**1、判别分析与聚类分析有何不同？**

聚类分析和判别分析有相似的作用，都是起到分类的作用。但是判别分析是已知分类然后总结出判别规则，是一种有指导的学习；而聚类分析则是有了一批样本，不知道它们的分类，甚至连分成几类都不知道，希望用某种方法把观测进行合理的分类，使得同一类的观测比较接近，不同类的观测相差较多，这是无指导的学习。所以聚类分析依赖于对观测间的接近程度（距离）或相似程度的理解，定义不同的距离量度和相似性量度就可以产生不同的聚类结果。

**2、简述聚类分析的基本思想。有哪两类聚类分析？各自的作用？**

聚类分析就是根据空间点群的“亲疏”关系进行分类的一种方法。为此要给出表示空间点与点之间“亲疏”关系的相似性度量，然后讨论根据相似性度量进行点群簇分的方法和应用。

聚类分析的目的是把分类对象按一定规则分成若干类，这些类不是事先给定的，而是根据数据的特征确定的。在同一类中这些对象在某种意义上趋向于彼此相似，而在不同类中对象趋向于不相似。聚类分析根据对象不同分为$Q$型聚类分析（对样本进行聚类）和$R$型聚类（对变量进行聚类）。

对样品或变量进行聚类时，我们常用距离和相似系数来对样品或变量之间的相似性进行度量。距离用来度量样品之间的相似性，而相似系数常用来度量变量间的相似性。

**3、距离系数需要满足的基本条件？**

点$i$和点$j$之间的距离$d_{ij}$可有各种不同的定义，只要其满足所谓的距离公理：

对一切$i, j$，$d_{ij} \geq 0$；

$d_{ij} = 0$等价于点$i$和点$j$为同一点，即$x(i) = x(j)$；

对一切的$i, j$，$d_{ij} = d_{ji}$；

三角不等式成立，即对一切的$i, j, k$，有$d_{ij} \leq d_{ik} + d_{kj}$。

**4、系统聚类法的基本思想和步骤。有哪些常用的系统聚类法？**

基本思想：

（1）将聚类的$n$个样品（或者变量）各自看成一类，共有$n$类；

（2）按照事先选定的方法计算每两类之间的聚类统计量，即某种距离（或者相似系数），将关系最密切的两类并为一类，其余不变，即得$n-1$类；

（3）按前面的计算方法计算新类与其他类之间的距离（或者相似系数），将关系最密切的两类并为一类，其余不变，即得$n-2$类；

（4）如此继续下去，直到最后所有样品（或者变量）归为一类为止。

基本步骤：

（1）$n$个样品（或者变量）各自成一类，一共有$n$类。计算两两之间的距离，显然$D(G_p,G_q)=d_{pq}$，构成一个对称矩阵$D_{(0)}=(d_{ij})_{n\times n}$，其对角线上的元素全为0。

（2）选择$D_{(0)}$中对角线元素以外的上（或者下）三角部分中的最小元素，设其为D（$G_p,G_q$），与其下标相对应，将类$G_p$与$G_q$合并成一个新类，记为$G_r$。计算$G_r$与其他类$G_k$（$k\neq p,q$）之间的距离。

（3）在$D_{(0)}$中划去与$G_p$、$G_q$所对应的两行和两列，并加入由新类$G_r$与其他各类之间的距离所组成的一行和一列，得到一个新的n-1阶对称距离矩阵$D_{(1)}$。

（4）由D(1)出发，重复步骤（2）（3）得到对称矩阵D(2)；再由D(2)出发，重复步骤（2）（3）得到对称矩阵D(3),...，依次类推，直到n个样品（或者变量）聚为一个大类为止。

（5）在合并某两类的过程中记下两类样品（或者变量）的编号以及所对应的距离（或者相似系数），并绘制成果聚类图。

（6）决定类的个数以及聚类结果。

常用的系统聚类法有：最短距离法、最长距离法、中间距离法、重心法、来平均法、离差平方和法。

**5、如何确定合理的聚类数目？**

聚类数目的真正确定在于研究的问题是什么，以及事先有无一个大致的判断标准。分类的数目应该符合使用的目的。确定聚类数的问题属于聚类有效性问题。比如在模糊聚类分析中，可以根据方差分析理论，应用混合F统计量来确定最佳分类数。

