聚类分析实验

信171 李金哲

PART. 1 ——

层次聚类

操作过程

选择菜单->【分析】->【分类】->【系统聚类】(图1.1)

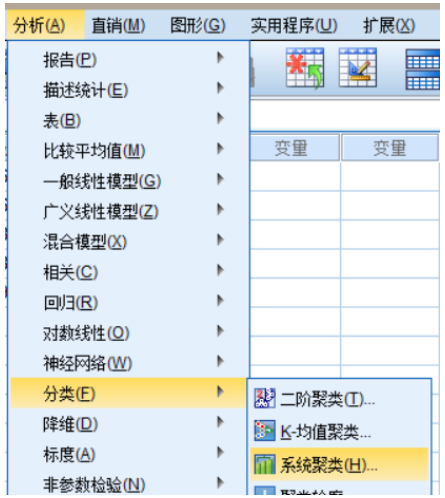

图1.1 选择菜单

将购货环境和服务质量放入变量；商店变化放入个案标注中（图1.2）

图1.2 层次分析窗口

点击统计，勾选解的范围，即分的类数，分别填2和3（图1.3）

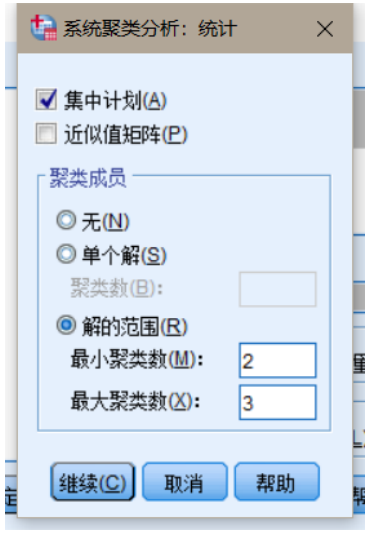

图1.3 统计窗口

点击方法，选择基础的欧式距离即可（图1.4）

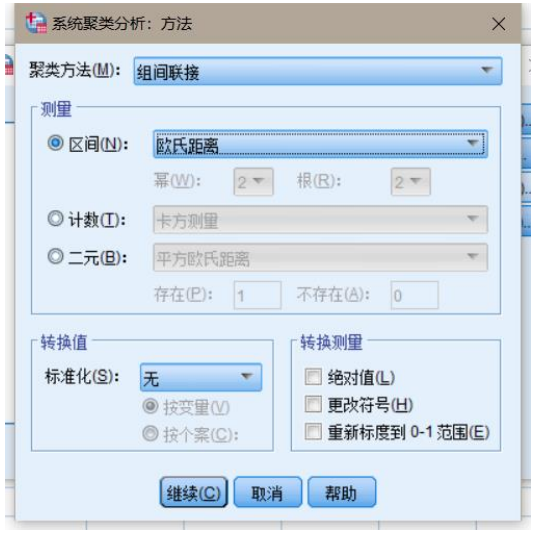

图1.4 方法窗口

结果分析

表1.1 凝聚状态表

集中计划

阶段	组合聚类 聚类 1	组合聚类 聚类 2	系数	首次出现聚类的阶段 聚类 1	首次出现聚类的阶段 聚类 2	下一个阶段
1	4	5	3.606	0	0	3
2	1	2	8.062	0	0	4
3	3	4	11.013	0	1	4
4	1	3	28.908	2	3	0

表1.2 类成员表

聚类成员

个案	3 个聚类	2 个聚类
1:A商厦	1	1
2:B商厦	1	1
3:C商厦	2	2
4:D商厦	3	2
5:E商厦	3	2

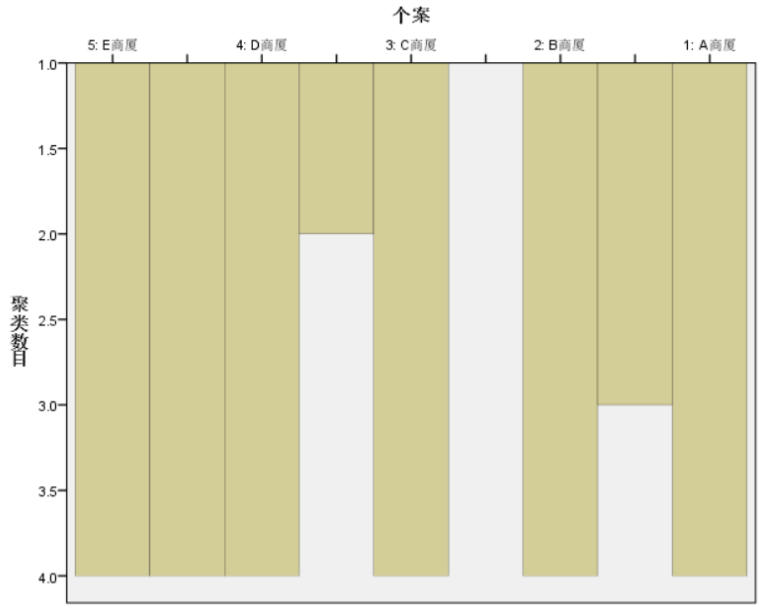

图1.5 冰柱图

在表1.1中，第一列表示聚类分析的第几步;第二列、第三列表示本步聚类中哪两个观测个体或小类聚成一类;第四列是个体距离或小类距离;第五列、第六列表示本步聚类中参与聚类的是个体还是小类，0表示个体(样本)，非0表示由第几步聚类生成的小类参与本步聚类;第七列表示本步聚类的结果将在以下第几步中用到。

