MULTIVARIATE STATISTICS

零、知识点

- [ ] (P49,e.g. 3.2.8)已知x2时，x1的条件分布为：

$$\mu_{1.2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

- [ ] 总体均值和协方差阵极大似然估计、样本离差矩阵(A)、样本协方差矩阵(S)
- [ ] 估计量的性质（无偏性、有效性、一致性、充分性）
- [ ] 复相关系数、偏相关系数
- [ ] 三大分布（Wishart分布、Hotelling T-square分布、Wilks Lambda分布）
- [ ] T-Square分布与F分布之间的关系

$$T_\alpha^2(p,n) = \frac{np}{n-p+1}F_\alpha(p, n-p+1)$$

- [ ] 单总体均值的检验（协方差已知、未知）
- [ ] 两总体均值的检验（协方差已知、未知）
- [ ] 置信区域
- [ ] 置信区间（联合置信区间、Bonferroni联合置信区间、）
- [ ] 单总体轮廓分析（对比矩阵）

$$H_0 : C\mu = 0; H_1 : C\mu \neq 0;$$

$$C=\begin{pmatrix}1 & -1 & 0 & ... & 0\\ 1 & 0 & -1 & ... & 0\\ ... & ... & ... & ... & ...\\ 1 & 0 & 0 & ... & -1\end{pmatrix}$$

检验统计量为：

$$T^2=n\bar{x}'C'(CSC')^{-1}C\bar{x}\to T^2(p-1,n-1)$$

- [ ] 两总体的轮廓分析（平行吗？重合吗？水平嘛？C要会写）
- [ ] 多元方差分析（考吗？）

$$E=SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i})(x_{ij}-\bar{x_i})'$$

$$H=SSTR=\sum_{i=1}^{k}n_i(\bar{x_i}-\bar{x})(\bar{x_i}-\bar{x})'$$

$$T=E+H$$

检验统计量为：

$$\Lambda=\frac{E}{E+H}$$

- [ ] BOX-M检验：协方差矩阵相等性的检验。
- [ ] 判别分析（距离判别、Fisher判别、Bayes判别，逐步判别）
- [ ] 最大后验概率法、最小期望误判代价法ECM
- [ ] 聚类分析中距离与相似度的定义
- [ ] 系统聚类法（Q型聚类（对样本）、R型聚类（对变量））
- [ ] 动态聚类法
- [ ] 主成分分析中的重要概念（主成分、贡献率、累积贡献率、原始变量与主成分之间的相关系数、主成分对原始变量的贡献率、主成分在原始变量上的载荷）
- [ ] 正交因子模型及假定：

$$x=\mu+Af+\epsilon$$

$$\begin{cases} E(f)=0 \\ E(\epsilon)=0 \\ V(f)=0 \\ V(\epsilon)=D=diag(\sigma_1^2,\sigma_2^2,...,\sigma_p^2) \\ cov(f,\epsilon)=E(f\epsilon')=0 \end{cases}$$

☐ 正交因子模型的性质

1.协方差阵有分解：$\Sigma = AA' + D$。

2.模型不受单位的影响。

3.因子载荷不是唯一的。

☐ 因子载荷矩阵A的统计意义

A的元素：$a_{ij} = Cov(x_i, y_j)$。经过标准化后，aij表示相关系数。

A的行元素平方和：$h_i^2 = \sum_{j=1}^{m} a_{ij}^2$称为共性方差，表示公共因子$f_1, f_2, ..., f_m$对$x_i$的方差贡献。

A的列元素平方和：$g_j^2 = \sum_{i=1}^{p} a_{ij}^2$反映了公共因子$f_j$对$x_1, x_2, ..., x_p$的影响。是衡量公共因子重要性的一个尺度。

A的元素平方和：$tr(AA') = tr(A'A) = \sum_{i=1}^{p}\sum_{j=1}^{m} a_{ij}^2$是m个公共因子对总方差的累积贡献。

