第十五章 对应分析

15.1 模型简介

15.1.1 问题的提出

研究分类变量间的联系是统计分析中常见的工作，卡方检验、二分类logistic模型等是常用的方法。但是，当所涉及的分类变量类别较多，或者分类变量个数较多时，这些方法就会显得力不从心。例如，研究全国56个民族的职业分布规律，在通过抽样收集到数据后，我们可以很容易的使用卡方检验得出各民族间职业分布有差异的结论。但是，这样的结论又有什么用呢？我们更希望得知的是每个民族更倾向于从事何种职业，例如蒙古族更倾向于从事农牧业、朝鲜族更倾向于教师职业等，这种结论才是真正有用的。要得到这种精确、全面的结果，也就是对分类变量各类别间的联系进行清楚的呈现，就需要在方法学上有相应的突破。人们也一直在寻找针对此类问题的适当统计分析方法，大致而言有两种解决思路：

1．采用对数线性模型、多分类logistic模型等对类别间的联系进行精确建模，通过假设检验的方法确定分类变量各类别之间究竟有无联系。这样虽然精确，但相应模型的操作和解释都非常复杂，并非普通用户轻易能够掌握的，同时分析结果在呈现上也较为困难。

2．采用图形化呈现的方式，通过对应分析将交叉表转换为相应的对应分析图，这样虽然没有涉及到假设检验，无法得到确切的统计结论。但是结果更为直观，而且操作简单、对结果的解释也更加容易。显然，这种方式更为广大读者所喜闻乐见。

对应分析的起源较多，现在一般认为它起源于上一世纪30～40年代的一批互相独立的文献，如Richardson和Kuder（1933）、Hirshfeld（1935）、Horst（1935）、Fisher（1940）、Guttman（1941）等，很难说哪位统计学家是该方法的真正作者，有的认为是Fisher，有的则认为是French。同时因其来源众多，它的别名也较多，如同质性分析、数量化方法等。但所有这些方法的基本原理是相同的。长期以来，对应分析在法国和日本都非常流行，这部分是因国情所致，但更重要的是几位统计学家在各自国内所起的推动作用。现在，随着国内统计软件的逐渐普及，对应分析的优势越来越为人所知，在国内也正在得到越来越广泛的应用。

15.1.2 模型入门

对应分析的实质就是将行、列变量的交叉表变换为一张散点图，从而将表格中包含的类别关联信息用各散点空间位置关系的形式表现出来，但是，其中所使用的算法较为复杂，其整个分析过程可以被大致分为以下五步：

1．数据的变换与标准化

由于对应分析的主要目的是呈现出各类别间的联系，因此它首先需要对数据进行变换，使得这种联系能够被凸现出来。具体的方式是假设行、列变量间无关联，随后在绘制出两变量相应的交叉表后，基于该原假设计算各单元格内的标准化残差：

$$\text{标准化残差}=\frac{\text{观察频数}-\text{理论频数}}{\sqrt{\text{理论频数}}}$$

这样就将原始的频数阵转换为了一个新的数据阵Z。在变换后，每个单元格内的数据反映当前单元格偏离该无关联假设的程度，相应的两变量类别间联系越强，则单元格内数据的绝对值就越大，数据的正负则反映了是正向还是负向联系。对应分析随后的分析步骤就是将变换后的数据阵转换为相应的散点图而已。

随后的四步数学味较重，读者不需要完全理解，只须知道每一步的结果是什么即可。

2．奇异值分解

对矩阵Z进行奇异值分解（Singular value decomposition），公式如下：

$$Z = K\Lambda L'$$

其中 $K'K = L'L = 1$，而 Λ 则为对角阵，其中包含一些奇异值，且沿主对角线从大到小排列，每个奇异值就对应了结果中的一个维度。该步骤实际上确定了分析结果的最大维度数，以及每个维度所携带的信息量。