**6、在进行系统聚类分析时，不同的类间距离计算方法有何区别？请举例说明。**

（1） 最短距离法

定义类$p$与类$q$之间的距离为两类最近样品的距离，即：

$$d_{pq}=\min_{i\in p,j\in q}\{d_{ij}\}$$

设类 $p$ 与 $q$ 合并成一个新类，记为 $k$，则 $k$ 与任一类 $r$ 的距离是：

$$d_{kr}=\min\{d_{pr},d_{qr}\}$$

（2） 最长距离法

定义类 $p$ 与类 $q$ 之间的距离为两类最远样品的距离，即：

$$d_{pq}=\max_{i\in p,j\in q}\{d_{ij}\}$$

设类 $p$ 与 $q$ 合并成一个新类，记为 $k$，则 $k$ 与任一类 $r$ 的距离是：

$$d_{kr}=\max\{d_{pr},d_{qr}\}$$

（3） 中间距离法

设类 $p$ 与类 $q$ 合并成一个新类，记为 $k$，则 $k$ 与任一类 $r$ 的距离是：

$$d_{kr}^2=\frac{1}{2}d_{pr}^2+\frac{1}{2}d_{qr}^2-\frac{1}{4}d_{pq}^2$$

（4） 重心法

设 $p$ 和 $q$ 的重心分别是 $\bar{x}_p$ 和 $\bar{x}_q$，则类 $p$ 和 $q$ 的距离是 $d_{pq}=d_{\bar{x}_p\bar{x}_q}$

设聚类到某一步，类 $p$ 与类 $q$ 分别有样品 $n_p$、$n_q$ 个，将 $p$ 和 $q$ 合并成 $k$，则 $k$ 类的样品个数为 $n_k=n_p+n_q$，它的重心是 $\bar{x}_k=\frac{1}{n_k}(n_p,\bar{x}_p+n_q\bar{x}_q)$，某一类 $r$ 的重心是 $\bar{x}$，它与新类 $k$ 的距离是：

$$d_{kr}^2=\frac{n_p}{n_q}d_{pr}^2+\frac{n_q}{n_k}d_{qr}^2-\frac{n_pn_q}{n_k^2}d_{pq}^2$$

（5） 类平均法

它定义两类之间的距离平方为这两类元素两两之间距离平方的平均，即：

$$d_{pq}^2=\frac{1}{n_pn_q}\sum_{i\in p}\sum_{j\in q}d_{ij}^2$$

设聚类到某一步，类 $p$ 与 $q$ 分别有样品 $n_p$、$n_q$ 个，将 $p$ 和 $q$ 合并为 $k$，则 $k$ 类的样品个数为 $n_k=n_p+n_q$，$k$ 类与任一类 $r$ 的距离为：

$$d_{kr}^2=\frac{1}{n_k n_r}\sum_{i\in p}\sum_{j\in q}d_{ij}^2=\frac{n_p}{n_k}d_{pr}^2+\frac{n_q}{n_k}d_{qr}^2$$

（6）离差平方和法

$$S_t=\sum_{t=1}^{n_t}(X_{it}-\bar{X}_t)'(X_{it}-\bar{X}_t)$$

$$D_{pq}^2=\frac{n_p n_q}{nr}=(\bar{x}_p-\bar{x}_q)^{'(\bar{x}_p-\bar{x}_q)}=S_r-S_p-S_q$$

$$D_{kr}^2=\frac{n_k+n_p}{n_r+n_k}D_{kp}^2+\frac{n_k+n_q}{n_r+n_k}D_{kq}^2-\frac{n_k}{n_r+n_k}D_{pq}^2$$

**7、Q 型聚类统计量**

考虑对样品进行聚类，描述变量之间的接近程度常用“距离”来度量。两个样品之间的距离越小，表示两者之间的共同点越多；距离越大，共同点越少。常用距离有：绝对值距离、欧式距离、闵克夫斯基距离、切比雪夫距离、马哈拉诺比斯距离

**8、R 型聚类统计量**

考虑对样品进行聚类，描述变量之间的接近程度常用“相似系数”来度量。两个变量之间的相似系数的绝对值越接近于 1，表示两者关系越密切；绝对值越接近于 0，关系越疏远。常用相似距离有：夹角余弦和相似系数。

**9、简述主成分分析的基本思想。**

主成分分析的基本思想是构造原始变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少量几个新变量并使它们含有足够多的原始变量带有的信息，从而使得用这几个新变量代替原始变量分析问题和解决问题成为可能。

**10、主成分的求取**

首先，求其协方差矩阵 Σ 的各特征值及相应的正交单位化特征向量，然后，以特征值从大到小所对应的特征向量为组合系数所得到的 X1, X2, . . . , Xp 的线性组合分别取作 X 的第一、第二、直至第 p 个主成分，而各主成分的方差等于相应的特征值。

**11、主成分分析的基本思想，可以做什么应用及在应用中要选几个主成分？**

主成分分析的基本思想：构造原始变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少量几个新变量并使它们含有足够多的原始变量带有

