第十六讲 对应分析（SVD的应用）

2024.5.6

列联表是一种强度矩阵，恰当的标准化是列联表分析的关键。

内容

1. 预备知识：SVD 与双标图
2. 预备知识：列联表及其中心标准化
3. 对应分析

1. 预备知识： SVD与双标图

经典双标图

第12讲的双标图与SVD有关。假设数据矩阵$X_{n\times p}$ 以及中心化，假设其SVD为$X_{n\times p} = U_{n\times r}D_{r\times r}V^{\top}{}_{r\times p}$，其中$UD$为所有样本的$r$个主成分。

以2阶SVD近似替代原始数据

$$X \approx U_2 D_2 V_2^{\top} = (\mathbf{u}_1, \mathbf{u}_2)\begin{pmatrix}\sqrt{\lambda_1} & \\ & \sqrt{\lambda_2}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1^{\top} \\ \mathbf{v}_2^{\top}\end{pmatrix} = CR^{\top}$$

其中

$$C_{n\times 2} = \left(\mathbf{u}_1\sqrt{\lambda_1}, \mathbf{u}_2\sqrt{\lambda_2}\right),\ R_{p\times 2} = (\mathbf{v}_1, \mathbf{v}_2)$$

分别为n个样本点的最优的前两个主成分表示和p个变量的二维箭头表示。

注意：$C_{n\times 2} = \left(\mathbf{u}_1\sqrt{\lambda_1}, \mathbf{u}_2\sqrt{\lambda_2}\right)$是样本点最优的2维表示，$R_{p\times 2} = (\mathbf{v}_1, \mathbf{v}_2)$不是变量的最优表示，在biplot中以箭头表示之。

根据对称性，变量最好的二维表示为：

$$R_{p\times 2} = \left(\mathbf{v}_1\sqrt{\lambda_1}, \mathbf{v}_2\sqrt{\lambda_2}\right)$$

我们可称之为“**主样本**”，若以此表示变量，

$$X \approx U_2 D_2 V_2^{\top} = (\mathbf{u}_1, \mathbf{u}_2)\begin{pmatrix}\sqrt{\lambda_1} & \\ & \sqrt{\lambda_2}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1^{\top} \\ \mathbf{v}_2^{\top}\end{pmatrix} = CR^{\top}$$

则我们以$\mathbf{C} = (\mathbf{u}_1, \mathbf{u}_2)$箭头/方向表示样本。

SVD	对偶关系	列	行	双标图(前两列)
$X = UDV^{\top}$	$Y = XV = UD$	主成分	个体	$Y = UD$
$X^{\top} = VDU^{\top}$	$Z = X^{\top}U = VD$	主样本	变量	$Z = VD$或$Z = V$

其它双标图

R软件中biplot函数将2阶SVD的对角阵拆分（缺省值$s = 1$），以

$$C = \left(\mathbf{u}_1\lambda_1{}^{s/2}, \mathbf{u}_2\lambda_2{}^{s/2}\right), R = \left(\mathbf{v}_1\lambda_1{}^{(1-s)/2}, \mathbf{v}_2\lambda_2{}^{(1-s)/2}\right)$$

分别表示样本和变量。

2. 预备知识-列联表及其中心标准化

列联表

假设两个属性变量x, y各有I, J个类别或水平。记x_{ij}为x取值i、y取值j 的样本频数， $X_{I\times J} = (x_{ij})$称为$I \times J$ 列联表(contingency table) 或交叉分类表(cross tabulation)。记

$$p_{ij} = P(x = i, y = j),\ P_{I\times J} = (p_{ij})$$

通常假设所有计数服从服从多项分布

$$X = (x_{ij}) \sim M_{IJ}(n, P),\ \sum x_{ij} = n,\ \sum p_{ij} = 1$$

	x_{ij}		

$r_i = \sum_{j=1}^{J} x_{ij}$

$c_j = \sum_{i=1}^{I} x_{ij}$

有时，根据实际设计，也可假设每一行/列服从独立的J/I项分布（参见例1评注）。

例1. 二元伯努利随机向量$(x_i, y_i), i = 1, \dots, n$ iid,交叉分类

		y		
		1	0	
x	1	x_{11}	x_{10}	r_1
	0	x_{01}	x_{00}	r_0
		c_1	c_0	

其中计数(count)

$$x_{uv} = \#\{i: x_i = u, y_i = v\}, u, v = 1,0$$

记

$$p_{uv} = P(x_i = u, y_i = v), \qquad u, v = 1,0$$

$$\Rightarrow (x_{11}, x_{10}, x_{01}, x_{00}) \sim M_4\big(n, (p_{11}, p_{10}, p_{01}, p_{00})\big),$$ 四项分布