3．行、列尺度的调整

按照行、列变量相应的类别构成比，对K、L矩阵中包含的奇异向量进行标准化，使之具有单位长度。标化后K、L实际上就分别将行、列变量各散点的坐标确定了下来。

4．估计方差与协方差

这一步的实际含义是通过对方差、协方差的估计，初步得到各类别所对应的散点坐标。

5．行、列评分的标准化

按照所选定的标准化方法，对计算出的行、列变量坐标进行标化，该步骤进行完后，得到的就是最终在图形中呈现的散点坐标。

在这五步中，后四部基本上被固定下来，没有太多的选项可供调整，最为重要的是第一步，其变化将在第三节中详细讲解。

15.1.3 SPSS中的相应功能

对应分析可根据所分析变量的数目被分为简单对应分析和多重对应分析两种：简单对应分析用于分析两个分类变量间的联系，在SPSS中可以使用属于base模块的Correspondence Analysis过程完成。多重对应分析则用于是分析多个分类变量之间的类别联系，SPSS中提供的是基于最优尺度变换的多重对应分析，该功能由属于category模块的Optimal Scaling过程来实现。

15.2 简单分析实例

www.MedStatStar.com

这里我们以较为经典的头发颜色与眼睛颜色的研究案例来说明SPSS中对应分析的实现方法，该数据由Fisher在1940年首次引用。研究者收集了苏格兰北部Caithness郡5387名小学生眼睛与头发颜色的数据，见下表，其中眼睛有深、棕、蓝、浅四种颜色，头发有金、红、棕、深、黑五种颜色。研究者希望知道头发和眼睛的颜色间存在何种关联，即某种头发颜色的人起眼睛更倾向于何种颜色？

表15.1　头发颜色与眼睛颜色的交叉表

		头发颜色					合 计
		金色	红色	棕色	深色	黑色	
眼	深色	98	48	403	681	85	1315
睛	棕色	343	84	909	412	26	1774
颜	蓝色	326	38	241	110	3	718
色	浅色	688	116	584	188	4	1580
合计		1455	286	2137	1391	118	5387

15.2.1 对数据的初步分析

数据见文件hair&eye.sav，对于两个分类变量间是否有联系这一问题，我们可以首先对数据进行描述，以获得初步信息。除交叉表以外，统计图也是用于描述的重要工具，对该问题我们可以分别做出条图和马赛克图如下：

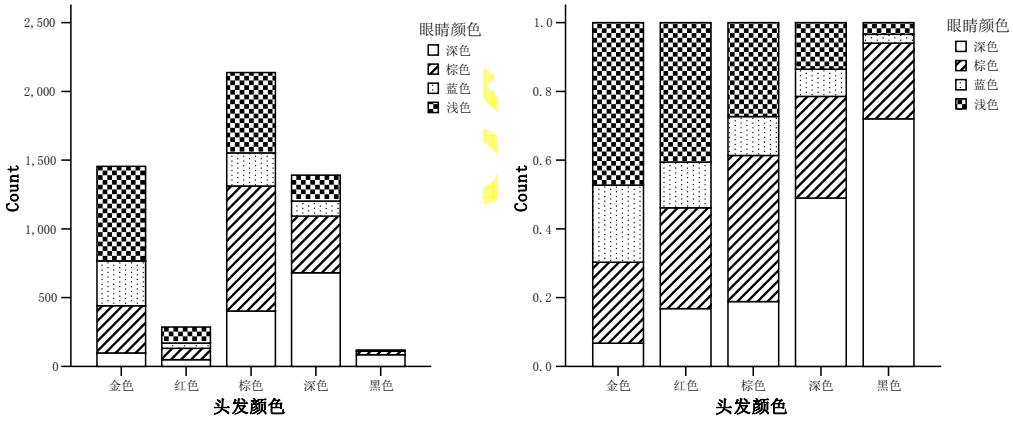


图15.1　头发颜色与眼睛颜色的条图与马赛克图

条图反映的是原始频数大小，从中可见深色头发中深色眼睛的人远比金色头发众多。但是，由于各类别的总人数并不相同，直接比较原始频数并不合理。更妥当的办法是进行各类别构成比的比较。马赛克图反映的就是这一信息。从图中我们可以看到：随着头发颜色有金色、红色逐渐变为神色、黑色，人群中眼睛颜色为浅色的比例越来越低，而眼睛深色的比例越来越高。显然，这一信息提示头发颜色和眼睛颜色之间是有关联的。

以上信息是通过对样本的直接观察而来，那么这种联系究竟是真实存在，还是由抽样误差导致的假象？这可以通过检验来加以验证。对于两分类变量的关联问题一般使用卡方检验来验证，在SPSS中使用crosstab过程实现，结果如下，可见卡方值为1240.4，P值远小于0.05，因此可以认为眼睛颜色和头发颜色间存在关联。但是，究竟是怎样的联系方

式？是其中仅某两类间存在联系，例如浅色头发的人浅色眼睛的较多；还是两两都有联系？这是卡方检验所不能回答的，我们需要采用更复杂的分析方法才能得到进一步的分析结果，而对应分析就是一个很好的选择。

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1240.039[a]	12	.000
Likelihood Ratio	1218.314	12	.000
Linear-by-Linear Association	920.631	1	.000
N of Valid Cases	5387		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.73.

