第 1 章　P20

1. 在数据处理时，为什么通常要进行标准化处理？

对数据进行标准化处理主要为了消除变量的量纲以及量纲差别较大时所带来的影响，尤其当变量间的单位不同且量级差别特别大时，使用不做任何处理的数据进行计算，可能会得到极不合理的结果。

2. 欧氏距离与马氏距离的优缺点是什么？

欧氏距离是计算点与点之间距离的常用方法，其缺点是坐标的各维度对计算距离的贡献是同等的，距离的大小与各维度对应的指标变量的单位有关。因此，对于大部分统计问题，欧氏距离不太适合。而马氏距离弥补了欧氏距离在统计问题上的缺陷，马氏距离的计算中会将各指标变量转化为无量纲的数值，而且当变量服从或渐近服从多元正态分布时，马氏距离具有良好的统计性质。

3. 当变量X_1和X_2方向上的变差相等，且X_1与X_2互相独立时，采用欧氏距离与统计距离是否一致？

当变量X_1和X_2方向上的变差相等，且X_1与X_2互相独立时，采用欧氏距离与统计距离的计算结果会相差一个常数倍，即欧氏距离=统计距离*C，该常数项 C 为变量X_1和X_2的标准差。

4. 如果正态随机向量$X=(x_1,x_2,\dots,x_p)'$的协方差阵Σ是对角阵，证明X的分量是相互独立的随机变量。

证明：不妨设$\Sigma=\mathrm{diag}(\sigma_1^2,\dots,\sigma_p^2)$，X的均值向量为$\mu=(\mu_1,\dots,\mu_p)$，则$X\sim N(\mu,\Sigma)$. X的概率密度函数为：

$$
\begin{aligned}
f(x_1,\dots,x_p) &= (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right\} \\
&= (2\pi)^{-\frac{p}{2}}\sigma_1^{-1}\cdots\sigma_p^{-1} \\
&\cdot\exp\left\{-\frac{1}{2}(x_1-\mu_1,\dots,x_p-\mu_p)\begin{pmatrix}\sigma_1^{-2} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_p^{-2}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ \vdots\\ x_p-\mu_p\end{pmatrix}\right\} \\
&= (2\pi)^{-\frac{p}{2}}\sigma_1^{-1}\cdots\sigma_p^{-1}\cdot\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right\} \\
&= \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right\}=\prod_{i=1}^{p}f(x_i)
\end{aligned}
$$

因此，X的分量是相互独立的随机变量。

5. y_1与y_2是相互独立的随机变量，且$y_1\sim N(0,1)$，$y_2\sim N(3,4)$。

(a)求y_1^2的分布。

(b) 如果$\mathrm{y}=\begin{bmatrix}\mathrm{y}_1\\(\mathrm{y}_2\text{-}3)/2\end{bmatrix}$，写出$\mathrm{y}'\mathrm{y}$关于$\mathrm{y}_1$与$\mathrm{y}_2$的表达式，并写出$\mathrm{y}'\mathrm{y}$的分布。

(c) 如果$\mathrm{y}=\begin{bmatrix}\mathrm{y}_1\\\mathrm{y}_2\end{bmatrix}$且$\mathrm{y}\sim\mathrm{N}(\mu,\Sigma)$，写出$\mathrm{y}'\Sigma^{-1}\mathrm{y}$关于$\mathrm{y}_1$与$\mathrm{y}_2$的表达式，并写出$\mathrm{y}'\Sigma^{-1}\mathrm{y}$的分布。

解：（a）$\mathrm{y}_1\sim\mathrm{N}(0,1)$，记$\mathrm{y}_1$的分布函数为$\Phi(\mathrm{y}_1)$， y_1的密度函数为$\phi(\mathrm{y}_1)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{y_1^2}{2})$。

$$P(y_1^2\le y)=P\left(-\sqrt{y}\le y_1\le\sqrt{y}\right)=\Phi\left(\sqrt{y}\right)-\Phi\left(-\sqrt{y}\right)=2\Phi\left(\sqrt{y}\right)-1,\quad y\ge 0$$

则y_1^2的密度函数为：

$$f(y)=2\phi\left(\sqrt{y}\right)\cdot\frac{1}{2\sqrt{y}}=\frac{2}{\sqrt{2\pi}}\exp\left(-\frac{y}{2}\right)\cdot\frac{1}{2\sqrt{y}}=\frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{y}{2}\right),\quad y\ge 0$$

即 $f(y)=\begin{cases}\frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{y}{2}\right), & y\ge 0\\ 0, & y<0\end{cases}$，

因此，y_1^2服从自由度为 1 的卡方分布。

（b）由$\mathrm{y}_2\sim\mathrm{N}(3,4)$可知，$(y_2-3)/2\sim\mathrm{N}(0,\ 1)$，则$\frac{(y_2-3)^2}{4}\sim\chi^2(1)$，

$$\mathrm{y}'\mathrm{y}=\begin{pmatrix}y_1 & (y_2-3)/2\end{pmatrix}\begin{pmatrix}y_1\\(y_2-3)/2\end{pmatrix}=y_1^2+\frac{(y_2-3)^2}{4}$$

由于y_1与y_2是相互独立的，所以$y_1^2+\frac{(y_2-3)^2}{4}\sim\chi^2(2)$，即$\mathrm{y}'\mathrm{y}$服从自由度为 2 的卡方分布。