表1.1 显示了五座商厦聚类的情况。聚类分析的第1步中，4号观测(D商厦)与5号观测(E商厦)聚成一小类，它们的个体距离(这里采用欧氏距离)是3.606，这个小类将在下面第3步用到;同理，聚类分析的第3步中，3号观测(C商厦)与第1步聚成的小类(以该小类中第1个观测号4为标记)又聚成一小类，它们的距离(个体与小类的距离，这里采用组间平均链锁距离)是11.013，形成的小类将在下面第4步用到。经过4步聚类过程，5个样本最后聚成了一大类。**n个观测需n—1步聚成一个大类，第k步完成时可形成“n—k”个类**。

由表1.2可知，当聚成3类时，A，B两个商厦为一类，C商度自成一类，D，E两个商厦为一类;当聚成2类时，A，B两个商厦为一类，C，D，E三个商厦为一类。可见，SPSS的层次聚类能够产生任意类数的分类结果。

图1.5是五座商厦聚类的纵向显示的冰柱图。观察冰柱图应从最后一行开始，横向切一条线，在同一区域的为一类。如图1.5，当聚成4类时，D，E商厦为一类，其他各商厦自成一类;当聚成3类时，A，B商厦为一类，D，E商厦为一类，C商厦自成类;当聚成2类时，A，B商厦为一类，C，D，E商厦为一类。

PART. 2

K-M聚类

操作过程

选择菜单->【分析】->【分类】->【K均值聚类】(图2.1)

图2.1 选择菜单

将省事放入个案标注依据中，其他的放入变量中（图2.2）

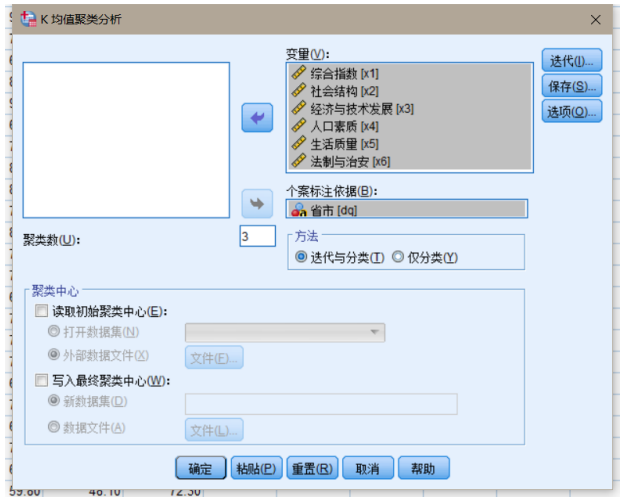

图2.2 K-M聚类分析窗口

点击迭代，设置迭代次数和新旧聚类点的间距差（图2.3）

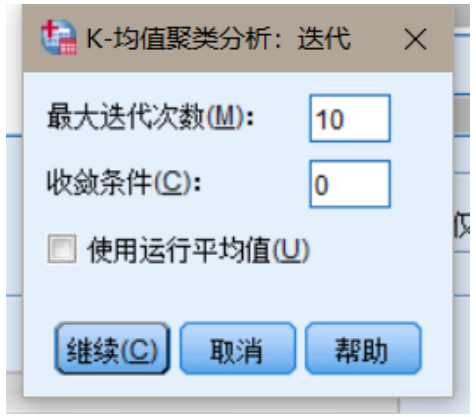

图2.3 迭代窗口

点击选项，将ANONA表勾选上（图2.4）

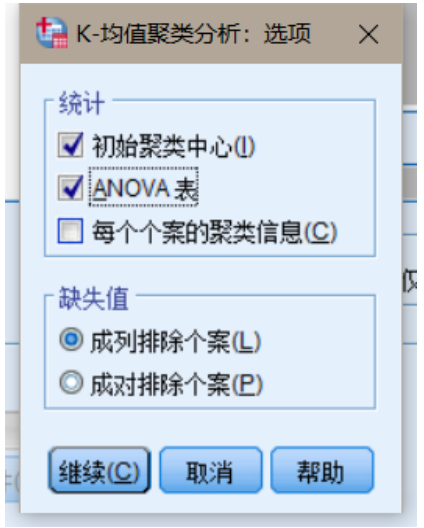

图2.4 选项窗口

结果分析1

表2.1 初始聚类中心

初始聚类中心

	聚类		
	1	2	3
综合指数	79.20	92.30	51.10
社会结构	90.40	95.10	61.90
经济与技术发展	86.90	92.70	31.50
人口素质	65.90	112.00	56.00
生活质量	86.50	95.40	41.00
法制与治安	59.40	57.50	75.60