☐ A与D的估计（会考！）

主成分法

$$\begin{aligned} S &= \hat{\lambda_1}\hat{t_1}\hat{t_1}' + ... + \hat{\lambda_m}\hat{t_m}\hat{t_m}' + \hat{\lambda_{m+1}}\hat{t_{m+1}}\hat{t_{m+1}}' + ... + \hat{\lambda_p}\hat{t_p}\hat{t_p}' \\ &\approx \hat{\lambda_1}\hat{t_1}\hat{t_1}' + ... + \hat{\lambda_m}\hat{t_m}\hat{t_m}' + \hat{D} = \hat{A}\hat{A}' + \hat{D} \end{aligned}$$

其中$A = (\sqrt{\hat{\lambda_1}}\hat{t_1}, ..., \sqrt{\hat{\lambda_m}}\hat{t_m})$，$\hat{D}$是由$\hat{\lambda_{m+1}}\hat{t_{m+1}}\hat{t_{m+1}}' + ... + \hat{\lambda_p}\hat{t_p}\hat{t_p}'$的对角线元素构成的对角矩阵。

2.主因子法

Step1. 先算出x的相关矩阵R。

Step2. 对特殊方差取一个合适的初始估计$\hat{\sigma_i^2}$。有三种选择：

一：取$\hat{\sigma_i^2} = \frac{1}{r^{ii}}$。其中$r^{ii}$是$\hat{R}^{-1}$的第i个对角线元素。此时的共性方差是xi与其他p-1个变量间样本复相关系数的平方。该估计要求R要满秩。

二：取$\hat{\sigma_i^2} = 1 - max_{j\neq i}|r_{ij}|$。

三：取$\hat{\sigma_i^2} = 0$，得到的是主成分解。

Step3. 算出$\hat{R}^*$.

$$\hat{R}^* = \begin{pmatrix} 1-\hat{\sigma_1^2} & r_{12} & ... & r_{1p} \\ r_{21} & 1-\hat{\sigma_2^2} & ... & r_{2p} \\ ... & ... & ... & ... \\ r_{p1} & r_{p2} & ... & 1-\hat{\sigma_p^2} \end{pmatrix}$$

Step4. 求出$\hat{R}^*$的特征值特征向量。此时，

$$\hat{A} = (\sqrt{\hat{\lambda_1^*}}\hat{t_1^*}, ..., \sqrt{\hat{\lambda_m^*}}\hat{t_m^*})$$

Step5. 令$\hat{R}^* = \hat{A}\hat{A}'$。重复以上步骤，直到收敛为止。

注⚠️：（1） $\hat{R}^*$常常有小的负特征值。一是因为假定不一定能成立；二是矩阵相减。（2）当因子数增加时，原来因子的估计载荷不变，fj对x的总方差贡献率仍然是 令$g_j^2 = \lambda_j$。（3）若从样本协方差矩阵S出发来求解主因子解，则可以采用$\hat{\sigma_i^2} = \frac{1}{s^{ii}}$来做估计。

3.极大似然法

$$\begin{gathered} \mu = \bar{x} \\ \hat{\Sigma}\hat{D}^{-1}\hat{A} = \hat{A}(I_m + \hat{A}'\hat{D}^{-1}\hat{A}) \\ \hat{D} = diag(\hat{\Sigma} - \hat{A}\hat{A}') \end{gathered}$$

唯一性条件：

$$A'D^{-1}A\text{是对角矩阵}$$

☐ 选择因子旋转的角度的标准为：最大方差旋转法。这里的方差指的是相对方差之和（P244）。

☐ 因子得分（加权最小二乘法、回归法）

1.加权最小二乘法

偏差平方和的矩阵表示为: $(x-\mu-A\hat{f})'D^{-1}(x-\mu-A\hat{f})$

此时的加权最小二乘法得分（巴特莱特得分）为：$\hat{f} = (A'D^{-1}A)^{-1}A'D^{-1}(x-\mu)$

2.回归法

回归法得分（Tompson得分）：

$$\hat{f} = A'(AA'+D)^{-1}(x-\mu) = (I+A'D^{-1}A)^{-1}A'D^{-1}(x-\mu)$$

⚠️：（1）加权最小二乘法得分是无偏的。回归法得分是有偏的。（2）回归法比加权最小二乘法更有效。

☐ 对应分析

▼ 对应分析

▼ 行轮廓和列轮廓

列联表

对应矩阵

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1q} \\ p_{21} & p_{22} & \cdots & p_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ p_{p1} & p_{p2} & \cdots & p_{pq} \end{pmatrix}$$