15.2.2 正式分析

操作说明

下面使用对应分析对本例进行分析，由于只涉及到两个分类变量，因此可以考虑使用简单对应分析，相应的操作如下：

Analyze➔Data Reduction➔Correspondence Analysis
Row框：hair
选中hair：Define Range：Minimum框：1：Maximum框：5：Update：Continue
Column框：eye
选中eye：Define Range：Minimum框：1：Maximum框：4：Update：Continue
OK

可见该过程的操作并不复杂，只需要将相应的行、列变量选入，并确定相应的取值范围即可。简单对应分析中有许多选项，例如在model自对话框中的功能就对应了前面对应分析计算中的各个步骤。但是在多数情况下均使用默认值即可，因此这里我们均使用默认值，随后再对较为有用的功能作进一步解释。

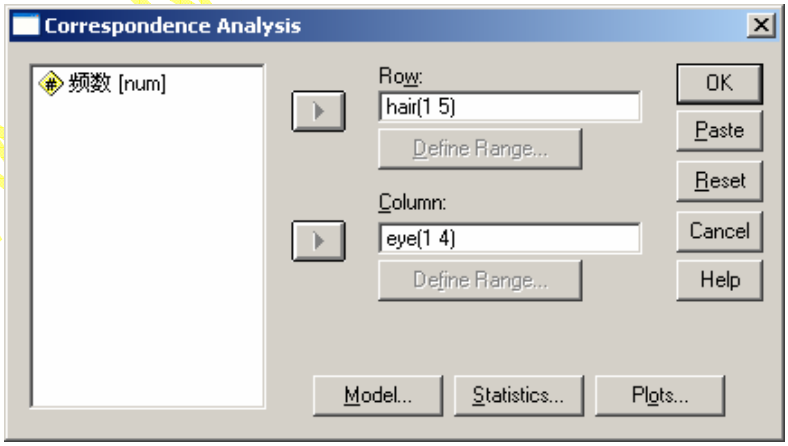


图15.2　Correspondence过程的主对话框

结果解释

对应分析的结果中首先会输出该模块的版权信息，说明该模块是由荷兰Leiden大学DTSS课题组编制的。此处略。随后才是正式的分析结果，这里我们一一解释如下：

Correspondence Table

	眼睛颜色				
头发颜色	深色	棕色	蓝色	浅色	Active Margin
金色	98	343	326	688	1455
红色	48	84	38	116	286
棕色	403	909	241	584	2137
深色	681	412	110	188	1391
黑色	85	26	3	4	118
Active Margin	1315	1774	718	1580	5387

首先输出的这张表格被称为对应分析表，但是大家可以看出，这实际上就两个变量的行×列表。由于对应分析随后的计算是完全基于该表格而来，所以首先将其输出，便于对变量间的关联进行大致的观察，也可用于检查有无数据录入错误。

Summary

					Proportion of Inertia	
Dimension	Singular Value	Inertia	Chi Square	Sig.	Accounted for	Cumulative
1	.446	.199			.866	.866
2	.173	.030			.131	.996
3	.029	.001			.004	1.000
Total		.230	1240.039	.000[a]	1.000	1.000

a. 12 degrees of freedom

上表为整个对应分析的结果汇总表（右侧有删节），在对应分析中，最多可以提取的维度数等于两变量最小类别数－1。但是，往往前两、三个维度就携带了绝大多数信息，因此可以只对前几个维度进行观察，以方便结果的解释。这里的汇总表主要用于给出所提取的每个维度（因子）所携带的信息量，从而可以帮助确定需要使用多少个维度（因子）对结果进行解释。从左到右的前六个指标依次是维数、奇异值、惯量、总的卡方检验及P值、方差解释比例，依次说明如下：