表2.2 迭代记录

迭代历史记录[a]

	聚类中心中的变动		
迭代	1	2	3
1	24.387	6.307	23.579
2	.000	.000	.000

a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 2。初始中心之间的最小距离为 49.349。

表2.3 最终聚类中心

最终聚类中心

	聚类		
	1	2	3
综合指数	75.49	91.13	60.02
社会结构	82.86	96.17	66.86
经济与技术发展	72.41	92.03	44.03
人口素质	77.74	106.13	69.32
生活质量	75.84	94.27	51.81
法制与治安	67.17	58.57	76.15

表2.4 聚类个案数

每个聚类中的个案数目

聚类	1	7.000
	2	3.000
	3	21.000
有效		31.000
缺失		.000

表2.5 聚类结果

聚类成员

个案号	省市	聚类	距离
1	北京	2	7.102
2	上海	2	6.307
3	天津	2	11.431
4	浙江	1	20.820
5	广东	1	24.387
6	江苏	1	8.875
7	辽宁	1	18.871
8	福建	1	12.532
9	山东	1	18.045
10	黑龙江	1	20.799
11	吉林	3	26.199
12	湖北	3	15.427
13	陕西	3	18.023
14	河北	3	19.819
15	山西	3	15.668
16	海南	3	26.012
17	重庆	3	14.438
18	内蒙古	3	12.907
19	湖南	3	12.853
20	青海	3	10.796
21	四川	3	7.236
22	宁夏	3	17.136
23	新疆	3	20.272
24	安徽	3	9.844
25	云南	3	14.805
26	甘肃	3	11.144
27	广西	3	14.372
28	江西	3	15.614
29	河南	3	19.973
30	贵州	3	23.579
31	西藏	3	31.619

表2.1展示了3个类的初始类中心点的情况。3个初始类中心点的数据分别是(79.20，90.40，86.90，65.90，86.50，59.40)，(92.30，95.10，92.70，112.00，95.40、57.50)，(51.10，61.90，31.50，56.00，41.00，75.60)。可得**第2类各指数均是最优的，第1类次之，第3类各指数最不理想**。

表2.2展示了3个类中心点每次送代时的偏移情况。由表2.2可知，第1次送代后，3个类的中心点分别偏移了24.387，6.307，23.579，**第1类中心点偏移最大**。第2次选代后，3个类的中心点的偏移均小于指定的判定标准(0.02)，聚类分析结束。

表2.3展示了3个类的最终类中心点的情况。3个最终类中心点的数据分别是(75.49,82.86,72.41，77.74,75.84,67.17),(91.13,96.17,92.03,106.13,94.27,58.57)，(60.02,6.86,44.03,69.32,51.81,76.15)。仍然可见，**第2类各指数均是最优的，第1类次之，第3类各指数最**。

表2.4展示了3个类的类成员情况。第1类(中游水平)有7个省市自治区，第2类(上游水平)有3个省市自治区，第3类(下游水平)有21个省市自治区。这里，**聚类结果见表2.5**。

结果分析2

表2.5 聚类结果分析

ANOVA

	聚类		误差			
	均方	自由度	均方	自由度	F	显著性
综合指数	1633.823	2	22.518	28	72.556	.000
社会结构	1539.872	2	47.312	28	32.547	.000
经济与技术发展	4381.296	2	56.760	28	77.190	.000
人口素质	1817.856	2	74.363	28	24.446	.000
生活质量	3315.174	2	59.276	28	55.928	.000
法制与治安	530.188	2	76.284	28	6.950	.004

由于已选择聚类以使不同聚类中个案之间的差异最大化，因此 F 检验只应该用于描述目的。
实测显著性水平并未因此进行修正，所以无法解释为针对"聚类平均值相等"这一假设的检验。

表2.5展示了各指数(聚类变量)在不同类的均值比较情况，各数据项的含义依次为组间方差、组间自由度、组内方差、组内自由度、F统计量的观测值以及对应的概率P-值。该表显示各指数的总体均值在3类中有显著差异。应注意这里的单因素方差分析并非用于对各总体均值的对比，而需关注F值。F值大表明组间差大，组内差小，说明将数据聚成当前的K个类是合理的。而且层次聚类分析中观测所属类一旦确定就不会再改变，而K- Means聚类分析中观测的类归属会不断调整。

当前该表中所有元素的概率-P值均小于0.05，**说明将该数据聚成当前的3个类是合理的。**