行密度向量（行和组成的向量）$r=P1=(p_{1+},p_{2+},...,p_{p+})'$

列密度向量（列和组成的向量）$c=1'P=(p_{+1},p_{+2},...,p_{+q})'$

我们引入两个矩阵

$$R=\begin{pmatrix} \frac{p_{11}}{p_{1+}} & \frac{p_{12}}{p_{1+}} & \cdots & \frac{p_{1q}}{p_{1+}} \\ \frac{p_{21}}{p_{2+}} & \frac{p_{22}}{p_{2+}} & \cdots & \frac{p_{2q}}{p_{2+}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_{p1}}{p_{p+}} & \frac{p_{p2}}{p_{p+}} & \cdots & \frac{p_{pq}}{p_{p+}} \end{pmatrix}$$

$$C=\begin{pmatrix} \frac{p_{11}}{p_{+1}} & \frac{p_{12}}{p_{+2}} & \cdots & \frac{p_{1q}}{p_{+q}} \\ \frac{p_{21}}{p_{+1}} & \frac{p_{22}}{p_{+2}} & \cdots & \frac{p_{2q}}{p_{+q}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_{p1}}{p_{+1}} & \frac{p_{p2}}{p_{+2}} & \cdots & \frac{p_{pq}}{p_{+q}} \end{pmatrix}$$

行轮廓

上述R的每一行称之为行轮廓。即$r_i'=(\frac{p_{i1}}{p_{i+}},\frac{p_{i2}}{p_{i+}},...,\frac{p_{iq}}{p_{i+}})$。其元素之和等于1。

列轮廓

上述C的每一列称之为列轮廓。即$c_j'=(\frac{p_{1j}}{p_{+j}},\frac{p_{2j}}{p_{+j}},...,\frac{p_{pj}}{p_{+j}})$。其元素之和等于1。

边缘对角矩阵：$D_r=diag(p_{1+},p_{2+},...,p_{p+}),D_c=diag(p_{+1},p_{+2},...,p_{+q})$。则有

$$R=(r_1',r_2',...,r_p')'=D_r^{-1}P;$$
$$C=(c_1,c_2,...,c_q)=PD_c^{-1}$$

由此，我们可以推出很重要的结论：

$$r = P1 = (PD_c^{-1})(D_c 1) = (c_1, c_2, ..., c_q)(p_{+1}, p_{+2}, ..., p_{+q})' = \sum_{j=1}^{q} p_{+j} c_j$$

$$c' = 1'P = (1'D_r)(D_r^{-1}P) = \sum_{i=1}^{p} p_{i+} r_i'$$

即：r是各列轮廓的加权平均。c是各行轮廓的加权平均。

▼ 独立性的检验和总惯量

行、列独立的检验

$$\chi^2 = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{p_{ij} - p_{i+}p_{+j}}{p_{i+}p_{+j}} \sim \chi^2[(p-1)(q-1)]$$

总惯量

$$\text{总惯量} = \frac{\chi^2}{n} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{p_{ij} - p_{i+}p_{+j}}{p_{i+}p_{+j}}$$

引入卡方距离。

r_i到中心c的卡方距离定义为：$(r_i - c)' D_c^{-1} (r_i - c)$

c_j到中心r的卡方距离定义为：$(c_j - r)' D_r^{-1} (c_j - r)$

则此时 总惯量=$\sum_{i=1}^{p} p_{i+} (r_i - c)' D_c^{-1} (r_i - c) = \sum_{j=1}^{q} p_{+j} (c_j - r)' D_r^{-1} (c_j - r)$

P的中心化：$P* = P - rc'$

P的标准化：$Z = D_r^{-1/2}(P - rc')D_c^{-1/2} = U\Lambda V'$

$$\text{总惯量} = \sum_{i=1}^{p} \sum_{j=1}^{q} z_{ij}^2 = tr(ZZ') = \sum_{i=1}^{k} \lambda_i^2$$

▼ 行、列轮廓的坐标

由$Z = D_r^{-1/2}(P - rc')D_c^{-1/2} = U\Lambda V'$可以推出$P - rc' = D_r^{1/2} U \Lambda V' D_c^{1/2}$