（c）不妨设$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}$，则$\Sigma^{-1}=\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{pmatrix}$，

$$y'\Sigma^{-1}y=\begin{pmatrix}y_1 & y_2\end{pmatrix}\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}=\frac{\sigma_{22}y_1^2-2\sigma_{12}y_1y_2+\sigma_{11}y_2^2}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}$$

令$z=\Sigma^{-1/2}y$，由多元正态分布的性质（3）可得，$z\sim\mathrm{N}(\Sigma^{-\frac{1}{2}}\mu,\ \mathrm{I})$其中I是单位矩阵。根据习题 4 的结论可知，$z$的各分量相互独立且服从正态分布，因此$\mathrm{z}'\mathrm{z}$服从自由度为 2 的非中心卡方分布，而$\mathrm{z}'\mathrm{z}=\mathrm{y}'\Sigma^{-1}\mathrm{y}$。

第 2 章　P34

1. 试举出两个可以运用多元均值检验的实际问题。

实际问题 1：试从多个方面比较两家上市公司的综合实力；实际问题 2：试基于药物的有效性和可能会引起的不良反应等方面比较新研发的药物相较于已有药物的效果。

2. 试谈 Wilks 统计量在多元方差分析中的重要意义。

Wilks 统计量是两个广义方差之比，在多元方差分析中主要用于检验不同水平间多个指标是否存在显著差异。类似于一元方差分析中的方差分解方式，多元方差分析中也将总方差（变异）分为组间方差（变异）和组内方差（变异），只是对于多元的情况，组间变异和组内变异均为矩阵，Wilks 统计量等于组内变异的广义方差与总变异的广义方差之比。因此，正如 F 分布对于一元方差分析的重要性一样，Wilks 统计量对于多元方差分析也是十分重要的。

3. 现选取内蒙古、广西、贵州、云南、西藏、宁夏、新疆、甘肃和青海等 9 个内陆边远省区。选取人均 GDP、第三产业比重、人均消费支出、人口自然增长率及文盲半文盲人口占 15 岁以上人口的比例等 5 项能够较好地说明各地区社会经济发展水平的指标，验证边远及少数民族聚居区的社会经济发展水平与全国平均水平有无显著差异。

边远及少数民族聚居区社会经济发展水平的指标数据

地区	人均 GDP（元）	第三产业比重（%）	人均消费支出（元）	人口自然增长率（%）	文盲半文盲人口占比(%)
内蒙古	5 068	31.1	2 141	8.23	15.83
广西	4 076	34.2	2 040	9.01	13.32
贵州	2 342	29.8	1 551	14.26	28.98
云南	4 355	31.1	2 059	12.10	25.48
西藏	3 716	43.5	1 551	15.90	57.97
宁夏	4 270	37.3	1 947	13.08	25.56
新疆	6 229	35.4	2 745	12.81	11.44
甘肃	3 456	32.8	1 612	10.04	28.65
青海	4 367	40.9	2 047	14.48	42.92

资料来源:中华人民共和国国家统计局.中国统计年鉴:1998.北京，中国统计出版社，1998.

5 项指标的全国平均水平为:

$$\boldsymbol{\mu_0} = (6212.01, 32.87, 2972, 9.5, 15.78)'$$

解：由于协方差未知，根据 2.1.2 节中内容可知，需要计算统计量$T^2 = n(n-1)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{L}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$，其中$\bar{\boldsymbol{X}}$是样本均值，$L$是样本离差阵。

$\bar{\boldsymbol{X}} = (4208.77778, 35.12222, 1965.88889, 12.21222, 27.79444)'$，

$\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0 = (-2003.232222, 2.252222, -1006.111111, 2.712222, 12.014444)'$，

$$\boldsymbol{L} = \begin{bmatrix} 9181717.556 & 5242.14444 & 2985241.7778 & -5398.05556 & -57206.9211 \\ 5242.144 & 175.31556 & -949.1778 & 59.62156 & 393.2721 \\ 2985241.778 & -949.17778 & 1128278.8889 & -1922.70778 & -28170.3456 \\ -5398.056 & 59.62156 & -1922.7078 & 54.88976 & 228.6913 \\ -57206.921 & 393.27211 & -28170.3456 & 228.69131 & 1771.9588 \end{bmatrix}$$

$$\boldsymbol{L^{-1}} = \begin{bmatrix} 0.02089345 & -0.4891093 & -0.06775808 & 3.110548 & -0.6955698 \\ -0.48910928 & 230.04163 & -0.23460510 & -26.269187 & -67.1859664 \\ -0.06775808 & -0.2346051 & 0.25261392 & -11.670613 & 3.3867726 \\ 3.11054818 & -26.2691867 & -11.67061277 & 947.787363 & -201.6074538 \\ -0.69556979 & -67.1859664 & 3.38677256 & -201.607454 & 77.9608336 \end{bmatrix} \times 10^{-4}$$

由此可计算得：

$$\frac{n-p}{(n-1)p}T^2 = \frac{n(n-p)}{p}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{L}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0) = \frac{9\times 4}{5} \times 5.569546 = 40.10073$$

由于$\frac{n-p}{(n-1)p}T^2$服从分布$F(5,4)$，从网上查得$F_{5,4}(0.01) = 15.52$，$F_{5,4}(0.05) = 6.26$，

由$40.10073 > F_{5,4}(0.01)$可知，在 0.01 的显著性水平上边远及少数民族聚居区的社会经济发展水平与全国平均水平有显著差异。

4. 试针对某一实际问题具体运用多元方差分析方法。（答案略）

第 3 章　P82

1.聚类分析的基本思想和功能是什么？

聚类分析首先假定所研究的样品或指标（变量）之间存在不同程度的相似性（亲疏关系），然后对于给定的一批有多个观测指标的样品，可以根据一些能够度量样品或指标之间相似程度的统计量作为划分类型的依据，最终把相似程度接近的样品（指标）聚合为同一类。聚类分析的目的就是把研究对象根据相似程度进行归类，使同类中对象的相似最大化，而类与类之间的差异性最大化。

