对应分析

相关概念

1. 列联表
2. 对应矩阵

	1	2	...	j	...	q	合计（行边缘）
1	p_{11}	p_{12}	...	p_{1j}	...	p_{1q}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2j}	...	p_{2q}	$p_{2\cdot}$
...	...	...	...	...	...	...	...
i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{iq}	$p_{i\cdot}$
...	...	...	...	...	...	...	...
p	p_{p1}	p_{p2}	...	p_{pj}	...	p_{pq}	$p_{p\cdot}$
合计（列边缘）	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot j}$	...	$p_{\cdot q}$	1

对应矩阵相当于将列联表中的数据计算频率对应的结果

另外，最后一列记为$r = P1 = (p_{1\cdot}, \ldots, p_{p\cdot})'$，即表示对该向量记录了每个row的元素和；最后一行记为$c = 1'P = (p_{\cdot 1}, \ldots, p_{\cdot q})$，即表示对该向量记录了每个column的元素和

称r, c分别为*行**边缘频率***，*列边缘频率*；或*行密度*，*列密度* *（或质量）*

3. 行/列轮廓

- 称对应矩阵中，某行每个元素都概率值/该行合计概率组成的向量为该行的*行轮廓*，记为r_i，即 $r_i = (p_{i1}/p_{i\cdot}, \ldots, p_{iq}/p_{i\cdot})'$
- 同理，称某列每个元素都概率值/该列合计概率组成的向量为该列的*列轮廓*，记为c_j，即 $c_j = (p_{1j}/p_{\cdot j}, \ldots, p_{pj}/p_{\cdot j})$

（由列联表和对应矩阵的关系可知行列轮廓也可以通过列联表类似求得）

可以证明，某一行/列轮廓的各元素之和为1

注：个人理解，轮廓的这个操作也类似于一个归一化，相当于把每行/列整理成一个100%的数值，再看每个元素在这行的占比是多少。它在某种程度上也反应了数据的一种分布占比情况

4. 对应矩阵、边缘频率、行列轮廓的计算关系
   如下的计算关系都可以通过简单的矩阵运算得到，这里仅给出结论
   记$D_r = \mathrm{diag}(p_{1\cdot}, \ldots, p_{p\cdot})$，$D_c = \mathrm{diag}(p_{\cdot 1}, \ldots, p_{\cdot q})$
   记$R = (r_1', \ldots, r_p')'$，$C = (c_1, \ldots, c_q)$
   则 $R = D_r^{-1}P, C = PD_c^{-1}$
   这个推到仅仅是一种记号上的简单关系，例如对对应矩阵中的每个元素除以该行的边缘概率，就得到了行轮廓矩阵；对每个元素除以该列的边缘概率，就得到了列轮廓矩阵。其实只要知道这两个东西的定义，都可以自

由推知。

这一部分的概念定义由于有许多矩阵，相应元素的记号可能有点绕。实际上将例题的数据在Excel中计算一下即可轻松理解，对于具体数据而言还是很友好的，记清楚每个概念的含义即可

独立性检验（卡方检验）

- 检验的内容为：行变量与列变量之间是否是独立的（例如对于例题9.2.1给出的列联表，相当于检验心理健康状态与父母的经济地位之间是否存在联系）
- H_0：行列是独立的
- 检验思路：对于列联表中的每一个数据元素，我们有一个该数据的真实存在频率p_{ij}，又有该行、列分别的边缘概率$p_{i\cdot},p_{\cdot j}$。如果真的是独立的，那么$p_{ij}\approx p_{i\cdot}\cdot p_{\cdot j}$。因此对二者进行做差，相差越大则说明原假设不成立的可能性越大。如此遍历列联表中的每个数据，就可以得到整体的一个独立与否的大致判断。
- 检验统计量：
  $\chi^2=n\sum_i\sum_j\frac{(p_{ij}-p_{i\cdot}\cdot p_{\cdot j})^2}{p_{i\cdot}\cdot p_{\cdot j}}$
- 拒绝域：$\chi^2\geq\chi^2_\alpha((p-1)(q-1))$
- 适用性要求：n足够大、$np_{i\cdot}\cdot p_{\cdot j}\geq 5$

总惯量(inertia)

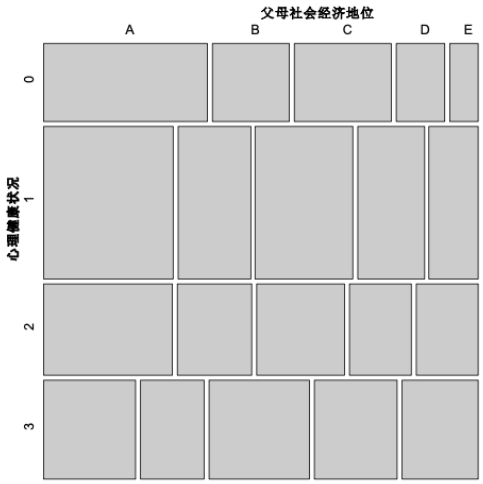
$$\text{inertia}=\frac{\chi^2}{n}=\sum_i\sum_j\frac{(p_{ij}-p_{i\cdot}\cdot p_{\cdot j})^2}{p_{i\cdot}\cdot p_{\cdot j}}$$

- 总惯量是一种衡量行列变量之间关联程度的指标，其值越大，说明真实情况和独立假设相差越大，说明行列变量之间的关联程度越大
- 总惯量的表达式还可以通过 边缘频率、行列轮廓等之间的关系进行进一步推导

总惯量为0的情况
总惯量为0相当于行列之间基本是独立的（因为相当于和理论假设独立情况的分布结果无差别），而这又也下列三种叙述等价：

1. $p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$
2. 所有的行轮廓相同
3. 所有的列轮廓相同