令$A = D_r^{1/2} U; B = D_c^{1/2} V'$。

则$r_i' - c' = x_{i1} b_1' + ... + x_{ik} b_k'$， 即在$(b_1, b_2, ..., b_k)$坐标系中的坐标为$(x_{i1}, x_{i2}, ..., x_{ik})$;

$c_j - r = y_{j1}a_1 + ... + x_{jk}a_k$，即在$(a_1, ..., a_k)$坐标系中的坐标为$(y_{j1}, y_{j2}, ..., y_{jk})$。

此时有：

$$\sum_{j=1}^{p} p_{j+}x_{ji} = \sum_{j=1}^{q} p_{+j}y_{ji} = 0$$

$$\sum_{j=1}^{p} p_{j+}x_{ji}^2 = \sum_{j=1}^{q} p_{+j}y_{ji}^2 = \lambda_i^2$$

λ_i^2称之为第i主惯量。度量了在第i坐标轴上的变差，它反映了列联表数据在第i维上的信息量。

▼ 对应分析图

当$\sum_{i=1}^{m} \lambda_i^2 / \sum_{i=1}^{k} \lambda_i^2$足够大时，可将后面的k-m项去掉。即有：

$$r_i' - c' \approx x_{i1}b_1' + ... + x_{im}b_m'$$
$$c_j - r \approx y_{j1}a_1 + ... + y_{jm}a_m$$

为了作图，通常取m=1，2，3。

- 对应分析图的构建

$r_i' - c'$在由b_1, b_2构成的平面中的坐标为(x_{i1}, x_{i2});

$c_j - r$在由a_1, a_2构成的平面中的坐标为(y_{j1}, y_{j2});

a_1, b_1重叠，此轴上的贡献率为$\lambda_1^2 / \sum_{i=1}^{k} \lambda_i^2$；$a_2, b_2$重叠，此轴上的贡献率为$\lambda_2^2 / \sum_{i=1}^{k} \lambda_i^2$。累计贡献率为$\lambda_1^2 + \lambda_2^2 / \sum_{i=1}^{k} \lambda_i^2$。

- 行（列）点之间的距离

行点之间的距离：$d_{ij}^2(r) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \approx (r_i - r_j)' D_c^{-1} (r_i - r_j)$

列点之间的距离：$d_{ij}^2(c) = (y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2 \approx (c_i - c_j)' D_r^{-1} (c_i - c_j)$

- 行点和列点相近的意涵

（1）两个行（列）点越相近，说明两个行（列）轮廓越相似。

（2）如果一个行点和一个列点相近，则表明行、列两个变量的相应类别组合发生的实际频数一般会高于这两个变量相互独立情形下的期望频数，也就意味着该行类别与该列类别相关联。这种关联程度约为行点或列点到原点的加权平方欧式距离。

（3）综上，对于相近的行点（或列点），他们离原点越远，其关联性就越强，也就是其类别组合的实际频数越是明显高于独立的情形。如果他们都在原点附近，则其关联性一般较弱。

☐ 典型相关分析

▼ 典型相关分析

▼ 总体典型相关

- 典型相关的定义及导出

$x=(x_1,x_2,...,x_p),y=(y_1,y_2,...,y_n)$是两组随机变量，满足：

$$V=\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$$

想法：想找出x_i的线性组合与y_i的线性组合，使得这两个线性组合之间的相关系数是最大的。即

$$\rho(u,v)=a'\Sigma_{12}b$$

达到最大。

设$\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$的特征值为$\rho_1^2,...,\rho_m^2$，其特征向量为$\beta_1,...,\beta_m$。

令

$$\alpha_i=\frac{1}{\rho_i}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{1/2}\beta_i$$

此时令，$a_i=\Sigma_{11}^{-1/2}\alpha_i$，$b_i=\Sigma_{22}^{-1/2}\beta_i$，称为第i对典型相关向量。