的信息，从而使得用这几个新变量代替原始变量分析问题和解决问题成为可能。通常变量中所含信息的多少用该变量的方差（或样本方差）来度量，这是经典的信息量的表示方法。

解决的问题：

（1）研究的问题当中，随机变量的个数比较大，将增大计算量和分析问题的复杂性；

（2）随机变量之间存在着一定的相关性，它们的观测样本所反映的信息在一定程度上存在着重叠的。

一般地，在约束条件① $l_i^T l_i = 1$ ②$Cov(Y_i, Y_k) = l_i^T \Sigma l_k = 0, k = 1, 2, \ldots, i-1$ 之下，使得 $Var(Y_i)$ 达到最大，由此 $l_i$ 确定的 $Y_i = l_i^T X$ 称为 $X_1, X_2, \ldots, X_p$ 的第 i 个主成分。

**12、比较主成分分析与判别分析的基本思想。**

主成分分析就是一种通过降维技术把多个指标约化为少数几个综合指标的统计分析方法。其基本思想是：设法将原来众多具有一定相关性的指标（设为 p 个），重新组合成一组新的相互无关的综合指标来代替原来指标。数学上的处理就是将原来 P 个指标作线性组合，作为新的指标。第一个线性组合，即第一个综合指标记为 $Y_1$，为了使该线性组合具有唯一性，要求在所有线性组合中 $Y_1$ 的方差最大，即 $Var(Y_1)$ 越大，那么包含的信息越多。如果第一个主成分不足以代表原来 p 个指标的信息，再考虑选取第二个主成分 $Y_2$，并要求 $Y_1$ 已有的信息不出现在 $Y_2$ 中，即主成分分析是将分散在一组变量上的信息集中到某几个综合指标上的探索性统计分析方法。以便利用主成分描述数据集内部结构，实际上也起着数据降维作用。

聚类分析的目的是把分类对象按一定规则分成若干类，这些类不是事先给定的，而是根据数据的特征确定的。在同一类中这些对象在某种意义上趋向于彼此相似，而在不同类中对象趋向于不相似。聚类分析根据对象不同可分为 Q 型聚类分析（对样本进行聚类）和 R 型聚类分析（对变量进行聚类）。

对样本或变量进行聚类时，我们常用距离和相似系数来对样品或变量之间的相似性进行度量。距离常用来度量样品之间的相似性，而相似系数常用来度量变量间的相似性。

**13、简述典型变量与典型相关系数的概念，并说明典型相关分析的基本思想。**

在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数。选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对，如此下去直到两组之间的相关性被提取完毕为止。被选出的线性组合配对称为典型变量，它们的相关系数称为典型相关系数。

**14、因子分析的基本思想？**

因子分析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结构出发，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多元统计分析方法，因子分析的基本思想是根据相关性大小将变量分组，使得同组内的变量之间相关性较高，不同组的变量相关性较低。每一组变量代表一个基本结构，用一个不可观测的综合变量表示，这个基本结构称为公共因子。对于所研究的问题就可用最少个数的不可观测的所谓公共因子的线性函数与特殊因子之和来描述原来观测的每一分量。

**15、比较主成分分析与因子分析的异同点。**

**相同点**：①两种分析方法都是一种降维、简化数据的技术。②两种分析的求解过程是类似的，都是从一个协方差阵出发，利用特征值、特征向量求解。因子分析可以说是主成分分析的姐妹篇，将主成分分析向前推进一步便导致因子分析。因子分析也可以说成是主成分分析的逆问题。如果说主成分分析是将原指标综合、归纳，那么因子分析可以说是将原指标给予分解、演绎。

**主要区别是**：主成分分析本质上是一种线性变换，将原始坐标变换到变异程度大的方向上为止，突出数据变异的方向，归纳重要信息。而因子分析是从显在变量去提炼潜在因子的过程。此外，主成分分析不需要构造分析模型而因子分析要构造因子模型。

**16、简述相应分析的基本思想。**

相应分析指受制于某个载体总体的两个因素为 A 和 B，其中因素 A 包含 r 个水平，即 $A_1, A_2, \cdots, A_r$；因素 B 包含即 c 个水平，即 $B_1, B_2, \cdots, B_c$。对这两组因素作随机抽样调查， 记为得到一个 r×c 的二维列联表，记为 $K=(K_{ij})_{r\times c}$，主要目的是寻求列联表行因素 A 和列因素 B 的基本分析特征和它们的最优联立表示。基本思想为通过列联表的转换，使得因素 A 和列因素 B 具有对等性，这样就可以用相同的因子轴同时描述两个因素各个水平的情况，把两个因素的各个水平的状