该点通过轮廓来考虑也是非常自然的。以例题的这个行轮廓马赛克图为例。行轮廓马赛克图中，A～E的宽度表示对应不同心理状况行的行轮廓值。如果所有行轮廓相同（例如想象四行都想心理健康=1、2两行那样），那么就说明心理状况无论是多少，经济地位都是有类似的分布，那么也就是说心理健康状况完全不会影响父母的经济地位（在统计学的意义而非因果意义上），二者是独立的。而又由于独立的反身性，因此反之亦然。（通过这一例子也正好了解一下该马赛克图的读图方法）



标准化后的总惯量

若对对应矩阵进行标准化，即$Z = (\frac{p_{ij}-p_{i\cdot}\cdot p_{\cdot j}}{\sqrt{p_{i\cdot}\cdot p_{\cdot j}}})$，通过奇异值分解（$Z = U\Lambda V'$）等代数运算可知：
$\text{inertia} = \sum_{i=1}^{k} \lambda_i^2$其中$\lambda_i^2$为$ZZ'$的正特征值

简而言之，对于标准化后的对应矩阵Z，总惯量等于Z的奇异值之和，而奇异值之和相当于$Z'Z$的正特征值之和。

轮廓坐标与对应分析图

行/列坐标

在上面的对于标准化对应矩阵Z的奇异值分解中，我们得到$Z = U\Lambda V'$，由奇异值分解的数学意义可知，这里的U, V分别可以认为是一组基向量组成的矩阵。

> *回忆记号：*
>
> $r = P1 = (p_{1\cdot}, \ldots, p_{p\cdot})'$，$c = 1'P = (p_{\cdot 1}, \ldots, p_{\cdot q})$，分别表示**边缘概率**，是对应矩阵每一行（列）的概率求和
>
> $D_r = \text{diag}(p_{1\cdot}, \ldots, p_{p\cdot})$，$D_c = \text{diag}(p_{\cdot 1}, > \ldots, p_{\cdot q})$，分别为行列**边缘概率**组成的**对角矩阵**
>
> $r_i = (p_{i1}/p_{i\cdot}, \ldots, p_{iq}/p_{i\cdot})'$，$c_j = (p_{1j}/p_{\cdot j}, \ldots, p_{p>j}/p_{\cdot j})$，分别表示**行（列）轮廓**，相当于是一个条件概率，数值上为对应矩阵中的某一概率除以该行（列）的边缘概率

则再加上上述奇异值分解结果，我们可以得到两组基：
$A_{p\times k} = (a_1, \ldots, a_k) = D_r^{\frac{1}{2}}U$

$B_{q\times k} = (b_1, \ldots, b_k) = D_c^{\frac{1}{2}}V$

用这两组基，即可分别表示前面得到的（中心化后的）行、列轮廓，具体地：

第i行轮廓：
$r_i' - c' = x_{i1}b_1' + \cdots + x_{ik}b_k'$

第j列轮廓：
$c_j - r = y_{j1}a_1 + \cdots + y_{jk}a_k$

这里的a_i, b_j为基，则对应的x_i, y_j就是行/列轮廓对应的坐标。有时也会按照几何意义上坐标上点 称行/列轮廓为行/列点。

通过数学计算还可以证明，行轮廓坐标依照行边缘概率进行加权平均后为0；列轮廓坐标按照列列边缘概率加权亦为0

第i惯量

数学上还有如下等式成立：
$\sum_{j=1}^{p} p_{j\cdot} x_{ji}^2 = \sum_{j=1}^{q} p_{\cdot j} y_{ji}^2 = \lambda_i^2 \quad (i = 1, \ldots, k)$

其具体含义为，第i行轮廓坐标的平方按照行边缘概率加权平均后等于λ_i^2。而这里的λ_i^2不是别的，恰恰是总惯量中的$\text{inertia} = \sum_{i=1}^{k} \lambda_i^2$。

因此，我们可以将λ_i^2称为第i惯量。换言之，总惯量可以被拆分为各个惯量的和。而这些分惯量对于行/列来说都是广义上对称等价的，都等于各自的坐标平方按照各自边缘概率加权平均的结果。
（有点类似于对于一组数据，均值为0，而总的方差可以进行分解的感觉；因为如果直接进行加权平均就相当于是求期望（一阶矩），平方再求期望就类似于求方差并对方差进行分解（二阶矩））

对应分析图

（此处为ht授课重点）

对应分析图的构建

回到上面总惯量的讨论中，可以发现总惯量是一系列分惯量（相当于是在一系列坐标轴下）的和，而这里的λ_i^2是从大到小进行排列的，类似于PCA的降维理念，我们也可以通过选取前几项最为重要的惯量来近似替代总体的惯量信息，以起到精简的作用。

这里主要选取前两个惯量，即λ_1^2, λ_2^2，来近似替代总惯量，从而得到一个二维的坐标系。具体构建时，将$a_1 b_1$ 重合在一起，作为一个坐标轴，$a_2 b_2$重合在一起，作为另一个坐标轴，这样就得到了对应分析图。

对应分析图的理解

对应分析图中几何上的两点欧氏距离在数学上相当于两个轮廓之间的卡方距离。

- 同类点接近（行点-行点 / 列点-列点接近）：
  - 对应分析图中的点越近，说明对应的两个轮廓越相似，反之则越不相似。
- 异类点接近（行点-列点接近）：
  - 两个点越接近，说明该行×该列交叉项的那个数据的实际频数高于独立假定下相乘的频数，反之则低于。
  - 但是这种关系还与二者与原点距离有关。距离原点近的，这种相关性不强；距离远点越远，这种关联性会越强，实际高于理论频数的差额就会越大。