第一对典型变量：$u_1=a_1'x,\ v_1=b_1'x$。

第i对典型变量：$u_i=a_i'x,\ v_i=b_i'x$。

- 典型变量的性质

(1)同一组的典型变量互不相关且均有单位方差。

(2)不同组的典型变量之间的相关性：若$i=j$,则相关系数为ρ_i。若$i\neq j$,则相关系数为0。

(3)原始变量与典型变量之间的相关系数：

记

$$A=(a_1,a_2,...,a_m),B=(b_1,b_2,...,b_m)$$

则

$$u = A'x, v = B'y$$

u和v与x和y的相关系数

$$\rho(x,u) = D_1^{-1}\Sigma_{11}A; \rho(x,v) = D_1^{-1}\Sigma_{12}B$$
$$\rho(y,u) = D_2^{-1}\Sigma_{21}A; \rho(y,v) = D_2^{-1}\Sigma_{22}B$$

(4)典型相关系数是某种复相关系数

$$\rho_{u_i,y} = \rho_i$$
$$\rho_{v_j,x} = \rho_i$$

(5)简单相关、复相关和典型相关之间的关系:

(1)$p = q = 1$时，x与y之间的典型相关就是他们的简单相关。

(2)$p = 1$或$q = 1$时，x与y之间的典型相关就是他们之间的复相关。

▼ 样本典型相关

需注意的是，此时所有统计量都是对于样本而言的。

▼ 典型相关系数的显著性检验

检验问题：$H_0: \rho_1 = \rho_2 = ... = \rho_m = 0; H_1:$ ρ_i不全相等。

等价于：$H_0: \Sigma_{12} = 0; H_1: \Sigma_{12} \neq 0$

检验统计量为：$Q_1 = -[n - \frac{1}{2}(p+q+3)]ln\Lambda_1$,其中$\Lambda_1 = \prod_{i=1}^{m}(1-r_i^2)$。
服从卡方分布。自由度为pq。

多元统计假设检验部分

- 单总体的均值检验

1.Σ已知

检验统计量：$n(\bar{x}-\mu_0)'\Sigma^{-1}(\bar{x}-\mu_0) \sim \chi^2(p)$

2.Σ未知

检验统计量：$n(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0) \sim T^2(p, n-1)$

3.置信区域

$$n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \leq T_\alpha^2(p, n-1)$$

4.置信区间

(1)T2联合置信区间

$$a_i'\bar{x} - T_\alpha^2(p, n-1)\sqrt{a_i'Sa_i}/\sqrt{n} \le a_i'\mu \le a_i'\bar{x} + T_\alpha^2(p, n-1)\sqrt{a_i'Sa_i}/\sqrt{n}$$

(2)Bonferroni联合置信区间

$$a_i'\bar{x} - t_{\alpha/2k}(n-1)\sqrt{a_i'Sa_i}/\sqrt{n} \le a_i'\mu \le a_i'\bar{x} + t_{\alpha/2k}(n-1)\sqrt{a_i'Sa_i}/\sqrt{n}$$

- 双总体的均值检验

1.两个独立样本

检验：

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

$$T^2 = \frac{n_1n_2}{n_1+n_2}(\bar{x}-\bar{y})'S_p^{-1}(\bar{x}-\bar{y}) \sim T_\alpha^2(p, n_1+n_2-2)$$

置信区域：

$$\frac{n_1n_2}{n_1+n_2}(\bar{x}-\bar{y}-\mu_1+\mu_2))'S_p^{-1}(\bar{x}-\bar{y}-\mu_1+\mu_2)) \le T_\alpha^2(p, n_1+n_2-2)$$

置信区间：

T2:

$$a_i'(\bar{x}-\bar{y}) \pm T_\alpha(p, n_1+n_2-2)\sqrt{\frac{n_1+n_2}{n_1n_2}}\sqrt{a_i'S_pa_i}$$

Bonferroni:

$$a_i'(\bar{x}-\bar{y}) \pm t_{2\alpha/k}(n_1+n_2-2)\sqrt{\frac{n_1+n_2}{n_1n_2}}\sqrt{a_i'S_pa_i}$$

2.成对样本

$$(x_i, y_i) \to d_i = x_i - y_i \sim N_p(\delta = \mu_1 - \mu_2, \Sigma)$$

希望检验$\mu_1 = \mu_2$。等价于检验$\delta = 0$。检验统计量为：

$$T^2 = n\bar{d}' S_d^{-1} \bar{d} \sim T^2(p, n-1)$$
$$\text{其中} \bar{d} = \bar{x} - \bar{y}, S_d = \frac{1}{n-1}\Sigma_{i=1}^{n}(d_i - \bar{d})(d_i - \bar{d})'$$