2.试述系统聚类法的原理和具体步骤。

系统聚类的原理是根据样品（或指标变量）间的距离（或相似性）进行类的合并，首先将各样品或（变量）当作一类，然后每次将距离最近（或相似度最高）的两类（或变量）聚合成一类，如此重复进行下去，直至每个样品（或变量）最终被聚成一个大类。

系统聚类的具体步骤如下：

（1）将每个样品（或变量）独自作为一类，如此构造n个类；

（2）计算n个类两两之间的距离$\{d_{ij}\}$；

（3）合并距离最近的两类为一新类，并重新计算新类与当前各类之间的距离；

（4）重复步骤（3），直至最后将所有的样品（或变量）全被聚成一个类。

3.试述 K-均值聚类的方法原理。

K-均值聚类方法的思想是把每个样品聚集到其最近质心（均值）的类中，它是一种迭代求解的聚类分析算法。其步骤是：首先从数据集中随机选取K个点作为初始聚类中心，然后计算各个样本到聚类中心的距离，并把样本归到离它最近的那个聚类中心所在的类，最后计算新形成的每一个类所包含对象的平均值作为新的聚类中心。重复前面的操作，直至相邻两次的聚类中心没有任何变化，说明样本调整结束。

4.试述模糊聚类的思想方法。

模糊聚类分析是根据研究对象的亲疏程度或相似性，通过建立模糊相似关系对研究对象进行聚类分析的方法。在模糊聚类中，每个样本不再仅属于某一类，而是以一定的隶属度属于每一类，意味着通过模糊聚类分析，可得到样本属于各个类别的不确定性程度，即建立起了样本对于类别的不确定性的描述，这样就更能准确地反映实际情况。

5.试运用 SPSS 软件进行一个实际问题的分类研究。

（答案略）

第 4 章　P104-P105

1.应用判别分析应该具备什么样的条件？

判别分析最基本的要求是：分组类型在两组以上；每组案例的规模必须至少在一个以上；解释变量必须是可测量的，才能够计算其平均值和方差，使其能合理地应用于统计函数。另外，判别分析的假设条件有：判别变量间不存在多重共线性；各判别变量服从多元正态分布，且各组的协方差矩阵相等。

2.试述贝叶斯判别方法的思路。

贝叶斯统计的思想是假定对研究对象已有一定的认识，而且常用先验概率分布来描述这种认识，然后对于取得的一个样本，可以用样本来修正已有的认识（先验概率分布）从而得到后验概率分布，各种统计推断都可以通过后验概率分布来进行。将贝叶斯统计思想用于判别分析，就是贝叶斯判别，具体为：假设 k 个总体分别具有 p 维的密度函数，并且 k 个总体的先验分布是已知的，k 个总体对应 R^p 上的一个划分。通过建立判别规则和相应的损失函数，可以求得使平均损失（后验风险）最小的一个划分。

3.试述费歇判别方法的思想。

费歇判别的核心思想是投影，即将k组p维数据投影到某一个方向，使得组与组之间的投影尽可能地分开，其中费歇尔判别借用了一元方差分析的思想来衡量组与组之间的分开程度，进而求解使分开程度最大化的投影向量。

4.什么是逐步判别分析？

凡具有筛选变量能力的判别方法统称为逐步判别法。逐步判别法的基本思想是：逐步引入变量，每次引入一个使检验统计量取得最优值的变量，同时也检验先前引入的变量，如果先前引入的变量其判别能力随新变量的引入而变得不显著，则需及时将其剔除，直到判别式中的变量都很显著，且剩下来的变量也再没有其他重要的变量可引入时，逐步筛选结束。

5.简要叙述判别分析的步骤及流程。

判别分析的逻辑步骤如下：

（1）明确研究问题：这一步骤主要根据待研究的问题来确定具体的研究内容。（2）研究设计要点的确定：主要包括解释变量的选择、估计判别函数所需样本量的确定和用于后续的验证中的测试样本的保留。（3）假定条件的验证：检验解释变量的多元正态性、协方差是否相等以及解释变量间是否存在多重共线性等。（4）估计判别函数：确定具体的判别分析方法，估计判别函数。（5）结果的解释：说明判别函数中每个解释变量的相对重要性，其中可以通过标准化判别权重、判别载荷、偏 F 值等方法来确定其重要性。（6）判别结果的验证：通常采用分割样本或者交叉验证法。