况同时反映到具有相同坐标轴的因子平面上，直观地描述两个因素 A 和因素 B 以及各个水平之间的相关关系。

**17、进行相应分析时在对因素*A*和因素*B*进行相应分析之前有没有必要进行独立性检验？为什么？**

有必要，如果因素A和因素B独立，则没有必要进行相应分析；如果因素A和因素B不独立，可以进一步通过相应分析考察两因素各个水平之间的相关关系。

**18、解释因子分析模型中，变量共同度与公因子方差贡献的统计意义。为什么有时候需要作因子旋转？有哪些估计因子得分的方法？因子得分的计算是不是通常意义下的参数估计？**

变量共同度的统计意义：

$$X_i^* = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i$$

两边求方差 $Var(X_i)=a^2_{i1}Var(F_1)+...+a^2_{im}Var(F_m)+Var(\varepsilon_i)$

$$1=\Sigma a_{ij}^2+\sigma_i^2=h_i^2+\sigma_i^2$$

所有的公共因子和特殊因子对变量 $X_i^*$的贡献为 1。$h_i^2$反映了全部公共因子对变量 $X_i^*$影响，是全部公共因子对变量方差所作出的贡献，或者说 $X_i^*$对公共因子的共同依赖程度，称为公共因子对变量 $X_i^*$的方差贡献。

$h_i^2$接近于 1，表明该变量的原始信息几乎都被选取的公共因子说明了。

$\sigma_i^2$特殊因子的方差，反映了原有变量方差中无法被公共因子描述的比例。

公因子方差贡献的统计意义：

是衡量公共因子相对重要性的指标，$g_j^2$越大，表明公共因子 $F_j$对 x 的贡献越大，或者说对 x 的影响和作用就越大。

一个正交变换对应坐标系的旋转，而且主因子的任一解均可由已求得的 A 经过旋转（右乘一个正交阵）得到。经过旋转后，公共因子对 $x_i$的贡献 $h_i^2$并不改变，但公共因子本身可能有较大变化，即 $g_j^2$不再与原来的值相同，从而可通过适当的旋转来得到我们比较满意的公共因子。

估计因子得分的方法较多，常用的有回归估计法，Bartlett 估计法，Thomson 估计法。

⑴回归估计法

$F = X\ b = X\ (X\ ¢X)-1A¢ = XR-1A¢$ （这里 R 为相关阵，且 $R = X\ ¢X$ ）。

⑵Bartlett 估计法

Bartlett 估计因子得分可由最小二乘法或极大似然法导出。

F = [(W-1/2A) ∁ W-1/2A]-1(W-1/2A) ∁ W-1/2X = (A∁W-1A)-1A∁W-1X

⑶Thomson 估计法

在回归估计法中，实际上是忽略特殊因子的作用，取 R = X ∁X，若考虑特殊因子的作用，此时 R = X ∁X+W，于是有：

F = XR-1A∁ = X (X ∁X+W)-1A∁

这就是 Thomson 估计的因子得分，使用矩阵求逆算法（参考线性代数文献）可以将其转换为：

F = XR-1A∁ = X (I+A∁W-1A)-1W-1A∁

将公共因子用变量的线性组合来表示，也即由地区经济的各项指标值来估计它的因子得分。

设公共因子 F 由变量 x 表示的线性组合为：

$F_j = u_{j1}\ x_{j1} + u_{j2}\ x_{j2} + \cdots + u_{jp}x_{jp}\quad j=1, 2, \cdots, m$

但因子得分函数中方程的个数 m 小于变量的个数 p，所以并不能精确计算出因子得分，只能对因子得分进行估计。

**19、试比较主成分分析、因子分析、对应分析这三种方法的异同之处并简要介绍它们的应用。**

主成分分析的基本思想是构造原始变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少量几个新变量并使它们含有足够多的原始变量带有的信息，从而使得用这几个新变量代替原始变量分析问题和解决问题成为可能。通常变量中所含信息的多少用该变量的方差（或样本方差）来度量，这是经典的信息量的表示方法。例如，高校科研状况评价中的立项课题数与项目经费、经费支出等之间会存在较高的相关性；学生综合评价研究中的专业基础课成绩与专业课成绩、获奖学金次数等之间也会存在较高的相关性。利用主成分分析既可以大大减少参与建模的变量个数，同时也不会造成信息的大量丢失。能够有效降低变量维数。