- 轮廓分析**（Hotelling T方分布的第一个参数是p-1！p-1！p-1！！是对应矩阵的秩！！）**

(1)单总体的轮廓分析

$$H_0 : \mu_1 = \mu_2 = ... = \mu_p$$

$$C = \begin{pmatrix} 1 & -1 & 0 & ... & 0 \\ 1 & 0 & -1 & ... & 0 \\ .. & .. & .. & .. & .. \\ 1 & 0 & 0 & ... & -1 \end{pmatrix}$$

$$T^2 = n(C\bar{x})'(CSC')^{-1}(C\bar{x}) \sim T^2(p-1, n-1)$$

(2)两总体的轮廓分析

平行吗？

$$C = \begin{pmatrix} -1 & 1 & 0 & ... & 0 \\ 0 & -1 & 1 & ... & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & ... & 1 \end{pmatrix}$$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(C(\bar{x} - \bar{y}))'(CS_pC')^{-1}(C(\bar{x} - \bar{y})) \sim T^2(p-1, n_1 + n_2 - 1)$$

重合吗？

$$H_0 : \frac{\mu_{11} + \mu_{12} + ... + \mu_{1p}}{p} = \frac{\mu_{21} + \mu_{22} + ... + \mu_{2p}}{p}$$

等价于

$$H_0 : 1'\mu_1 = 1'\mu_2$$

即$C = 1'.$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(1'(\bar{x} - \bar{y}))'(1'S_p1)^{-1}(1'(\bar{x} - \bar{y})) \sim F(1, n_1 + n_2 - 1)$$

水平吗？

$$\bar{Z}=\frac{n_1}{n_1+n_2}\bar{x}+\frac{n_2}{n_1+n_2}\bar{y}$$

新样本协方差矩阵记为S

$$C=\begin{pmatrix}1 & -1 & 0 & ... & 0\\ 1 & 0 & -1 & ... & 0\\ .. & .. & .. & .. & ..\\ 1 & 0 & 0 & ... & -1\end{pmatrix}$$

$$T^2=(n_1+n_2)(C\bar{z})'(CSC')^{-1}C\bar{z}$$

一、判断题

- [ ] 1.协方差矩阵的性质

$$cov(X,Y)=cov^{'}(Y,X)$$

- [ ] 2.马氏距离是x和y经标准化之后的欧式距离。
- [ ] 3.x服从多元正态分布，当且仅当它的任何线性函数$a'x$均服从一元正态分布。
- [ ] 4.多元正态分布的任何边缘分布仍为多元正态分布。但是，一个随机向量的任何边缘分布均为正态，并不表明它服从多元正态分布。
- [ ] 5.多元正态变量的线性变换、边缘分布、条件分布仍是多元正态的。
- [ ] 6.y与预测值$\hat{y}$的样本相关系数等于x与$x_1,x_2,...,x_p$的样本复相关系数。
- [ ] 7.偏相关系数度量了剔除$x_{k+1},...,x_p$的影响后，x_i与x_j间相关关系的强弱。
- [ ] 8.复相关系数度量了一个随机变量y和一组随机变量$x_1,x_2,...,x_n$之间的相关关系的强弱，它也是y和用$x_1,x_2,...,x_n$对y的最优线性预测的相关系数。复相关可看成是典型相关的一个特例。
- [ ] 9.对于多元正态变量，偏相关系数与条件相关系数是相同的，此时的条件相关系数与已知条件变量的取值无关。
- [ ] 10.偏相关系数为0，并不意味着相关系数为0。反之亦然。偏相关系数和相关系数未必同号。偏相关系数和相关系数之间孰大孰小没有必然的结论。
- [ ] 11.BOX-M检验对某些非正态情形非常敏感。即便M检验拒绝了H0,我们仍可继续使用多元方差分析。
- [ ] 12.对于成对实验的数据，减少了抽样误差，往往可以得到比独立样本方法更精确的结论。
- [ ] 13.多元方差分析中的检验统计量并不惟一，可以有多个。

