项经济指标。

### 2.2.3 特征值分析

**說明的變異數總計**

| | 起始特徵值 | | | 擷取平方和載入 | | |
|---|---|---|---|---|---|---|
| 元件 | 總計 | 變異的 % | 累加 % | 總計 | 變異的 % | 累加 % |
| 1 | 8.387 | 76.246 | 76.246 | 8.387 | 76.246 | 76.246 |
| 2 | 1.502 | 13.654 | 89.900 | 1.502 | 13.654 | 89.900 |
| 3 | .589 | 5.352 | 95.252 | | | |
| 4 | .182 | 1.659 | 96.911 | | | |
| 5 | .162 | 1.469 | 98.380 | | | |
| 6 | .094 | .855 | 99.235 | | | |
| 7 | .043 | .388 | 99.623 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | .031 | .284 | 99.907 | | | |
| 9 | .010 | .092 | 99.999 | | | |
| 10 | 9.719E-5 | .001 | 100.000 | | | |
| 11 | 1.205E-16 | 1.095E-15 | 100.000 | | | |

**表 3 特征值分析**

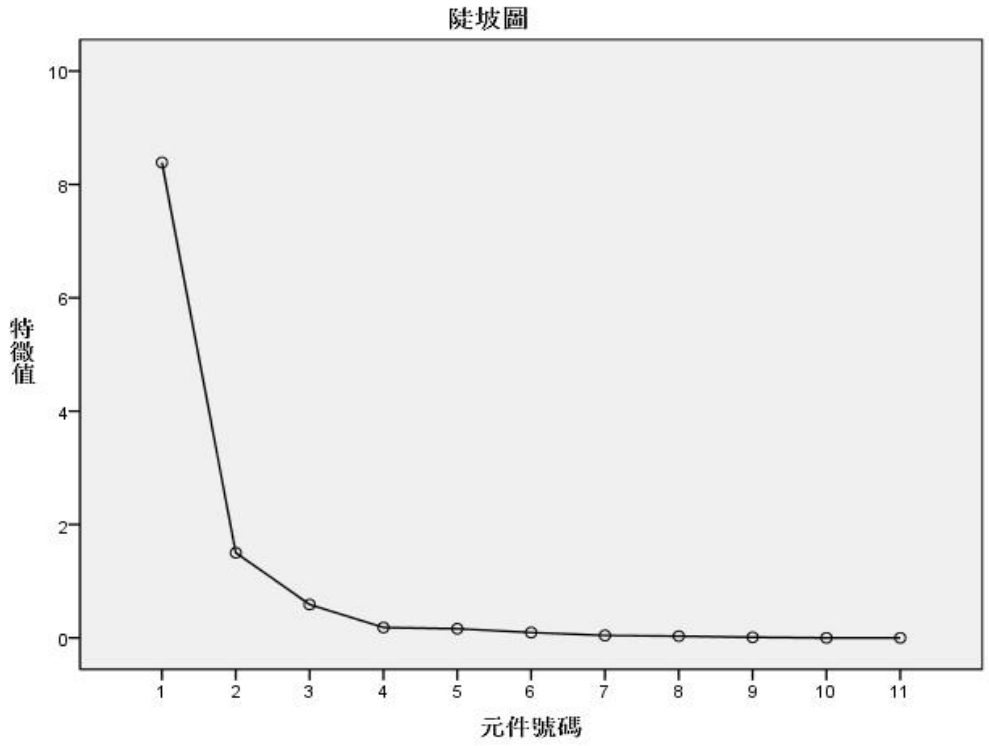


**图 3 碎石图**

从表 3 和可知图 3 碎石图可知相关系数矩阵的特征值从大到小进行排列，且下降的速度很快，我们选取大于 1 的前两个特征值 $\lambda_1 = 8.387$， $\lambda_2 = 1.502$ 从表中我们可以看出前两个特征对于总方差的贡献率为 0.76 和 0.13 个总的贡献率为 0.89。我们分别将其命名为 $F_1$ 和 $F_2$，它们就是我们找到的公共因子。