独立性假设H_0：x, y独立。Pearson卡方检验：

$$\chi^2 = \frac{n(x_{11}x_{00} - x_{10}x_{01})^2}{r_1 r_0 c_1 c_0} = nr^2$$

r: 样本相关系数

齐一性假设和独立性假设的检验形式相同，都是Pearson卡方（参见下页）。

评注：容易验证,给定行边际总数时，每行都服从独立二项分布

$$x_{11}|_{x_{11}+x_{10}=\,r_1} \sim B(r_1, p_1), x_{01}|_{x_{01}+x_{00}=\,r_0} \sim B(r_0, p_0),$$

其中$p_1 = p_{11}/\,(p_{11}+p_{10}), p_0 = p_{01}/\,(p_{01}+p_{00})$.

反之，两组二项分布问题也可生成上述列联表，并可应用同样的卡方检验。

两个二项分布的相等性/齐一性检验

假设两组伯努利变量：

第一组: $y_i, i = 1, \ldots, r_1$ iid, $a = \sum_{i=1}^{n_1} y_i \sim B(r_1, p_1)$

第二组: $y_i, i = n_1 + 1, \ldots, r_1 + r_0 = n$ iid, $c = \sum_{i=n_1+1}^{n_1+n_0} y_i \sim B(r_0, p_0)$

齐一性假设H_0：$p_1 = p_0$.

记 $b = r_1 - a, d = r_0 - c$, 列表如右，该表仍可以看作是交叉分类表（每个样本定义组号x_i）。

H_0：$p_1 = p_0$的检验仍是Pearson 卡方。

		y		
		1	0	
第一组	1	a	b	r_1
第二组	0	c	d	r_0

两个二项分布

一般表格有类似结论，参见 下页

IJ-项分布与I个J-项分布

类似于例1，一般的$I \times J$ 列联表既可以是两个变量x, y（各有I，J个水平）生成的交叉分类表（IJ-项分布），独立性假设

$$H_0: x, y\text{独立}$$

$I \times J$ 列联表也可由I个独立的J-项分布生成, 每一行的J个计数联合服从J-项分布:

$$\left(x_{ij}, j = 1, \dots, J\right) \sim M_J(n_i, \mathbf{p}_i), i = 1, \dots, I$$

$\mathbf{p}_i = \left(p_{i1}, \dots, p_{iJ}\right)$, $p_{i1} + \cdots + p_{iJ} = 1$

齐一性假设(I个概率分布相同)：

$$H_0: \mathbf{p}_1 = \cdots = \mathbf{p}_I$$

多项分布的性质：

若$(x_1, \dots, x_m) \sim M_m(n, (p_1, \dots, p_m))$, $p_1 + \cdots + p_m = 1$. 对任何下标集$S = \{i_1, \dots, i_k\} \subset \{1, \dots, m\}$，$x_S = \left(x_{i_1}, \dots, x_{i_k}\right)$

$$x_S|_{x_{i_1}+\cdots+x_{i_k}=N} \sim M_k(N, \left(p_{i_1}, \dots, p_{i_k}\right)/(p_{i_1} + \cdots + p_{i_k})),$$

且x_S与x_{-S}条件独立

几乎

IJ-项分布 $\Longleftrightarrow$ I个独立的$J-$项分布，

$\Rightarrow$独立性检验和齐一性检验形式完全相同。

为什么称为contingency table？

Contingency：一个事件可能引发的事件(依赖性)、应急、临时、附加条款。

Pearson将属性变量之间的关联性度量称为contingency。

Pearson（1900)提出了(一个）属性变量概率分布（即多项分布）的拟合优度Pearson卡方检验。

Pearson(1904)提出以Pearson卡方χ^2作为两个属性变量的关联性的检验，以χ^2/n度量关联性大小。为了与连续变量情形的相关系数概念区分，Pearson将属性变量之间的关联性度量称为contingency。计数表格称为contingency table。

- K. Pearson (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. 现代统计的开端
- K. Pearson (1904). On Contingency and its Relation to Association and Normal Correlation.