1．奇异值和惯量：奇异值这个术语来自于矩阵运算，就是前面讲对应分析计算步骤中进行奇异值分解所得到的东西。对矩阵运算原理不熟悉的朋友可以不去多考虑它。它的平方就是惯量（inertia），相当于因子分析中常说的特征根，用于说明对应分析各个维度的结果能够解释列联表中两变量联系的程度。但是，由于这里是对应分析，所以他的大小不再像因子分析中那样明白的代表该因子平均携带了多少个原始变量的信息。但是，所由维度惯量的总和则可以用来表示总信息量的大小，详后。

2．卡方检验及P值：大家从卡方值和P值大小即可看出，此处进行的卡方检验就是前面初步分析中crosstab过程所进行的Pearson卡方检验！也就是检验：行变量和列变量间是

否存在关联。因此，它可以被看成是对应分析适用条件的检验，因为只有当行变量和列变量间有关联时，我们才需要使用对应分析对这种联系加以详细分析，否则，就没有使用对应分析的必要了。

3．方差解释比例：表明每个维度所携带的信息量，实际上就是按照每个维度的惯量占惯量总和的比例计算而来，这一点上惯量倒是和因子分析中的特征根完全相同。从上表可见第一维占了总信息量的0.199/0.23=86.6%，第二维占了13.1%，第三维则仅占0.4%。三个维度加起来共携带了100％的原始信息量。显然，由于前两个维度即携带了绝大部分信息，而且二维图形要比三维图形更容易观察和理解，因此本例中完全可以使用二维空间进行分析结果的解释。

表格最右侧还会给出各奇异值的标准差及相关系数，对结果解释影响不大，可忽略。

Overview Row Points [a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
头发颜色	Mass	1	2	Inertia	1	2	1	2	Total
金色	.270	-.814	-.417	.088	.401	.271	.907	.093	1.000
红色	.053	-.349	-.116	.004	.014	.004	.770	.033	.803
棕色	.397	-.063	.500	.018	.004	.572	.039	.961	1.000
深色	.258	.881	-.250	.092	.449	.093	.969	.030	1.000
黑色	.022	1.638	-.688	.028	.132	.060	.934	.064	.998
Active Total	1.000			.230	1.000	1.000			

a. Symmetrical normalization

上表为行变量（头发颜色）各类别的分析结果概况，由于各类别均以散点的形式在空间中呈现，故称行点汇总表。表中主要给出各类别在各维度上的评分，以及相应的信息贡献量两大类信息，分述如下：

1．Mass：实际上就是各种类别的构成比，例如发色为金色的共1455人，占总数的构成比为1455/5387＝27%，依此类推。构成比的大小可以近似的反映计算出的相应指标是否稳定，因为构成比越高，说明频数越多，而相应的分析结果就越不易受个别极端样本值的影响。

2．Score in dimension：给出各类别在相关维度上的评分，首先给出的是在默认提取的两个维度上各类别的因子负荷值（空间坐标值），随后的inertia列则给出了惯量在行变量中的分解情况，它反映了总惯量（0.23）中分别由各行变量类别所提供的部分，数值越大，说明该类别对惯量的贡献越大。其大小既和相应的构成比有关，也和该类别与另一变量的关联度有关。例如黑色的构成比只有2.2%，而惯量比例则达到了0.028 / 0.23=12.2%，这说明他和列变量的关联可能较为明显。

3．Contribution：首先给出在各维度上信息量在各类别间的分解情况，本例中可见第一维度的信息主要被金色、深色和黑色三个类别所携带，反言之，这三个类别在第一维

度上的区分度较好，这从坐标值上即可得到验证。同理，在第二维度上则是金色和棕色的区分度较好。随后给出的是各类别的信息在各维度上的分布比例，例如金色的总信息量中有90.7%分布在第一维上，只有9.3%分布在第二维上。综合观察，可知除了棕色外，绝大多数类别的信息都分布在第一维上。最右侧给出的是各维度的信息比例之和，可见红色这一类别在前两维中只提取出了80.3%的信息量，因此如果红色的解释不理想或不合理，则可考虑是否加入第三维度以改善解释。