因子分析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结构出发，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 多元统计分析方法，因子分析的基本思想是根据相关性大小将变量分组，使得同组内的变量之间相关性较高，不同组的变量相关性较低。每一组变量代表一个基本结构，用一个不可观测的综合变量表示，这个基本结构称为公共因子。对于所研究的问题就可用最少个数的不可观测的所谓公共因子的线性函数与特殊因子之和来描述原来观测的每一分量。例如，某企业招聘人才，对每位应聘者进行外贸、申请书的形式、专业能力、讨人喜欢的能力、自信心、洞察力、诚信、推销本领、经验、工作态度、抱负、理解能力、潜在能力、实际能力、适应性的15个方面考核。这15个方面可归结为应聘者的表现力、亲和力、实践经验、专业能力4个方面，每一方面称为一个公告因子。企业可根据这4个公共因子的情况来衡量应聘者的综合水平。

对应分析是因子分析的进一步推广，也称关联分析、R-Q型因子分析，是近年新发展起来的一种多元相依变量统计分析技术，通过分析由定性变量构成的交互汇总表来揭示变量间的联系。可以揭示同一变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。对应分析的基本思想是将一个联列表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。

它最大特点是能把众多的样品和众多的变量同时作到同一张图解上，将样品的大类及其属性在图上直观而又明了地表示出来，具有直观性。另外，它还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程，可以从因子载荷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。

相应分析指受制于某个载体总体的两个因素为A和B，其中因素A包含r个水平，即 $A_1, A_2, \cdots, A_r$；因素B包含即c个水平，即 $B_1, B_2, \cdots, B_c$。对这两组因素作随机抽样调查， 记为得到一个r×c的二维列联表，记为 $K=(K_{ij})_{r\times c}$，主要目的是寻求列联表行因素A和列因素B的基本分析特征和它们的最优联立表示。基本思想为通过列联表的转换，使得因素A和列因素B具有对等性，这样就可以用相同的因子轴同时描述两个因素各个水平的情况，把两个因素的各个水平的状况同时反映到具有相同坐标轴的因子平面上，直观地描述两个因素 A 和因素 B 以及各个水平之间的相关关系。

共同点：

（1）都是用少数的几个变量（因子）来反映原始变量（因子）的主要信息。并且新的变量彼此不相关，消除了多重共线性。

（2）求解过程是类似的，都是从一个协方差阵出发，利用特征值、特征向量求解。

不同点：

（1）相对于主成分分析，因子分析更倾向于描述原始变量之间的相关关系。

（2）线性表示方向不同，因子分析和对应分析是把变量表示成公共因子的线性组合，而主成分分析则是把主成分表示成各变量的线性组合。

（3）主成分分析本质上是一种线性变换，将原始坐标变换到变异程度大的方向上为止，突出数据变异的方向，归纳重要信息。而因子分析和对应分析是从显在变量去提炼潜在因子的过程。此外，主成分分析不需要构造分析模型而因子分析和对应分析要构造因子模型。

（4）对应分析克服了因子分析的不足之处，可以寻找出 R 型和 Q 型分析间的内在联系，由 R 型分析的结果可以方便地得到 Q 型分析结果，克服了做 Q 型分析样品容量 n 很大时计算上的困难。

**20、因子分析的一般步骤**

1）将原始数据标准化

2）建立变量的相关系数矩阵 R

3）求 R 的特征根及相应的单位特征向量，根据累积贡献率要求，取前 m 个特征根及相应的特征向量，写出因子载荷阵 A

4）对 A 施行因子旋转

5）计算因子得分

**21、试述主成分分析的基本思想。由协方差矩阵出发和由相关系数矩阵出发求主成分有何不同？**

答：主成分分析的基本思想是构造原始变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少量几个新变量并使它们含有足够多的原始变量带有的信息，从而使得用这几个新变量代替原始变量分析问题和解决问题成为可能。一般而言，对于度量单位不同的指标或是取值范围彼此差异非常大的指标，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，由相关

阵出发求解主成分。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。相关阵求得的主成分与协差阵求得的主成分一般情况是不相同的。实际表明，这种差异有时很大。由协方差阵出发求解主成分所得的结果及由相关阵出发求解主成分所得的结果有很大不同，所得主成分解释原始变量方差比例与主成分表达式均有显著差别，且两者之间不存在简单的线性关系。

**22、简述动态聚类法的基本思想和步骤，在实际应用中如何确定合理的聚类数目？**

答：基本思想：首先选择若干个样本作为聚类中心，再按照事先确定的聚类准则进行聚类。在聚类过程中，根据聚类准则对聚类中心反复修改，直到分类合理为止。

步骤：

（1）选择凝聚点，凝聚点就是一批有代表性的样品。可以凭经验选择，或将所有样品随机分成 k 份，计算每一类的均值，将这些均值作为凝聚点；也可以采用最大最小原则或密度法。