Pearson卡方检验

对于$I \times J$ 列联表，独立性或齐一性的Pearson卡方检验统计量

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(x_{ij}-E_{ij})^2}{E_{ij}},\ E_{ij} = r_i c_j / n$$

其中$r_i = \sum_{j=1}^{J} x_{ij}, c_j = \sum_{i=1}^{I} x_{ij}$ 分别是行和、列和， E_{ij}是原假设 成立时$\mathrm{E}(x_{ij})$的估计。

在独立性或齐一性假设下，近似地$\chi^2 \sim \chi^2_{(I-1)(J-1)}$。

我们将在后面说明为什么这样构建检验统计量以及为什么渐近服从卡方分布

列联表的独立性或齐一性Pearson卡方检验是一种**整体上**的omnibus检验。对应分析更关心表格的细节特点：已知x, y不独立条件下，属性变量x 的哪些类别与y的哪些类别联系更紧密？

列联表的中心标准化

列联表矩阵与一般的数据矩阵略有不同，其元素非负，代表了关联大小，与此类似的是强度矩阵（intensity）、丰度矩阵(abundance）、网络的邻接矩阵等矩阵，矩阵(i,j)元素x_{ij}代表行标i与列标j的联系紧密程度，恰当的标准化对于此类矩阵分析尤其重要。

例2. $n = 781$件出土陶器按考古地点(0-6)和类型(a-d)两个属性变量交叉分类得到下述列联表。(i,j)格子元素x_{ij}为地点i出土的第j类陶器的个数/计数。

	a	b	c	d	总计
0	30	10	10	39	89
1	53	4	16	2	75
2	73	1	41	1	116
3	20	6	1	4	31
4	46	36	37	13	132
5	45	6	59	10	120
6	16	28	169	5	218
总计	283	91	333	74	781

不同考古地点代表不同的时期或文化，因此陶器类型分布相似的考古地点，其年代可能接近，我们甚至可以通过研究陶器类型分布判断考古地点的年代次序关系和交流情况。

关心的问题：

- 独立性：考古地点与陶器类型是否存在关联？ Pearson 卡方
- 齐一性：不同考古地点的陶器分布是否相同？不同的陶器个各个考古地点分布是否相同？Pearson 卡方
- 考古地点0和1出土陶器类型是否相似？a,b 陶器在各个考古地点出现的机率/分布是否大致相同？
- 哪些地点的哪些陶器比较多？对应分析

	a	b	c	d	总计
0	30	10	10	39	89
1	53	4	16	2	75
2	73	1	41	1	116
3	20	6	1	4	31
4	46	36	37	13	132
5	45	6	59	10	120
6	16	28	169	5	218
总计	283	91	333	74	781

2号遗址比1号遗址有更多的a陶器(73>53)？需要参考出土陶器总数：53/75 > 73/116

比较比率而不是计数

对于列联表$X_{I\times J}=(x_{ij})$，定义

- 行和： $\mathbf{r}=X\mathbf{1}_J$， $D_r=diag(\mathbf{r})$;
- 列和: $\mathbf{c}=X^{\top}\mathbf{1}_I, D_c=diag(\mathbf{c})$
- 总和(样本量)： $n=\mathbf{r}^{\top}\mathbf{1}_I=\mathbf{c}^{\top}\mathbf{1}_J=\mathbf{1}_I{}^{\top}X\mathbf{1}_J$

行归一化 列归一化

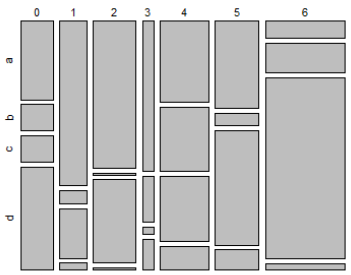
衡量计数x_{ij}的大小,需要考虑其所在行的总和r_i，或所在列的总和c_j:

- 行归一化：$P=D_r^{-1}X$， $p_{j|i}=x_{ij}/r_i, \sum_j p_{j|i}=1$(行条件概率分布)
- 列归一化：$Q=XD_c^{-1}$， $q_{i|j}=x_{ij}/c_j, \sum_j q_{i|j}=1$(列条件概率分布)

马赛克图

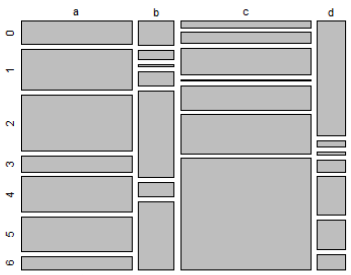
Mosaic plot（马赛克图）描述行或列条件概率分布，高度等于概率，宽度与边际总和成正比。

各个考古地点4种陶器的分布



0号遗址4种陶器的分布，高度代表概率，宽度与该遗址陶器总数89成正比。

各种陶器的在7个考古地点的分布



a类陶器在7个地点的分布。

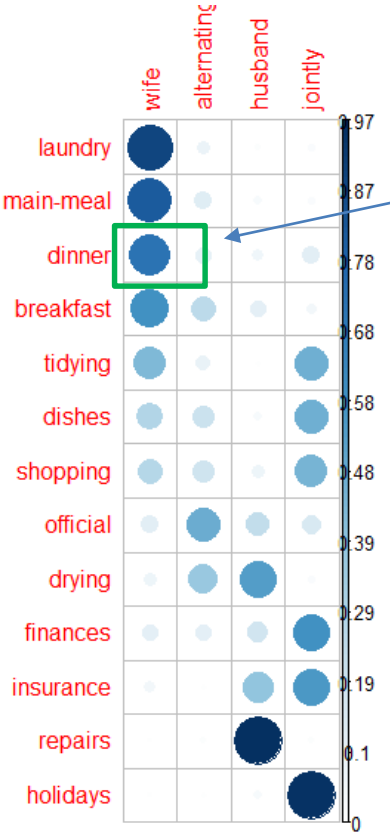
例3（家务分配）一项社会问卷调查询问了1774个家庭的13项家务活分配情况，共有4种分配方案：wife，alternating，husband，jointly. 原始数据如下：

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

比如，有156个家庭的妻子负责洗衣(laundry)，14个家庭轮流(alternating).

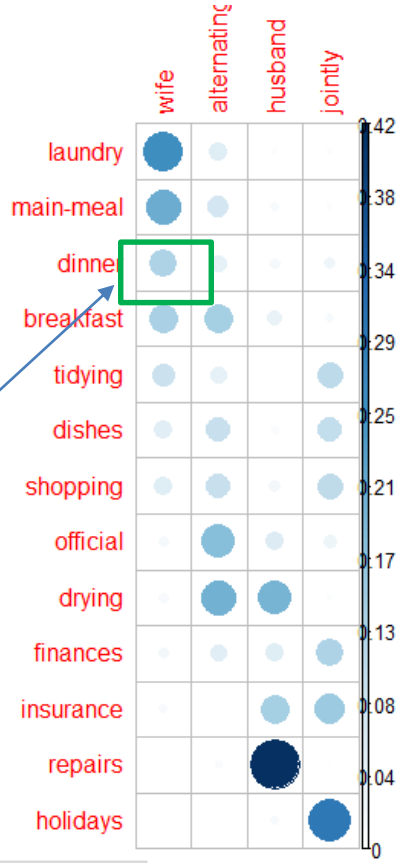
行归一化（行和为1），前4行类似，都是同一类型的家务(圆圈大小表示概率），与马赛克图类似，更直观：

列归一化（列和为1）：



(dinner, wife):
左图第三行， 在dinner的4种完成方式中，wife完成的比例最大

而右图圆圈较小,在wife的家务活中，dinner占有稍小的（第4大）比例。



Key：对应关系/圆圈的大小要结合左右两个图来看，这正是对应分析方法的关键，即同时考虑行和与列和。

- 行和： $\mathbf{r} = X\mathbf{1}_J$， $D_r = diag(\mathbf{r})$;
- 列和: $\mathbf{c} = X^{\top}\mathbf{1}_I, D_c = diag(\mathbf{c})$
- 总和(样本量)： $n = \mathbf{r}^{\top}\mathbf{1}_I = \mathbf{c}^{\top}\mathbf{1}_J = \mathbf{1}_I{}^{\top}X\mathbf{1}_J$