Overview Column Points [a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
眼睛颜色	Mass	1	2	Inertia	1	2	1	2	Total
深色	.244	1.052	-.322	.125	.605	.145	.965	.035	1.000
棕色	.329	.050	.588	.020	.002	.657	.018	.981	.999
蓝色	.133	-.599	-.397	.026	.107	.121	.836	.143	.979
浅色	.293	-.660	-.212	.060	.286	.076	.956	.039	.995
Active Total	1.000			.230	1.000	1.000			

a. Symmetrical normalization

上表为列变量各类别分析情况的汇总，阅读方式和前面相同，不再重复解释，大家只须注意惯量总和也是0.23，即行、列变量的分解是在相同的解释空间中进行的，故此，相应的类别散点才能被放在同一个空间中加以阅读。

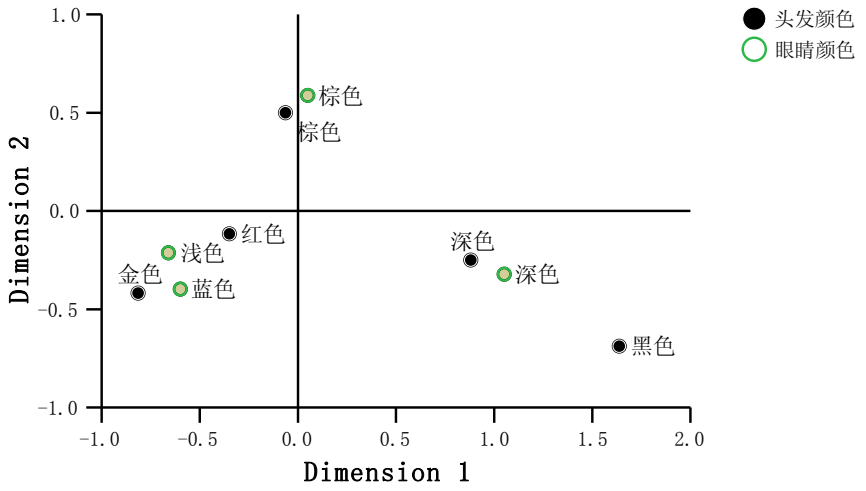


图15.3　对应分析图

分析结果中最后给出的是对应分析图（为便于阅读，此处已经过编辑），实际上对于对应分析而言，由于所有主要信息均反映在了该图形中，各类别散点在空间中的距离和位置就反映了各自间的关系，因此多数分析报告均只使用这张图进行描述。阅读该图形主要可以了解同一变量各类别的区分程度，以及不同变量各类别间的关联程度如何，因此对应分析图的阅读可按如下顺序进行：

1．考察同一变量的区分度：首先分别考察行变量、列变量各类别间是否被清晰的分开了，可以分别检查在各个维度上的区分情况，如果同一变量不同类别在某个方向上靠得较近，则说明这些类别在该维度上区别不大。在本例中，可以看到无论是头发颜色还是眼睛颜色在空间位置上都分得比较开。分维度考察，则可以注意他们在第二维度上的区分度稍差一些。

2．考察不同变量的类别联系：这才是对应分析所真正关心的问题！一般而言，落在从图形原点(0,0)处出发相同方位上大致相同区域内的不同变量的分类点彼此有联系。散点间距离越近，说明关联倾向越明显；散点离原点越远，也说明关联倾向越明显。

下面我们就根据以上原则对图形结果加以解释：

1．眼睛棕色和头发棕色两个散点靠得非常近。显然这两种特征之间存在关联。

2．同理，我们也容易确定眼睛深色和头发黑色存在关联。但是头发黑色是否也有关联？它离眼睛深色有一定的距离，但是要注意到如果考察相对于原点的位置，则头发黑色这一散点基本上在原点和眼睛黑色散点连线的延长线上，因此可以认为头发神色、黑色都和眼睛深色有联系。

3．头发金色和眼睛蓝色、浅色的散点在一起，因此可以认为它们之间存在联系。虽然头发红色的散点空间位置也在它们附近，但是注意到该散点距离原点较近，因此对它的解释要小心，即它的联系可能比较弱。

这样，通过对应分析图，我们就可以非常直观而简明的得到头发颜色和眼睛颜色间的关联特征，显然，借助于图形化的结果，对应分析要比对数线性模型等建模方法更容易理解和应用。

15.2.3 对引例的进一步分析

加入更多维度

前面的分析结果共使用了两个维度，他们累计携带了99.6%的信息量，应当说已经非常充分了。但考虑到现有结果中对头发红色的解释度较弱，而该类别尚有一些信息在第三维度中，因此可以要求将第三维度也加入解释。该选项在model子对话框中更改，最上方的dimensions in solution框中用于指定希望提取的维度数，默认为两维，更改为三维后分析结果中即会输出相应的三维分析结果，此处给出第三个维度的图形（经过编辑）如下：