- [ ] 14.偏相关系数与简单相关系数的检验方法是完全类似的，主要区别是检验统计量分布的自由度有所不同。
- [ ] 15.判别分析中，回代法给出的估计值通常偏低。交叉验证法是接近无偏的估计量，最值得推荐的方法。
- [ ] 16.样本容量小时，选择线性判别函数。容量大时，选择二次判别函数。
- [ ] 17.对于正态组及各协方差矩阵相同的情形下，距离判别等价于各先验概率均相等时的贝叶斯判别（最大后验概率法）。
- [ ] 18.密度函数比、误判代价比、先验概率比中，误判代价比最具有实际意义。
- [ ] 19.代价相同C(1|2)=C(2|1)=c，此时的判别规则将使总的误判概率达到最小。
- [ ] 20.概率比=代价比，即p1/p2=c(1|2)/c(2/1)，通常取p1=p2=0.5且c(1|2)=c(2|1)，此时的判别规则将使两个误判概率之和P(1|2)+P(2|1)达到最小。
- [ ] 21.最大后验概率法可看成是所有误判代价均相同时的最小期望误判代价法。
- [ ] 22.在Fisher判别中，我们需要假定各组的协方差阵都相同。
- [ ] 23.逐步判别法中，确定临界值时让“F出”比“F进”略小一点。（严进宽出）
- [ ] 24.逐步判别实际上是在做逐步多元方差分析，在确定好变量后再计算判别函数，确定判别规则。
- [ ] 25.距离判别和贝叶斯判别只能用于分类。Fisher判别可用于分类，也能用于分离。
- [ ] 26.对于两组的判别，Fisher判别等价于协方差矩阵相等的距离判别。
- [ ] 27.对于两组正态，Fisher判别等价于协方差相等且先验概率和误判都相等代价的贝叶斯判别。
- [ ] 28.欧式距离对异常值较为敏锐。通常来说，在Minkowski距离中，q越大，则越容易被异常值影响。
- [ ] 29.马氏距离在聚类中的缺陷为：“类”一直变化着，使得类内的样本协差阵难以确定，除非有关于不同类的先验知识。
- [ ] 30.在经济变量分析中，常用相关系数来描述变量间的相似性程度。
- [ ] 31.主成分分析中，表明了主成分的方向，y_i是x在t_i上的投影值，λ_i是这些投影值的差异，反映了t_i上投影点的变异程度。
- [ ] 32.如果后几个主成分的贡献率都非常小，则可以表示变量之间有几个彼此独立的多重共线性关系（这句话可以说成：特征值很小的主成分能够揭示出原始变量间的多重共线性关系）。如果$V(y_p)=0$，则表明$x_1,...,x_p$之间以概率1存在线性关系。
- [ ] 33.在进行主成分分析前，应将变量标准化。
- [ ] 34.解释主成分时，既要考察主成分在原始变量上的载荷，也应考察主成分与原始变量的相关系数，而考察前者更为重要。若求出的主成分是从相关矩阵出发的，则对前者和

后者的考察是等价的。

- [ ] 35.在正交因子模型中，虽然因子$f_1, ..., f_m$能100%地解释原始变量之间所有协方差或相关系数，但并不能保证这些因子一定能解释$x_1, ..., x_p$总方差的多大比例。理论上比该比例可以是较低的。
- [ ] 36.对极大似然解，各因子所解释的总方差的比例未必像主成分解以及主因子解那样依次递减。当因子数增加时，原来因子的估计载荷及对x的贡献率将发生变化。
- [ ] 37.考虑因子旋转，使得旋转之后的载荷矩阵在每一列元素的绝对值尽量地大小拉开。
- [ ] 38.因子旋转并非一定能有利于因子的解释，它只是提供了因子解释成功的更大可能性和更多机会。
- [ ] 39.因子旋转不改变共性方差；因子旋转不改变m个因子的累积贡献率；因子旋转不改变残差矩阵。
- [ ] 40.因子得分的估计并不是通常意义下的参数估计，而是对不可观测的随机变量$f_1, f_2, ..., f_m$的取值做出估计。
- [ ] 41.因子分析是主成分分析的推广，也是一种降维技术。
- [ ] 42.主成分法和主因子法是在求解的过程中确定因子数m的。而极大似然法要在求解之前确定m。
- [ ] 43.对应分析是用于寻找列联表的行和列之间关联的一种低维图形表示法，它同时可以揭示同一分类变量的各个类别之间的差异。
- [ ] 44.列（行）边缘频率向量可以表示成各行（列）轮廓的加权平均，故可将其视为各行（列）轮廓的中心。
- [ ] 45.总惯量既度量了列联表中行、列变量之间关联性，也度量了行（列）轮廓之间的总变差。
- [ ] 46.主惯量度量了在一坐标轴上的变差。
- [ ] 47.典型相关分析是研究两组变量之间相关性的一种统计分析方法，它是一种降维技术。复相关是典型相关的特例，简单相关是复相关的一个特例。
- [ ] 48.第一对典型相关包含有最多的有关两组变量间相关的信息，第二对其次，其他对依次递减。各对典型相关变量所含的信息互不重复。
- [ ] 49.经标准化的两组变量间的典型相关系数与原始的两组变量间的典型相关系数是相同的。前者的典型变量是后者的中心化值。
- [ ] 50.第一典型相关系数至少同x的任意分量与y的复相关系数一样大，即便复相关系数都很小，第一典型相关系数仍可能很大。复相关系数与简单相关系数也是同样的关系。
- [ ] 51.行列轮廓都表示一个条件（频率）分布，两个行（列）轮廓相近意味着两个行（列）有相似的条件分布。