标准化

衡量x_{ij}的大小既要考虑其所在行的总和r_i，也要考虑所在列的总和c_j，综合$p_{j|i}$和 $q_{i|j}$,我们以$\sqrt{p_{j|i}q_{i|j}} = x_{ij}/\sqrt{r_i c_j}$作为$x_{ij}$的标准化：

$$X_s = D_r^{-1/2} X D_c^{-1/2} = \left(x_{ij}/\sqrt{r_i c_j}\right) \qquad (X\text{的标准化})$$

偏差/中心化

两个属性变量独立时，x_{ij} 的期望估计为$E_{ij} = n \times \frac{r_i}{n} \times \frac{c_j}{n} = \frac{r_i c_j}{n}$，$E = \frac{\mathbf{r}\mathbf{c}^{\top}}{n} = \left(E_{ij}\right)$，偏差或误差为$x_{ij} - r_i c_j/n$，所有偏差:

$$X_c = X - E = X - \frac{\mathbf{r}\mathbf{c}^{\top}}{n} = \left(x_{ij} - r_i c_j/n\right) \qquad (X\text{的中心化})$$

误差矩阵 X_c也可称为是X的双向中心化（行和列和都为0）。

标准化偏差/中心标准化

定义标准化误差

$$r_{ij} = \left(x_{ij} - r_i c_j / n\right) / \sqrt{r_i c_j}$$

它们的平方和（乘以n)即为 Pearson卡方。所有标准化误差

$$R = (r_{ij}) = D_r^{-1/2} X_c D_c^{-1/2}$$

称为X的中心标准化矩阵。

注1：也可以定义标准化：$r'_{ij} = \frac{\left(x_{ij} - \frac{r_i c_j}{n}\right)}{\sqrt{\frac{r_i c_j}{n}}} = \sqrt{n} r_{ij}$, 分母上的$\frac{r_i c_j}{n}$代表了方差（这与Poisson有关）

注2: r_{ij}近似为Pearson相关系数

验证：记n个研究对象属性变量为x_k，y_k，$k = 1, \ldots, n$，示性变量

$$u_k = 1_{(x_k = i)}, v_k = 1_{(y_k = j)}, k = 1, \ldots, n$$

的样本相关系数为

$$\left(x_{ij} - r_i c_j / n\right) / \sqrt{(r_i - r_i^2/n)(c_j - c_j^2/n)} \approx r_{ij}$$

注意$\sum_{k=1}^{n} u_k v_k = x_{ij}$，等等。

命题3.（课本*Result* 12.1） $X_s = D_r^{-1/2} X\, D_c^{-1/2} = D_r^{-1/2} E D_c^{-1/2} + R,$
平方误差意义下， $D_r^{-1/2} E D_c^{-1/2}$ 是 $X_s = D_r^{-1/2} X\ D_c^{-1/2}$ 的最佳秩1逼近，R 为两者之差。

证明：首先

$$D_r^{-1/2} E D_c^{-1/2} = D_r^{-1/2} \mathbf{r}\mathbf{c}^{\mathsf{T}} D_c^{-1/2} / n \overset{\Delta}{=} \mathbf{u}\mathbf{v}^{\mathsf{T}} / n,$$

其中 $\mathbf{u} = D_r^{-1/2}\mathbf{r} = \left(\sqrt{r_1}, ..., \sqrt{r_I}\right)^{\mathsf{T}}$， $\mathbf{v} = D_c^{-1/2}\mathbf{c} = \left(\sqrt{c_1}, ..., \sqrt{c_I}\right)^{\mathsf{T}}$，

注意 $\| \mathbf{u} \| = \| \mathbf{v} \| = \sqrt{n}$。

另外由 $D_r \mathbf{1} = \mathbf{r} \Rightarrow \mathbf{u} = D_r^{-1/2}\mathbf{r} = D_r^{1/2}\mathbf{1}$， $\mathbf{v} = D_c^{-1/2}\mathbf{c} = D_c^{1/2}\mathbf{1}$。