判别分析的流程：

明确研究问题→设计要点的确定→假定条件的验证→估计判别函数→使用分类矩阵评估预测的精度→判别函数的解释→判别结果的验证

6.为研究某地区人口死亡状况，已按某种方法将 15 个已知样品分为 3 类，指标及原始数据如下表所示，试建立判别函数并判定另外 4 个待判样品属于哪类。

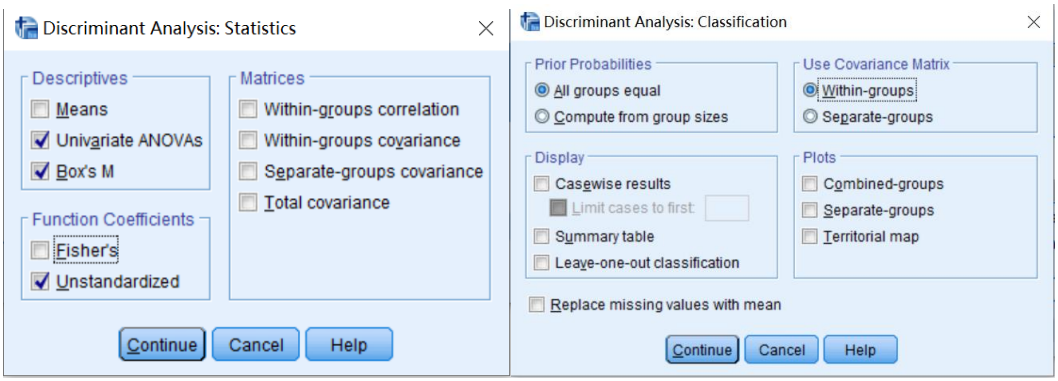
x_1:0 岁组死亡概率　　x_4:55 岁组死亡概率
x_2:1 岁组死亡概率　　x_5:80 岁组死亡概率
x_3:10 岁组死亡概率　　x_6:平均预期寿命

组别	序号	x_1	x_2	x_3	x_4	x_5	x_6
第一组	1	34.16	7.44	1.12	7.87	95.19	69.30
	2	33.06	6.34	1.08	6.77	94.08	69.70
	3	36.26	9.24	1.04	8.97	97.30	68.80
	4	40.17	13.45	1.43	13.88	101.20	66.20
	5	50.06	23.03	2.83	23.74	112.52	63.30
第二组	1	33.24	6.24	1.18	22.90	160.01	65.40
	2	32.22	4.22	1.06	20.70	124.70	68.70
	3	41.15	10.08	2.32	32.84	172.06	65.85
	4	53.04	25.74	4.06	34.87	152.03	63.50
	5	38.03	11.20	6.07	27.84	146.32	66.80
第三组	1	34.03	5.41	0.07	5.20	90.10	69.50
	2	32.11	3.02	0.09	3.14	85.15	70.80
	3	44.12	15.12	1.08	15.15	103.12	64.80
	4	54.17	25.03	2.11	25.15	110.14	63.70
	5	28.07	2.01	0.07	3.02	81.22	68.30
待判样品	1	50.22	6.66	1.08	22.54	170.60	65.20
	2	34.64	7.33	1.11	7.78	95.16	69.30
	3	33.42	6.22	1.12	22.95	160.31	68.30
	4	44.02	15.36	1.07	16.45	105.30	64.20

解：我们选择使用**费歇尔判别**方法来建立判别函数。

（1） 将上面表格中 6 个指标变量对应的数据复制粘贴到打开的 SPSS 数据框中，并定义一个新的变量 group，分别用 1、2、3 表示第一、二、三组，而待判样本对应的分组保持空着。

（2） 按本书 90 页例 4-1 的操作步骤所示，打开判别分析的对话框并进行相应设置。然后，点击右侧 Statistics 按钮，在新打开的对话框中，勾选如下边左图所示的选项；点击 Classify 按钮，在新打开的对话中，勾选如下边右图所示的选项。



（3） 点击 OK 运行后，其中部分输出结果如下所示：

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
x1	.997	.019	2	12	.981
x2	.990	.061	2	12	.941
x3	.645	3.301	2	12	.072
x4	.438	7.690	2	12	.007
x5	.174	28.557	2	12	.000
x6	.926	.478	2	12	.631

Box's Test of Equality of Covariance Matrices

Log Determinants

group	Rank	Log Determinant
1	.[a]	.[b]
2	.[a]	.[b]
3	.[a]	.[b]
Pooled within-groups	6	8.555

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

a. Rank < 5

b. Too few cases to be non-singular

从上面结果中可以看出，在 0.05 的显著性水平上，变量 x1、x2、x3、x6 对应的 p 值均大于 0.05，说明这四个均不能拒绝三个分组上均值相等的原假设。另外，从协方差阵的齐性检验结果也可看出，协方差阵是奇异矩阵，主要由于$p > n$，因此，考虑仅使用变量 x4 和 x5 建立判别函数。

（4） 将变量 x1、x2、x3、x6 从 Independents 框中移出，重新运行，其中得到有如下结果。

Test Results

Box's M		28.091
F	Approx.	3.547
	df1	6
	df2	3588.923
	Sig.	.002

Tests null hypothesis of equal population covariance matrices.

由上表可知，结果拒绝各组的协方差阵相等的原假设，认为各组的协方差阵不相等，因此在（2）中使用的协方差矩阵应该选择 Separate-groups。

（5） 将使用的协方差阵进行调整后，打开 Save 对话框，并选中第一个和第三个复选框，重新运行，可得到如下结果。另外，在数据框中也会出现 4 列新的变量，分别是对各样品的分组判别结果，以及分别被判为一、二、三组的概率。

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.156	21.358	4	.000
2	1.000	.000	1	.990

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
x4	-.531	1.441
x5	1.341	-.748

Canonical Discriminant Function Coefficients

	Function	
	1	2
x4	-.069	.186
x5	.102	-.057
(Constant)	-10.610	3.429

Unstandardized coefficients

Functions at Group Centroids

	Function	
group	1	2
1	-1.217	.004
2	2.927	.000
3	-1.710	-.004

Unstandardized canonical discriminant functions evaluated at group means

由以上第一张表可知判别函数 1 是显著的，而判别函数 2 是不显著的。其中，标准化的判别函数 1 为：$\mathrm{y}_1 = -0.531\mathrm{x}_4^* + 1.341\mathrm{x}_5^*$；非标准化的判别函数 1 为：$\mathrm{y}_1 = -0.069x_4 + 0.102x_5 - 10.610$。