- [ ] 52.对应分析的结果（包括总惯量）只依赖于对应矩阵P，而与n没有关系。独立性检验与n有关。
- [ ] 53.总惯量度量了行轮廓之间的总变差，也度量了列轮廓之间的总变差。行和列之间的关联性越强，则行（列）轮廓之间的差异性就越大。反之亦然。
- [ ] 54.各行点在坐标系bi上坐标的加权平均值为0。各列点在坐标系ai上坐标的加权平均值为0。
- [ ] 55.综上，对于相近的行点和列点，他们离原点越远，其关联性就越强，也就是其类别组合的实际频数越是明显高于独立的情形。如果他们都在原点附近，则其关联性一般较弱。
- [ ] 56.我们观察哪些行类别，哪些列类别之间关系密切，不能只是直接比较个类别组合的原始频数大小。这样的比较之所以不合理，是因为个类别的合计频数并不相同，合计频数高的类别，与其有关的类别组合频数自然会相对偏高。
- [ ] 57.当ni都超过20，p和k都不超过5时，BOX的卡方拟合度较高。

二、选择题

- [ ] 小测当中的选择题
- [ ] 在两组皆为正态且协方差矩阵相等的情形下，判别规则（）在使两个误判概率之和达到最小的意义上是最优的。

答案：

$$\begin{cases} x \in \pi_1, W(x) \geq 0 \\ x \in \pi_2, W(x) < 0 \end{cases}$$

$$W(x) = a'(x - \bar{\mu}); a = \Sigma^{-1}(\mu_1 - \mu_2); \bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$$

- [ ] Fisher判别函数的特点：

（1）各判别函数都具有单位方差

（2）各判别函数彼此之间不相关

（3）判别函数方向$t_1, t_2, ..., t_r$一般不正交。

（4）判别函数不受变量单位的影响。

- [ ] 被常称作“城市街区距离”的距离是：（A绝对值距离）

A（绝对值距离）

B（欧式距离）

C（切比雪夫距离）

D（马氏距离）

☐ 系统聚类的各个方法的特点：

最短距离法（最容易产生“结”，有链接倾向）

最长距离法（会被异常值夸大）

重心法（比较稳健）

离差平方和法（两个大类有大的距离，两个小类有小的距离。符合聚类的想法）

☐ 聚类方法的单调性：

有单调性：最短距离、最长距离、类平均、离差平方和、可变法、可变类平均法

没有单调性：重心法、中间距离法。

☐ 总惯量为0与以下三种情形都等价：

（1）$p_{ij} = p_{i+}p_{+j}$。也就是说，行和列完全独立。

（2）所有的列轮廓相等。

（3）所有的行轮廓相等。

三、大题

☐ 主成分分析（必考）

☐ 聚类分析（必考）

☐ 距离判别、贝叶斯判别（必考）

☐ 因子分析的主成分法与主因子法之间的关系推导

☐ 条件分布（条件期望、条件方差）

☐ 统计推断、联合置信区间、轮廓分析