### 2.2.4 载荷矩阵

**元件矩陣 [a]**

| | 元件 | |
|---|---|---|
| | 1 | 2 |
| 人口总数（万人） | .721 | -.564 |
| 生产总值（亿元） | .976 | -.003 |
| 全社会固定资产投资总额（亿元） | .947 | -.058 |
| 一般预算收入(亿元) | .945 | -.004 |

| | | |
|---|---|---|
| 城镇非私营单位就业人员工资总额（万元） | .934 | -.137 |
| 城镇居民人均可支配收入（元） | .573 | .797 |
| 农村居民可支配收入（元） | .725 | .558 |
| 农林牧副渔总产值（万元） | .898 | -.349 |
| 规模以上工业总产值（亿元） | .943 | -.064 |
| 社会消费品零售总额（万元） | .925 | -.142 |
| 进出口总额（千美元） | .920 | .262 |

**表 4 载荷矩阵 A**

通过观察表 4 我们可以看出 $F_1$ 与各项经济指标的贡献度都比较大，而 $F_2$ 对城镇居民人均可支配收入和农村居民可支配收入这两项经济指标的贡献度比较大。因此我们将 $F_1$ 命名为经济发展水平，$F_2$ 命名为居民收入水平。

### 2.2.5 因子得分

利用回归法计算各个城市的因子得分，

| 城市 | 经济发展水 | 居民收入水 |
|---|---|---|
| 石家庄 | 1.871323 | -0.47741 |
| 承德市 | -1.0921 | -0.27407 |
| 张家口市 | -0.79865 | -0.25122 |
| 秦皇岛市 | -0.55519 | 1.249469 |
| 唐山市 | 1.671359 | 0.873512 |
| 廊坊市 | -0.19316 | 2.097314 |
| 保定市 | 0.471253 | -1.337 |
| 沧州市 | -0.00749 | -0.14902 |
| 衡水市 | -0.79607 | -0.31686 |
| 邢台市 | -0.7904 | -0.71224 |
| 邯郸市 | 0.219135 | -0.70248 |

**图 4 各城市因子得分**

## 2.3 城市排行

我们根据因子得分结果对各城市的经济状况进行评分，计算公式如下

$$score = w_1 \times F_1 + w_2 \times F_2$$

其中 $w_1$ 和 $w_2$ 为 $F_1$ 和 $F_2$ 因子的权重，计算公式为：

$$w_i = \frac{F_i\text{因子对于总方差的贡献度}}{\text{总方差贡献度}}$$

经过计算得到河北省各市的排行情况为：

| 城市 | 得分 | 排名 |
|---|---|---|
| 唐山市 | 1.55 | 1 |
| 石家庄 | 1.51 | 2 |
| 保定市 | 0.2 | 3 |
| 廊坊市 | 0.15 | 4 |
| 邯郸市 | 0.08 | 5 |
| 沧州市 | -0.03 | 6 |
| 秦皇岛市 | -0.28 | 7 |
| 张家口市 | -0.72 | 8 |
| 衡水市 | -0.73 | 9 |
| 邢台市 | -0.78 | 10 |
| 承德市 | -0.97 | 11 |

**表 5 河北省所有城市排行**

# 第三章 聚类分析

## 3.1 K-means 聚类简介

K-means 是现在最为常用的聚类算法之一，主要思想是先随机选取 K 个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

## 3.2 对河北省所有城市进行聚类

我们采用 K-means 聚类算法按照各城市的因子得分对河北省的所有地级市进行聚类，分析不同类型城市之间的差异，为各个城市的发展提出建议。

利用 Spss 进行聚类分析的结果如下所示：

**起始叢集中心**

| | 叢集 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 经济发展水平 | .471253223 | -1.092104603 | 1.671359172 | -.193157417 |
| 居民收入水平 | -1.336996948 | -.274067657 | .873512481 | 2.097313783 |