（6） 根据数据框中输出的结果可知，待判样本中 1 和 3 被判为第二组、2 和 4 被判为第一组，而其他的样本中仅有第三组的第 3 个样本被判错。

第 5 章　P133

1.主成分的基本思想是什么？

主成分分析是研究如何通过原始变量的少数几个线性组合来解释原始变量绝大多数信息的一种多元统计方法。该方法主要基于众多变量之间有一定的相关性，则必然存在着起支配作用的共同因素这一想法，通过对原始变量相关矩阵或协方差矩阵内部结构关系进行研究，利用原始变量的线性组合形成几个综合指标（主成分）。利用主成分分析得到的主成分与原始变量之间有如下基本关系：

（1）每一个主成分都是各原始变量的线性组合；

（2）保留了原始变量绝大多数信息的主成分的数目远少于原始变量的数目；

（3）各主成分之间互不相关。

主成分分析的基本思想是在保留原始变量尽可能多的信息的前提下达到降维的目的，从而简化问题的复杂性并抓住问题的主要矛盾。

2.主成分在应用中的主要作用是什么？

主成分是原始变量的重新组合，少数的几个主成分就能包含原始变量的大部分信息，而且主成分之间互不相关，因此主成分主要可以解决实际应用中由指标变量的个数较多而且信息大量重叠所带来的复杂性增加、模型的建立和分析难度大等问题。信息的重叠有时甚至会抹杀事物的真正特征与内在规律，使用主成分分析，可以从事物之间错综复杂的关系中找出一些主要成分，从而能有效利用大量统计数据进行定量分析，揭示变量之间的内在关系，得到对事物特征及其发展规律的一些深层次的启发，把研究工作引向深入。

3.由协方差阵出发和由相关阵出发求主成分有什么不同？

由于对原始数据的各变量进行减均值除以标准差的标准化后，再对标准化后的数据求协方差阵即为原始数据的相关阵。因此，由协方差阵出发和由相关阵出发求主成分时的区别可以转化为使用标准化前后的数据求解主成分所带来的不同。由于对数据进行标准化的过程实际上就是抹杀原始变量离散程度差异的过程，标准化后的各变量方差相等且均为 1，即意味着原始各变量自身变异这一部分重要信息被抹杀，使得标准化后各变量在对主成分构成中的作用趋于相等，因此使用标准化前后的数据求解主成分会有较大差异。

一般而言，对于度量单位不同的指标变量或是取值范围彼此差异非常大的指标变量，我们不直接由其协方差矩阵出发进行主成分分析，而应该考虑将数据标准化。但是对于取值范围相差不大或是度量相同的指标，由于进行标准化处理后的数据会损失各变量自身方差这一重要信息，因此，对同度量或是取值范围在同量级的数据还是直接从协方差矩阵求解主成分为宜。

4.读者自己找一个实际问题的数据，应用 SPSS 软件试做主成分分析。（答案略）

第 6 章　P160

1.因子分析与主成分分析有什么本质不同？

因子分析可以看作是主成分分析的推广。它也是利用降维的思想，从研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。相比主成分分析，因子分析更倾向于描述原始变量之间的相关关系，因此因子分析的出发点是原始变量的相关矩阵。

二者的本质不同主要体现在以下几个方面：

(1) 因子分析把诸多变量看成是对每一个变量都有作用的一些公共因子和一些仅对某一个变量有作用的特殊因子的线性组合。因此，其目的就是要从数据中探查能对变量起解释作用的公共因子和特殊因子，以及公共因子和特殊因子的组合系数。主成分分析则简单一些，它只是从空间生成的角度寻找能解释诸多变量绝大部分变异的几组彼此不相关的新变量(主成分)，它是一种可逆的数据变换。

(2) 因子分析是把变量表示成各因子的线性组合，而主成分分析是把主成分表示成各变量的线性组合。

(3) 主成分分析中不需要有一些特定假设，而因子分析则需要一些假设。因子分析的假设包括:各个公共因子之间不相关，特殊因子之间不相关，公共因子和特殊因子之间不相关。

(4) 提取主因子的方法不仅可以使用主成分法，还有极大似然法等，不同的方法得到的结果一般也不同，而主成分只能用主成分法提取。

(5) 主成分分析中，当给定的协方差矩阵或者相关矩阵的特征根唯一时，主成分一般是固定的；而因子分析中因子是不固定的，可以使用旋转技术得到便于解释的不同因子。

2.因子载荷a_{ij}的统计定义是什么？它在实际问题分析中的作用是什么？

因子载荷a_{ij}的统计意义就是第i个变量与第j个公共因子的相关系数，即表示变量X_i依赖公共因子F_j的比重，另外其平方也是公共因子F_j解释X_i方差的比例。

在实际问题分析中，可以通过因子载荷的大小来分析影响各变量的主要公共因子，使得对变量的分析更加深刻。同时，当某几个变量在同一公共因子上的载荷均较大时，也表明这几个变量的相关性会较强，因此也可以通过此种方法更方便地分析变量间的相关关系。