（2）初始分类

（3）判断分类是否合理，若不合理，则修改分类，重复步骤（2）

（4）至分类结果合理，结束分类。

聚类数目的真正确定在于研究的问题是什么，以及事先有无一个大致的判断标准。分类的数目应该符合使用的目的。确定聚类数的问题属于聚类有效性问题。比如在模糊聚类分析中，可以根据方差分析理论，应用混合 F 统计量来确定最佳分类数。

**23、简述典型相关分析的基本思想与步骤，试举例说明它的应用。**

答：基本思想：在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数。选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对，如此下去直到两组之间的相关性被提取完毕为止。

步骤：（1）确定典型相关分析的目标

（2）设计典型相关分析

（3）检验典型相关分析的基本假设

（4）估计典型模型，评价模型拟合程度

（5）解释典型变量

（6）验证模型

典型相关分析的用途很广。在实际分析问题中，当我们面临两组多变量数据，并希望研究两组变量之间的关系时，就要用到典型相关分析。 例如，为了研究扩张性财政政策实施以后对宏观经济发展的影响,就需要考察有关财政政策的一系列指标如财政支出总额的增长率、财政赤字增长率、国债发行额的增长率、税率降低率等与经济发展的一系列指标如国内生产总值增长率、就业增长率、物价上涨率等两组变量之间的相关程度。

**24、作因子分析时，如何确定公共因子的个数？如何解释这些公共因子的实际意义？**

答：有 3 个方法可以用来确定因子的个数：

1）方差贡献率

2）设定特征值条件

3）碎石图

公共因子的实际意义，需结合具体问题来定。

**25、主成分分析与因子分析有哪些应用？**

答:主成分分析是构造原始变量的适当线性组合,以产生一系列互不相关的变量,并从中选取少量几个新变量来分析和解决问题,例如高校科研状况评价中的立项课题数与项目经费、经费支出等之间会存在较高的相关性；学生综合评价研究中的专业基础课成绩与专业课成绩、获奖学金次数等之间也会存在较高的相关性。利用主成分分析既可以大大减少参与建模的变量个数,同时也不会造成信息的大量丢失。能够有效降低变量维数。

因子分析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结构出发，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多元统计分析方法。例如，某企业招聘人才，对每位应聘者进行外贸、申请书的形式、专业能力、讨人喜欢的能力、自信心、洞察力、诚信、推销本领、经验、工作态度、抱负、理解能力、潜在能力、实际能力、适应性的 15 个方面考核。这 15 个方面可归结为应聘者的表现力、亲和力、实践经验、专业能力 4 个方面，

每一方面称为一个公告因子。企业可根据这 4 个公共因子的情况来衡量应聘者的综合水平。

**26、距离判别法采用何种距离？这种距离有什么特点？**

答：距离判别法采用马氏距离。

其特点有：

1）两点之间的马氏距离与原始数据的测量单位无关。

2）标准化数据和中心化数据(即原始数据与均值之差）计算出的二点之间的马氏矩离相同。

3）可以排除变量之间的相关性的干扰。

4）满足距离的四个基本公理：非负性、自反性、对称性和三角不等式。

**27、简述多元统计的主要内容，结合你的专业谈谈能用到哪些统计方法。**

答:多元统计分析是从经典统计学中发展起来的一个分支,是一种综合分析方法,它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律。主要内容包括多元正态分布及其抽样分布、多元正态总体的均值向量和协方差阵的假设检验、多元方差分析、直线回归与相关、多元线性回归与相关(Ⅰ)和(Ⅱ)、主成分分析与因子分析、判别分析与聚类分析、对应分析、典型相关分析、Shannon 信息量及其应用。

主成分分析作为多元统计分析的一种方法,作为数据分析和数据挖掘的工具,在遥感图像变化信息提取、遥感图像处理分析、地理要素分析等方面也具有广泛应用。主成分分析可以提取主要信息,使误差出现的机会大大减小。在分析影像数据特征和主成 分变换算法基础上,利用两次主成分变换的方式有效地实现了剔除原始影像中的部分噪声信息的目的,从而提供了一种新的方法实现动态监测变化信息自动发现,经验证此方法能使得数据结果达到应用精度的需求。针对在 TM 假彩色合成的影像中,河流像元和城市像元的混杂现象严重问题, 基于主成分分析和决策树的城市河流信息提取方法,对 TM 信息进行主成分变换,选取前三个分量组成假彩色图像,再分析城市和河流的光谱差别,建立规则,利用建立的规则提取城市中的河流信息,取得了很好结果。

**28、主成分分析和因子分析的十大不同**

1）原理不同

主成分分析基本原理：利用降维（线性变换)的思想，在损失很少信息的前提下把多个指标转化为几个不相关的综合指标(主成分),即每个主成分都是原始变量的线性组合,且各个主成分之间互不相关,使得主成分比原始变量具有某些更优越的性能（主成分必须保留原始变量90%以上的信息)，从而达到简化系统结构，抓住问题实质的目的。