**表 6 初始聚类中心**

**最終叢集中心**

| | 叢集 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 经济发展水平 | .345194089 | -.696943739 | 1.771341323 | -.374176066 |
| 居民收入水平 | -1.019740921 | -.340680272 | .198050199 | 1.673391402 |

**表 7 最终收敛中心**

从最终的收敛中心我们可以初步看出，第一类城市经济发展水平和居民收入水平位于全省平均水平附近，经济发展发展水平和居民收入水平比较均衡。第二类城市经济发展水平和居民收入水平均低于全省平均水平，第三类城市经济发展水平和居民收入水平均位于全省的前列，但是虽然经济发展水平高但是居民收入水平相对于经济总体而言略显不足，而第四类城市的经济体量虽然不大但是居民收入水平却相对较高。为了更直观的表示聚类结果我们利用 matplotlib 绘制散点图，更直观的展现各城市的分类状况。

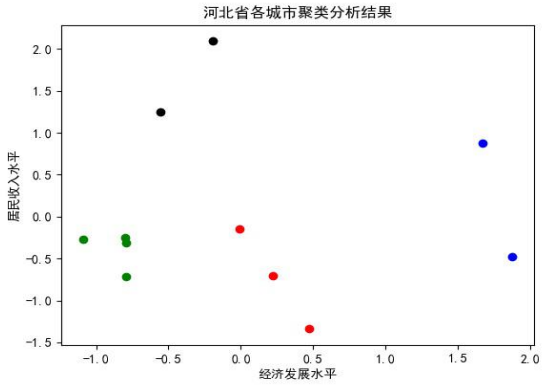


**图 5 聚类分析结果**

| 类别 | 城市 |
| --- | --- |
| 第一类 | 保定 沧州 邯郸 |
| 第二类 | 承德 张家口 衡水 邢台 |
| 第三类 | 石家庄 唐山 |
| 第四类 | 秦皇岛 廊坊 |

**表 8 城市分类表**

图 5 中红色的点表示第一类城市包括保定、沧州、邯郸，第一类城市对于经济发展和人民收入水平的提高之间的关系处理的比较好，因此主要任务是继续大力发展经济，做大蛋糕；绿色的点表示第二类城市包括承德、张家口、衡水、邢台，第二类城市经济发展水平和居民收入水平均低于全省的平均水平，这些城市应当结合自身的特点，找到适合自身的发展道路，积极发展经济改善人民生活，蓝色的点表示第三类城市包括石家庄、唐山，蓝色的点位于图的右方表示这一类城市的经济发展水平位于全省的前列，但相应的居民的收入水平与经济体量相比却稍显不足，因此第三类城市应当完善收入分配制度，将经济发展成果更好的普及更多的人民；黑色的点表示第四类城市包括秦皇岛、廊坊，第四类城市与第三类城市正好相反，此类城市的经济体量不大但是居民的收入水平却位于全省的前列，甚至高于石家庄和唐山，针对于这些城市我们应当进一步分析造成这种现象的原因，制定发展规划。

# 第四章 结论与展望

## 4.1 结论

本文针对河北省各地级市的 11 项经济指标进行因子分析，将 11 项指标归纳为经济发展水平和居民收入水平，对河北省所有城市进行评分和排名，接下来利用 K-means 对城市进行聚类分析，将所有城市分为 4 类并对于每一类城市的特点进行了分析，提出了发展建议

## 4.2 展望

本文只是对所有城市经济发展状况的初步分析，采集的数据只是最基础的衡量经济发展水平的指标，若想进一步分析各个城市的发展特点，应当采集更多的经济指标来进行分析，因子分析中我们只采用了两个公共因子，可以尝试增加公共因子的数量同时对载荷矩阵进行因子旋转，使得各因子的含义更加名确，同时还可以进一步对城市的各产业状况进行制定更加切实可行的发展规划。