3.试用 SPSS 软件对一个实际问题的研究应用因子分析。（答案略）

第 7 章　P186

1.试述对应分析的思想方法及特点。

对应分析是 R 型因子分析与 Q 型因子分析的结合，它也是利用降维的思想来达到简化数据结构的目的，与因子分析不同的是，它同时对数据表中的行与列进行处理，并寻求用低维图形来表示数据表中行与列之间的关系。其基本思想为：R 型因子分析与 Q 型因子分析是从不同角度出发对同一个整体进行研究，它们之间存在着一定的内在联系。对应分析通过一个矩阵 Z（Z 的行向量对应各行剖面，Z 的列向量对应各列剖面）将二者有机的结合起来。具体来说，对列剖面进行分析时的协方差阵为$A = Z'Z$，对行剖面进行分析时的协方差阵为$B = ZZ'$，而 A 和 B 有相同的非 0 特征根，且二者的相同特征根所对应的特征向量间也存在简单的线性关系。如果 A 的特征值λ所对应的特征向量为u，则 B 的特征值λ所对应的特征向量为Zu。因此，由 R 型因子分析的结果可以很方便地得到 Q 型因子分析的结果，从而大大减少了计算量，同时也可以用相同的因子轴表示行和列属性的不同状态，把其同时反映在具有相同坐标轴的因子平面上。

对应分析的一大特点就是可以在一张二维图上同时表示出行和列两类属性的各种状态，以直观描述原始数据结构。另外，它还可以直接从图上对样品进行直观的分类，而且能够指示分类的主要参数（主因子）以及分类的依据，是一种直观、简单、方便的多元统计方法。

2.试述对应分析中总惯量的意义。

总惯量不仅反映了行剖面集定义的各点与其重心加权距离的总和,同时与χ^2统计量仅相差一个常数,而χ^2统计量反映了列联表横联与纵联的相关关系,因此总惯量也反映了两个属性变量各状态之间的相关关系。对应分析就是在对总惯量信息损失最小的前提下，简化数据结构以反映两属性变量之间的相关关系。

3.试对一个实际问题运用 SPSS 软件进行对应分析。（答案略）

第 8 章　P209

1.试述典型相关分析的统计思想及该方法在研究实际问题中的作用。

典型相关分析研究两组变量间整体的线性相关关系，它是将每一组变量作为一个整体来进行研究，而不是分析每一组变量内部的各个变量。所研究的两组变量可以是一组变量为自变量，而另一组变量为因变量，也可以是两组变量处于同等地位，但典型相关分析要求两组变量都至少是间隔尺度的。它主要是借助于主成分分析的思想，对每一组变量分别寻找线性组合，使生成的新的综合变量能代表原始变量大部分的信息，同时与由另一组变量生成的新的综合变量的相关程度最大，这样一组新的综合变量称为第一对典型相关变量。同样的方法可以找到第二对，第三对……并且使各对典型相关变量之间互不相关，典型相关变量之间的简单相关系数称为典型相关系数。典型相关分析就是用典型相关关系数衡量两组变量之间的相关性。

在研究实际问题时，可以通过典型相关分析找出几对主要的典型相关变量，根据典型相关变量相关程度及各典型相关变量线性组合中原变量系数的大小，结合对所研究实际问题的定性分析，尽可能给出较为深刻的分析结果。

2.典型相关分析中的冗余度有什么作用？

为了克服在使用典型根(典型相关系数的平方)作为共同方差的测量中可能出现的有偏性和不稳定性，提出了冗余指数（冗余度），它可以辅助典型相关系数来分析典型变量。

冗余指数等价于在整个自变量组与因变量组的每一个因变量之间计算多元相关系数的平方，然后将这些平方系数平均得到一个平均的 R^2。这样，冗余测量就像多元回归的 R^2 统计量，作为一个指数的值也是类似的。但是，典型相关不同于多元回归，它不是处理单个因变量，而是处理因变量的组合，而且这个组合只有每个因变量的全部方差的一部分。因此，一个典型变量的冗余指数就是这个典型变量中所包含变量的共同方差比例乘以典型相关系数的平方，从而得到每个典型函数可以解释的共同方差部分。

3.典型变量的解释有什么具体方法？实际意义是什么？

一个典型变量对应一个典型函数，研究典型函数中原始变量的相对重要性主要使用以下三种方法：(1)典型权重(标准化系数)；(2)典型载荷(结构系数)；(3)典型交叉载荷。

(1)典型权重。传统的解释典型函数的方法包括观察每个原始变量在它的典型变量中的典型权重的符号和大小。有较大的典型权重，则说明原始变量对它的典型变量贡献较大，反之则相反。原始变量的典型权重有相反的符号，说明变量之间存在一种反向关系，反之则有正向关系。但是，这种解释遭到了很多批评，因此在解释典型相关的时候应慎用典型权重。

(2)典型载荷。由于典型权重的缺陷，典型载荷逐步成为解释典型相关分析结果的基础。典型载荷，也称典型结构相关系数，是原始变量(自变量或者因变量)与它的典型变量间的简单线性相关系数。典型载荷反映原始变量与典型变量的共同方差，它的解释类似于因子载荷，也就是每个原始变量对典型函数的相对贡献。

(3)典型交叉载荷。它的提出是作为典型载荷的替代，计算典型交叉载荷包括使每个原始因变量与自变量的典型变量直接相关。交叉载荷提供了一个更直接地测量因变量组与自变量组相关关系的指标。

4.运用 SPSS 或 SAS 软件（此处应该是 R 软件）试对一个实际问题的研究应用典型相关分析。（答案略）

第 9 章　P225

1.简述对数线性模型应用的原理。

对数线性模型是进一步用于离散型数据或整理成列联表格式的数据的统计分析工具。它可以把方差分析和线性模型的一些方法应用到对交叉列联表的分析中，从而对定性变量间的关系做进一步的描述和分析。列联表分析无法系统地评价变量间的联系，也无法估计变量间交互作用的大小，而对数线性模型是处理这些问题的最佳方法。