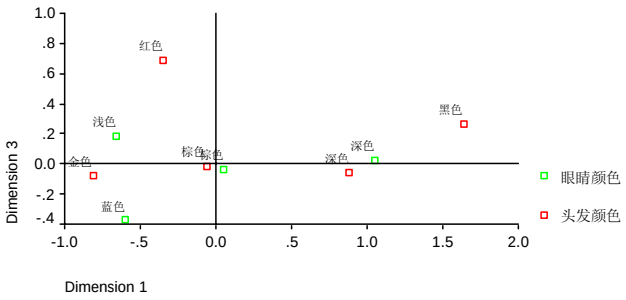


图15.4　第一、三维度的对应分析图

可见在第三个维度中，红色和其他各类别的距离均比较远，该维度提供的多余信息并未能改善对头发红色与眼睛颜色间关联性的解释。考虑到本身第三个维度的信息量就非常少，因此对于本问题而言，二维的解释是较为合理和充分的。

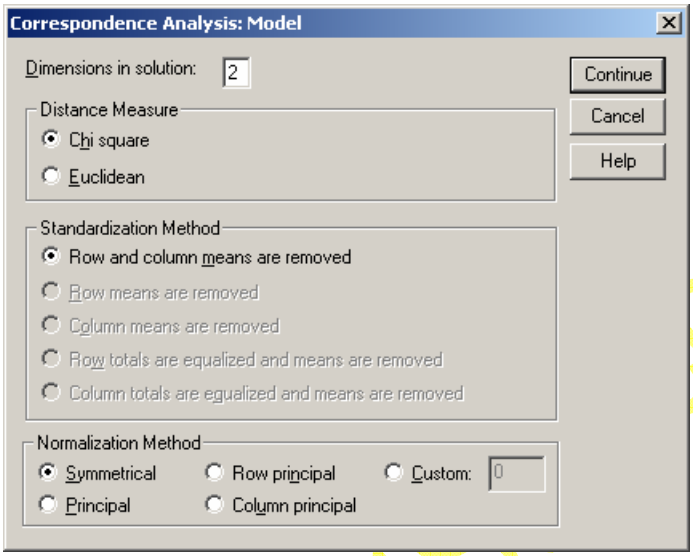


图15.5　Model子对话框

最优对应表

当参与对应分析的变量其类别间可能存在某种内在的次序关系时，分析者往往希望能够在表格中直接观察到这种次序，即各类别按照这种关系从达到小依次排列。对应分析中可以提供这种输出，由于各类别在各维度上都已算出相应的坐标值，因此只需要将各类别按照坐标值从小到大依次排列即可。相应的选项为Statistic子对话框中部的Permutations of the correspondence table复选框，生成的排序表格被称为permuted table。本例中的最优对应表如下：

Permuted Correspondence Table According to Dimension 1

	眼睛颜色				
头发颜色	浅色	蓝色	棕色	深色	Active Margin
金色	688	326	343	98	1455
红色	116	38	84	48	286
棕色	584	241	909	403	2137
深色	188	110	412	681	1391
黑色	4	3	26	85	118
Active Margin	1580	718	1774	1315	5387

联系各类别相应的第一维坐标，读者就可以发现眼睛颜色和头发颜色的各类别都是按照坐标值从小到大进行排列，可以发现眼睛颜色似乎是从浅到深、而头发颜色则是从金、红到深、黑，更为重要的是，表格中频数会集中在主对角线上，使得对应关系要比原来清楚一些。如现在可以立刻发现黑色头发和眼睛深色、棕色有较强的关联。

Permutations of the correspondence table复选框下部的Maximum dimension for框用于指定希望使用几个维度的坐标生成最优对应表，默认只使用一维。如果指定两维以上，则会对每个维度都生成一个单独的最优对应表。