下面证明 $\mathbf{u}, \mathbf{v}$ 满足 X_s 的SVD对偶方程，且1是最大奇异值： (待证)

$$X_s \mathbf{v} = D_r^{-1/2} X D_c^{-1/2} D_c^{1/2}\mathbf{1} = D_r^{-1/2} X \mathbf{1} = D_r^{-1/2}\mathbf{r} = D_r^{1/2} D_r^{-1}\mathbf{r} = D_r^{1/2}\mathbf{1} = \mathbf{u}$$

$$X_s^{\mathsf{T}} \mathbf{u} = D_c^{-1/2} X^{\mathsf{T}} D_r^{-1/2} D_r^{1/2}\mathbf{1} = D_c^{-1/2} X^{\mathsf{T}}\mathbf{1} = D_c^{-1/2}\mathbf{c} = D_c^{1/2} D_c^{-1}\mathbf{c} = D_c^{1/2}\mathbf{1} = \mathbf{v}$$

所以 $\mathbf{u}, \mathbf{v}$ 是 X_s 的奇异值分解中对应奇异值1的特征向量。

Pearson将偏差或标准化偏差或者它们的函数都称为contingency (关联性度量)。

Contingency：Pearson卡方

Pearson卡方是一种contingency度量：

$$\chi^2 = n\|R\|^2 = n\sum r_{ij}^2 = \sum \frac{(x_{ij}-E_{ij})^2}{E_{ij}},\ E_{ij}= r_i c_j/n,$$

独立性或者齐一性假设下χ^2近似服从卡方分布$\chi^2_{(I-1)(J-1)}$

给定总数n,所有计数服从多项分布（$m = IJ$项）：

$$X_{I\times J} = (x_{ij}) \sim \text{Multinominal}(n, (p_{ij}))$$

拉直X，以单下标表示

$$\mathbf{y} = (y_1,\ldots, y_m)^{\mathsf{T}} \sim \text{Multinominal}\left(n, \mathbf{p} = (p_1,\ldots, p_m)^{\mathsf{T}}\right)$$

则$E(\mathbf{y}) = n\mathbf{p} \overset{\Delta}{=} \boldsymbol{\mu}$

$$\text{var}(\mathbf{y}) = n\times\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_mp_1 & -p_mp_2 & \cdots & p_m(1-p_m) \end{pmatrix} = n\left[diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathsf{T}}\right] \overset{\Delta}{=} \Sigma,$$

将**y**标准化：$\mathbf{z}=(\Sigma^{-})^{1/2}(\mathbf{y}-\boldsymbol{\mu})$，（近似退化标准正态），其模长平方

$$\|\mathbf{z}\|^2=(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-}(\mathbf{y}-\boldsymbol{\mu})$$ （近似卡方分布）

度量了计数数据与其期望的差距（拟合优度），其中Σ^{-}是Σ的广义逆，满足$\Sigma\Sigma^{-}\Sigma=\Sigma$。容易验证

$$\Sigma^{-}=\mathrm{diag}(1/\mathbf{p})/n=\mathrm{diag}(1/np_1,...,1/np_m)$$

是广义逆，

$$\|\mathbf{z}\|^2=\sum(y_i-np_i)^2/np_i=\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-np_{ij})^2/np_{ij}$$

（二次型的值与广义逆选择无关)

在独立性原假设下，np_{ij}的估计为$n\times r_i/n\times c_j/n=r_ic_j/n$,代入$\|\mathbf{z}\|^2$ 得

Pearson卡方 $$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x_{ij}-r_ic_j/n)^2}{r_ic_j/n}.$$

对应分析

原文：Developed by the French, correspondence analysis is a graphical procedure for representing associations in a table of frequencies or counts.

中译版P557：对应分析是一种将频数或计数表中的各种联系用图来表示的方法，由弗伦奇所提出

French应译作“法国人”而不是“弗伦奇”！

对应分析

对列联表应用SVD和双标图可视化方法，研究行标和列标之间的对应或关联性的方法称为对应分析（correspondence analysis，课本12.7），由法国学者Jean-Paul Benzécri（1973）提出。对应分析是法国统计学派的代表性方法。

有一种观点认为，统计学作为一个学科是从法国人Pierre-Simon Laplace（1749-1827）和比利时-法国人Adolphe Quetelet(凯特勒1796-1896)开始的，后者提出了BMI。