在对数线性模型分析中，要先将列联表中的概率取对数再进行分解处理，可用公式表示如下：

$$\eta_{ij} = \ln p_{ij} = \ln\left(p_{i.}p_{.j}\frac{p_{ij}}{p_{i.}p_{.j}}\right) = \ln p_{i.} + \ln p_{.j} + \ln\frac{p_{ij}}{p_{i.}p_{.j}}, \quad i,j = 1,2$$

若把上式中的$\ln p_{i.}$，$\ln p_{.j}$，$\ln\frac{p_{ij}}{p_{i.}p_{.j}}$分别记为$\mathrm{A}_i$, B_j和$(\mathrm{AB})_{ij}$，则上式可写成

$$\eta_{ij} = \mathrm{A}_i + \mathrm{B}_j + (\mathrm{AB})_{ij}$$

该式的结构与有交互效应且各水平均为 2 的双因素方差分析模型的结构相似，而模仿方差分析可将其转化为与有交互效应的双因素方差分析数学模型等价的关系式。因此，可根据方差分析的模型估计方法，对模型的参数进行估计。

2.试建立一个实际问题的对数线性模型。（答案略）

3.Logistic 回归模型在处理问卷调查数据中有何应用？

问卷调查中的大部分数据为定性数据，尤其当调查问卷对应的研究目标是一定性的问题时，logistic 回归模型可以更好地综合各方面的指标信息对研究对象进行整体分析。例如，某一问卷是关于员工对公司的满意度的调查，其主要目的是为了公司能进一步在某些方面做出具体改善，以提高员工的工作积极性等。因此，如果能通过模型对调查问卷所搜集的数据进行深入透彻的分析，并指导公司做出有效的决策是非常有意义的。该研究的因变量是员工对公司的总体满意程度，那么因变量就是定性的变量，它可能包含多个分类，而自变量是影响员工对企业满意程度的不同因素。因此，可以使用 logistic 回归模型建立满意程度这一因变量关于不同影响因素（自变量）的回归模型，便于综合分析影响员工满意度的关键因素等。

4.试用 SPSS 软件建立一个实际问题的 Logistic 回归模型。（答案略）

第 10 章　P237

1.试述多变量图示法的思想方法和实际意义。

图形是对资料进行探索性研究的重要工具，人们在运用其他统计方法对所得资料进行分析之前，往往习惯于把资料所包含的信息在一张图形上展示出来，以直观地反映资料的分布情况及各变量之间的相关关系。当变量较少时（一般少于 3 个），可以采用直方图、条形图、饼图、散点图或是经验分布的密度图等方法。而当变量个数为 3 时，虽然仍可以作三维的散点图，但这样做已经不太便于分析。尤其当变量个数大于 3 时，就不能用通常的方法做图了，此时只能用多维变量的图表示方法，如散点图矩阵、脸谱图、雷达图、星图、星座图等多变量的图表示法。

散点图矩阵是借助两变量散点图的做图方法，它可以看作一个大的图形方阵，其中每一个非主对角元素的位置上是对应行的变量与对应列的变量的散点图，而主对角元素位置上是各变量名，这样可以清晰地看到所研究的多个变量两两之间的相关关系。

脸谱图是将观测的 p 个指标变量分别用脸部的某一部位的形状或大小来表示的。一个样品(观测)可以画成一张脸谱，其基本思想是由 15~18 个指标决定脸部特征，若实际资料变量更多将被忽略(有新的画图方法取消了脸的对称性并引入更多脸部特征，从而最多可以用 36 个变量来画脸谱)，若实际资料变量较少则脸部有些特征将被自动固定。脸谱容易给人们留下较深刻的印象，通过对脸谱的分析，就可以直观地对样品或观测进行归类或比较研究。

雷达图是目前应用较为广泛的多元变量做图方法，利用雷达图可以很方便地研究各样本点之间的关系并对样品进行归类。对任一样本点，可以分别在一个圆的 p 个半径轴上确定其坐标，在各坐标轴上点出其坐标后依次连接 p 个点，就可以得到一个 p 边形。这样每一个样本点可以用一个 p 边形表示出来，通过观察各个 p 边形的形状，就可以对各个样本点的相似性进行分析。当样本数目较小时，可以在一个图中画出所有的样本点，便于对各指标进行对比；当样本数目较大时，也可以每一个样本点画一个 p 边形进行分析。星图的形状与雷达图比较近似，尤其当样本数较多时，每一个样本点就可以画出一个多边形的星图或雷达图，此时二者无显著差异。

星座图是通过数据变换将所有样本点都画在一个半圆里面，就像天文学中表示星座的图像，根据样本点的位置可以直观地对各样本点之间的相关性进行分析。利用星座图可以方便地对样本点进行分类，在星座图上比较靠近的样本点比较相似，可以分为一类，相距较远的样本点的差异较大。

2.试对某一多变量实际问题分别画散点图矩阵、脸谱图、雷达图、星座图等。（答案略）

第 11 章　P260

1.简述多维标度法的基本思想。

多维标度法是以研究对象之间某种亲近关系(如距离、相似系数、亲疏程度的分类情况等)为依据，合理地将研究对象(样品或变量)在低维空间中给出标度或位置，以便于全面而又直观地再现原始各研究对象之间的关系，同时在此基础上也可根据各对象之间距离的远近实现对样品的分类。其基本思想是基于距离矩阵或相似（异）矩阵去寻找对应该矩阵的样本点（称作拟合构造点），使得两两样本点之间的距离等于或近似等于该距离矩阵中的各元素。通常这些构造样本点的维度会比较低，使得可以在图形上直观的画出各样本点，如此便于对样本间的相似性等进行分析。