因子分析基本原理：利用降维的思想，由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表示成少数的公共因子和仅对某一个变量有作用的特殊因子线性组合而成。就是要从数据中提取对变量起解释作用的少数公共因子（因子分析是主成分的推广，相对于主成分分析，更倾向于描述原始变量之间的相关关系）

2）线性表示方向不同

因子分析是把变量表示成各公因子的线性组合;而主成分分析中则是把主成分表示成各变量的线性组合。

3）假设条件不同

主成分分析：不需要有假设(assumptions),

因子分析：需要一些假设。因子分析的假设包括：各个共同因子之间不相关，特殊因子（specificfactor）之间也不相关，共同因子和特殊因子之间也不相关。

4）求解方法不同

求解主成分的方法：从协方差阵出发（协方差阵已知)，从相关阵出发（相关阵R已知)，采用的方法只有主成分法。

（实际研究中，总体协方差阵与相关阵是未知的，必须通过样本数据来估计）

注意事项：由协方差阵出发与由相关阵出发求解主成分所得结果不一致时，要恰当的选取某一种方法；一般当变量单位相同或者变量在同一数量等级的情况下，可以直接采用协方差阵进行计算;对于度量单位不同的指标或是取值范围彼此差异非常大的指标，应考虑将数据标准化，再由协方差阵求主成分；实际应用中应该尽可能的避免标准化,因为在标准化的过程中会抹杀一部分原本刻画变量之间离散程度差异的信息。此外,最理想的情况是主成分分析前的变量之间相关性高,且变量之间不存在多重共线性问题(会出现最小特征根接近0的情况)；

求解因子载荷的方法：主成分法，主轴因子法，极大似然法，最小二乘法，a 因子提取法。

5）主成分和因子的变化不同

主成分分析：当给定的协方差矩阵或者相关矩阵的特征值唯一时，主成分一般是固定的独特的；

因子分析：因子不是固定的，可以旋转得到不同的因子。

6）因子数量与主成分的数量

主成分分析：主成分的数量是一定的，一般有几个变量就有几个主成分（只是主成分所解释的信息量不等），实际应用时会根据碎石图提取前几个主要的主成分。

因子分析：因子个数需要分析者指定（SPSS 和 sas 根据一定的条件自动设定，只要是特征值大于 1 的因子主可进入分析），指定的因子数量不同而结果也不同；

7）解释重点不同：

主成分分析：重点在于解释个变量的总方差，

因子分析：则把重点放在解释各变量之间的协方差。

8）算法上的不同：
主成分分析：协方差矩阵的对角元素是变量的方差；

因子分析：所采用的协方差矩阵的对角元素不在是变量的方差，而是和变量对应的共同度（变量方差中被各因子所解释的部分）

9）优点不同：

因子分析：对于因子分析，可以使用旋转技术，使得因子更好的得到解释，因此在解释主成分方面因子分析更占优势；其次因子分析不是对原有变量的取舍，而是根据原始变量的信息进行重新组合，找出影响变量的共同因子，化简数据；

主成分分析：

第一：如果仅仅想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进入后续的分析，则可以使用主成分分析，不过一般情况下也可以使用因子分析；

第二：通过计算综合主成分函数得分，对客观经济现象进行科学评价；

第三：它在应用上侧重于信息贡献影响力综合评价。

第四：应用范围广，主成分分析不要求数据来自正态分布总体，其技术来源是矩阵运算的技术以及矩阵对角化和矩阵的谱分解技术，因而凡是涉及多维度问题，都可以应用主成分降维；