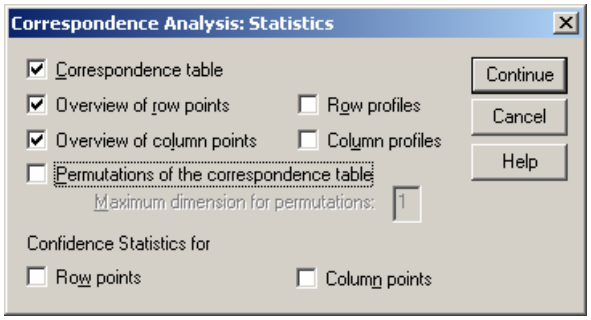


图15.6　Statistics子对话框

散点坐标的可信区间

在前面我们提到过，对应分析只是一个统计描述方法，它仅仅是将表格信息转化为图形的方式来展现。但许多时候我们都希望相应的分析结果在统计上能较为有力，而不仅仅是一种“可能”的情形。为此，对应分析过程中提供了计算各散点坐标可信区间的功能，通过观察各类别可信区间的大小及重叠程度，我们就可以对分析结果得到更清晰明了的认识。该功能位于Statistic子对话框的最下方，Confidence statistic复选框组由Row points和column points两个复选框构成，分别用于计算行、列类别散点的坐标可信区间，这里以行变量为例，相应的分析结果如下：

Confidence Row Points

	Standard Deviation in Dimension		Correlation
头发颜色	1	2	1-2
金色	.026	.028	-.452
红色	.083	.133	.088
棕色	.036	.021	.078
深色	.023	.026	.502
黑色	.058	.069	.404

注意表格中给出的是各维度上坐标的标准差（实际上是标准误），通过它就可以计算出相应坐标的95%可信区间，相对应的矩形区域就是总体中该散点的95%可能分布范围。但是，如果右侧行／列坐标的相关系数较大，则建议计算行、列坐标的联合二元可信区间，这样结果会更为准确。

15.3 基于均数的对应分析

15.3.1 方法原理

15.3.2 分析实例

15.4 多重对应分析

15.4.1 方法原理

15.4.2 分析实例

操作说明

结果解释

15.5 对应分析中的其它问题

15.5.1 对应分析结果的正确解释

15.5.2 罕见类别和相似类别的处理

15.5.3 有序类别的处理

15.6 本章方法小结

15.6.1 对应分析与其它分析方法的关系

对应分析与卡方检验

对应分析与典型相关分析的等价性

对应分析与因子分析的关系

15.6.2 对应分析的优势与劣势

对应分析的优势

对应分析的劣势

思考与练习

1．92年美国大选时中出现了三位候选人，最终是克林顿击败了老布什和佩罗当选总统，那么不同教育程度的选民其倾向性如何？SPSS自带数据集voter.sav是一部分抽样调查数据，变量pres92纪录受访者选择了哪位候选人，degree则为受访者的文化程度，请使用对应分析考察不同文化程度的选民其倾向如何，并回答下列问题：

1）对于该数据，解释空间的维度设定为几维比较合适？一维可以吗？

2）如果希望简化结果解释，可以进行哪些操作？

3）对照原始交叉表中的构成比和图形分析的结果，仔细体会对应分析结果的实际含义是什么。

2．打开数据文件hair&eye，进行操作，并回答以下问题：

www.MedStatStar.com

1）使用同质性分析方法对该数据进行拟合，将分析结果和简单对应分析相对照，理解两种方法在分析结果上的联系。

2）使用典型相关对数据进行分析，仔细对照典型相关分析和对应分析的结果，理解两者间的区别和联系。

3．打开数据文件meancores.sav，对其进行因子分析和对应分析，并回答如下问题：

1）本例中提取几个公因子比较合适？

2）对照因子分析的解释空间和对应分析的解释空间，两者的异、同点在哪里？

3）采用不同的空间距离标化方法，对照各种方法所得到的对应分析结果，深入理解各种标准方法的作用，以及对分析结果的影响。

4．打开数据文件mcorres.sav，在分析中将结果空间指定为三个维度，考察第三个维度主要对结果进行了怎样的改善，并确认是否应当将该维度加入解释。

参考文献

1. Perceptual Mapping Using SPSS Categories (v8.0 Revised). SPSS Inc. Chicago, Illinois, 1998
2. Richard Johnson著，陆璇译. 实用多元统计分析. 第四版. 北京：清华大学出版社. 2001
3. 方积乾主编. 医学统计学与电脑试验. 第二版. 上海：上海科学技术出版社. 2001
4. 郭志刚主编. 21世纪社会学系列教材——社会统计分析方法：SPSS软件应用. 北京：中国人民大学出版社. 1999

www.MedStatStar.com