2.简述实现多维标度法的步骤。

多维标度法的实现主要有以下步骤：（1）确定研究的目的；（2）选择需要进行比较分析的样品和原始变量（或者距离矩阵）；（3）选择适当的求解方法，分析样品间的距离矩阵；（4）选择适当的维数，得到距离阵的古典解，将各个样品直观地表现出来并对结果进行解释；（5）检验模型的拟合情况。

3.给定距离阵

$$D=\begin{bmatrix}0\\1&0\\2&1&0\\2&2&1&0\\2&2&2&1&0\\1&2&2&2&1&0\\1&1&1&1&1&1&0\end{bmatrix}$$

求它的拟合构造点，并说明它是不是欧氏型的。

解：由$a_{ij}=-\frac{1}{2}d_{ij}^2$可得矩阵

$$A=\begin{bmatrix}0\\-0.5&0\\-2&-0.5&0\\-2&-2&-0.5&0\\-2&-2&-2&-0.5&0\\-0.5&-2&-2&-2&-0.5&0\\-0.5&-0.5&-0.5&-0.5&-0.5&-0.5&0\end{bmatrix}$$

$$B=HAH=\left(I_n-\frac{1}{n}1_n1_n'\right)A\left(I_n-\frac{1}{n}1_n1_n'\right)$$

$$=\frac{1}{49}\begin{bmatrix}57&32.5&-41&-41&-41&32.5&1\\32.5&57&32.5&-41&-41&-41&1\\-41&32.5&57&32.5&-41&-41&1\\-41&-41&32.5&57&32.5&-41&1\\-41&-41&-41&32.5&57&32.5&1\\32.5&-41&-41&-41&32.5&57&1\\1&1&1&1&1&1&-6\end{bmatrix}$$

求解 B 的特征根和特征向量，得到

$$\lambda_1 = \lambda_2 = 3.5, \qquad \lambda_3 = \lambda_4 = 0.5, \qquad \lambda_5 = 0, \qquad \lambda_6 = -\frac{1}{7}, \qquad \lambda_7 = -1$$

各特征值对应的特征向量按列进行组合构成的矩阵为：

$$\mathrm{U} = (u_1, u_2, \dots, u_7)$$

$$= \begin{bmatrix} -0.275 & -0.508 & 0.084 & 0.571 & -0.378 & 0.154 & -0.408 \\ 0.302 & -0.492 & -0.537 & -0.213 & -0.378 & 0.154 & 0.408 \\ 0.577 & 0.016 & 0.453 & -0.358 & -0.378 & 0.154 & -0.408 \\ 0.275 & 0.508 & 0.084 & 0.571 & -0.378 & 0.154 & 0.408 \\ -0.302 & 0.492 & -0.537 & -0.213 & -0.378 & 0.154 & -0.408 \\ -0.577 & -0.016 & 0.453 & -0.358 & -0.378 & 0.154 & 0.408 \\ 0.000 & 0.000 & 0.000 & 0.000 & -0.378 & -0.926 & 0.000 \end{bmatrix}$$

由于 B 有负的特征根，所以 D 一定不是欧氏型的。

另外，根据

$$a_{1,4} = \frac{\sum_{i=1}^{4} \lambda_i}{\sum_{i=1}^{7} |\lambda_i|} = 0.875$$

$$a_{2,4} = \frac{\sum_{i=1}^{4} \lambda_i^2}{\sum_{i=1}^{7} \lambda_i^2} = 0.961$$

可取$k = 4$，即选取前 4 个特征值对应的特征向量以得到古典解。由于 U 中各列是标准的正交特征向量，所以$\hat{x}_{(i)} = \sqrt{\lambda_i} u_i,\ i = 1,2,3,4$，如此可得

$$\hat{X} = \begin{bmatrix} -0.514 & -0.95 & 0.059 & 0.404 \\ 0.566 & -0.92 & -0.379 & -0.151 \\ 1.080 & 0.03 & 0.320 & -0.253 \\ 0.514 & 0.95 & 0.059 & 0.404 \\ -0.566 & 0.92 & -0.379 & -0.151 \\ -1.080 & -0.03 & 0.320 & -0.253 \\ 0.000 & 0.00 & 0.000 & 0.000 \end{bmatrix}$$

其中，$\hat{X}$的行向量即为矩阵 D 的拟构造点。

但是，一般为了能够直观的在平面图上展示各事物（研究对象），便于分析各事物之间的关系，可以选择$k = 2$，此时对应的各构造点分别为：$\hat{x}_{(1)} = (-0.514, -0.95)$，$\hat{x}_{(2)} = (0.566, -0.92)$，$\hat{x}_{(3)} = (1.080, 0.03)$，$\hat{x}_{(4)} = (0.514, 0.95)$，$\hat{x}_{(5)} = (-0.566, 0.92)$，$\hat{x}_{(6)} = (-1.080, -0.03)$，$\hat{x}_{(7)} = (0,0)$。

4. 试解释样本间相似性的含义。

样本间的相似性主要用来衡量样本间的接近程度。一个样本常常对应 p 个指标变量，故每个样本可看作是 p 维空间中的一个点，n 个样本就组成 p 维空间中的 n 个点，此时自然想到使用距离来度量样本点间的接近程度，其中常用的距离有欧氏距离、马氏距离等。