10）应用场景不同：

主成分分析：

可以用于系统运营状态做出评估，一般是将多个指标综合成一个变量，即将多维问题降维至一维，这样才能方便排序评估；

此外还可以应用于经济效益、经济发展水平、经济发展竞争力、生活水平、生活质量的评价研究上；

主成分还可以用于和回归分析相结合，进行主成分回归分析，甚至可以利用主成分分析进行挑选变量，选择少数变量再进行进一步的研究。

一般情况下主成分用于探索性分析，很少单独使用，用主成分来分析数据，可以让我们对数据有一个大致的了解。

几个常用组合：

主成分分析+判别分析，适用于变量多而记录数不多的情况；

主成分分析+多元回归分析，主成分分析可以帮助判断是否存在共线性，并用于处理共线性问题；

主成分分析+聚类分析，不过这种组合因子分析可以更好的发挥优势。

因子分析：

首先，因子分析+多元回归分析，可以利用因子分析解决共线性问题；

其次，可以利用因子分析，寻找变量之间的潜在结构；

再次，因子分析+聚类分析，可以通过因子分析寻找聚类变量，从而简化聚类变量；

此外，因子分析还可以用于内在结构证实

**29、在作判别分析时，如何检验判别效果的优良性？**

当一个判别准则提出以后，还要研究其优良性，即要考察误判概率。一般使用以训练样本为基础的回代估计法与交叉确认估计法。

（1）误判率回代估计法

回判过程中，用 n12 表示将本属于 G1 的样本误判为 G2 的个数，n21 表示将本属于 G2 的样本误判为 G1 的个数，总的误判个数是 n12+n21，误判率的回代估计为(n12+n21)/(n1+n2)，但往往比真实的误判率要小。

（2）误判率的交叉确认估计

每次剔除训练样本中的一个样本，利用其余容量为 n1+n2-1 个训练样本来建立判别准则，再利用所建立的判别准则对删除的那个样本作判别，对训练样本中的每个样本做上述分析，以其误判的比例作为误判概率的估计。

**30、简述费希尔判别法的基本思想。**

从 k 个总体中抽取具有 p 个指标的样品观测数据，借助方差分析的思想构造一个线性判别函数系数：确定的原则是使得总体之间区别最大，而使每个总体内部的离差最小。将新样品的 p 个指标值代入线性判别函数式中求出值，然后根据判别一定的规则，就可以判别新的样品属于哪个总体。

**31、简述费歇尔准则下两类判别分析的基本思想。**

答：费歇尔的判别方法，其基本思想是把 p 个变量 x1,x2,...,xp 综合成一个新变量 y，y=c1x1+c2x2+...+cpxp=c’x ，也即产生一个综合判别指标，要求已知的 g 个类 Gk，k=1,2,...,g 在这个新变量下能最大程度地区分开，于是可用这个综合判别指标判别未知样品的归属。其中 c=(c1,c2,cp)’为待定参数。判别方程除没有常数外，与回归方程非常相似，但两者有着本质的区别。在回归方程中，y 为因变量，是一个已知的随机变量，有其样本测试值，回归分析的任务是选择一组参数，使得根据回归方程预测的因变量的值与实测值尽可能地接近；而判别模型中 y 只是一个综合变量，实际上并不存在这样一个变量，因而也没有实测值。判别模型的几何意义是把 p 维空间的点投影到一维空间（直线）上去，使各已知类在该直线上的投影尽可能分离。

**32、欧式距离与马氏距离的优缺点：**

欧式距离:

优点:简单、易操作、广泛使用

缺点:每个坐标对欧式距离的贡献是同等的，当坐标轴表示测量值时，他们往往带有大小不等的随机波动。当各个分量为不同性质的量时,“距离”的大小与指标的单位有关。

马氏距离:

优点:它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关由标准化数据和中心化数据(即原始数据和均值之差）计算出的两点之间的马氏距离相同，马氏距离可以排除变量之间的相关性的干扰。

缺点:马氏距离建立在总体样本的基础上，否则最终两个样本的马氏距离不同:在计算马氏距离的过程中，要求总体样本数大于样本的维数，否则得到的总体样本协方差矩阵逆矩阵不存在，二维样本在其所处的平面内共线，协方差矩阵逆矩阵也不存在,由此可知协方差矩阵对马氏距离计算的重要性导致了马氏距离的不稳定。在很大程度上，马氏距离夸大了变化微小变量的作用。

**33、聚类分析计算步骤:**

（1）分析所需要研究的问题，确定聚类分析所需要的多元变量

（2）选择对样本聚类还是对指标聚类

（3）选择合适的聚类方法

（4）选择所需的输出结果

**34、主成分分析的基本思想:**

通过对原始变量相关矩阵或协方差矩阵内部结构关系等研究,利用原始变量的线性组合形成几个综合指标(主成分),在保留原始变量主要信息的前提下起到降维和简化问题的作用,使得在研究复杂问题时更容易抓住主要矛盾。利用主成分分析得到的主成分与原始变量之间如下基本关系:

（1） 每一个主成分都是个原始变量的线性组合

（2） 主成分的数目大大少于原始变量的数目

（3） 主成分保留了原始变量绝大多数信息

（4） 各主成分之间互不相关

**35、主成分分析步骤:**

（1）根据研究问题选取初始分析变量

（2）根据初始变量特性判断由协方差阵求主成分还是相关阵求主成分

（3）求协方差阵或相关阵的特征根与相应标准特征向量

（4）判断是否存在明显的多重共线性，若存在，回到第一步

（5）得到主成分的表达式并确定主成分个数，选取主成分

（6）结合主成分对研究问题进行分析并深入研究