主成分分析

1. 主成分分析的基本思想及其作用？

（1）基本思想：主成分分析是设法将原来众多具有一定相关性（比如 P 个指标），重新组合成一组新的互相无关的综合指标来代替原来的指标。

主成分分析，是考察多个变量间相关性一种多元统计方法，研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间互不相关.通常数学上的处理就是将原来 P 个指标作线性组合，作为新的综合指标。

最经典的做法就是用 F1（选取的第一个线性组合，即第一个综合指标）的方差来表达，即 Var(F1)越大，表示 F1 包含的信息越多。因此在所有的线性组合中选取的 F1 应该是方差最大的，故称 F1 为第一主成分。如果第一主成分不足以代表原来 P 个指标的信息，再考虑选取 F2 即选第二个线性组合，为了有效地反映原来信息，F1 已有的信息就不需要再出现在 F2 中，用数学语言表达就是要求 Cov(F1, F2)=0，则称 F2 为第二主成分，依此类推可以构造出第三、第四，……，第 P 个主成分。

（2）作用：主要是降维、简化数据结构。

①主成分分析能降低所研究的数据空间的维数。即用研究 m 维的 Y 空间代替 p 维的 X 空间(m<p)，而低维的 Y 空间代替高维的 x 空间所损失的信息很少。即：使只有一个主成分 Yl(即 m=1)时，这个 Yl 仍是使用全部 X 变量(p 个)得到的。例如要计算 Yl 的均值也得使用全部 x 的均值。在所选的前 m 个主成分中，如果某个 Xi 的系数全部近似于零的话，就可以把这个 Xi 删除，这也是一种删除多余变量的方法。

②有时可通过因子负荷 aij 的结论，弄清 X 变量间的某些关系。

③多维数据的一种图形表示方法。我们知道当维数大于 3 时便不能画出几何图形，多元统计研究的问题大都多于 3 个变量。要把研究的问题用图形表示出来是不可能的。然而，经过主成分分析后，我们可以选取前两个主成分或其中某两个主成分，根据主成分的得分，画出 n 个样品在二维平面上的分布况，由图形可直观地看出各样品在主分量中的地位，进而还可以对样本进行分类处理，可以由图形发现远离大多数样本点的离群点。

④由主成分分析法构造回归模型。即把各主成分作为新自变量代替原来自变量 x 做回归分析。

⑤用主成分分析筛选回归变量。回归变量的选择有着重的实际意义，为了使模型本身易于做结构分析、控制和预报，好从原始变量所构成的子集合中选择最佳变量，构成最佳变量集合。用主成分分析筛选变量，可以用较少的计算量来选择量，获得选择最佳变量子集合的效果。

2. 请阐述主成分分析法的基本步骤。

①将原始数据进行标准化处理

②计算样本相关矩阵 R

③求相关矩阵 R 的特征值与特征向量，并计算贡献率

④选择主成分

⑤对所选主成分做解释

3. 用主成分分析法进行综合评价时，如何构建综合评价函数？

①对所有原始指标进行标准化处理

②进行主成分分析，确定主成分的个数，并计算各主成分得分

$$F_1 = a_{11}x_1 + a_{21}x_2 + \quad + a_{\ 1}x_p$$

③计算综合指标，即主成分综合得分。主成分综合得$\alpha_i = \dfrac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$主成分所对应的方差贡献率×各主成分得分)。 $F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots \alpha_m F_m$

4. 简述主成分分析的应用。

主成分降维之后，再利用新的变量（主成分），进行聚类分析、回归分析、综合评价等等。

①综合评价：利用主成分构建综合评价函数，计算综合得分，比较样本的得分情况。

②主成分回归：将得到的主成分作为新的自变量放入回归模型中，可以解决回归分析中的多重共线性问题。

5. 简述提取样本主成分的原则。

①累计方差贡献率（80%-85%）。②主成分是否好解释

6. 简述主成分分析的适用范围。

适用于变量间有较强相关性的多变量数据。主成分分析法适用于解决人口统计学、数量地理学、分子动力学模拟、数学建模、数理分析等问题

7. 简述量纲对主成分分析的影响及消除方法。

（1）影响：数据标准化后，使每个变量的均值为 0，方差为 1。总体的协方差矩阵与总体的相关系数相等。

（2）消除方法：①标准化；②均值化；③功效系数法；④0-1

因子分析

1. 简述因子分析的基本思想。

因子分析是根据相关矩阵内部的依赖关系，把一些具有错综复杂关系的变量综合为数量较少的几个因子。通过不同因子来分析决定某些变量的本质及其分类的一种统计方法。

简单地说，就是根据相关性大小把变量分组，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。每组变量代表一个基本结构，这个基本结构称为因子。

通过因子分析，能减少变量的数量，将大批量的变量减少为少量的基本特性，通过因子分析，数据的基础结构被抽象出来。

可以利用因子分析得出的结果，进行其他的统计分析。比如：

回归分析：寻找关键的驱动因素

聚类分析：把目标分类为某些特征更加相似的细分群体

因子分析是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。

2. 简述因子载荷矩阵的含义、统计特征及其意义。

含义：

因子载荷矩阵的估计是因子分析的主要问题之一。可以从两方面来理解因子载荷矩阵的含义：

（1）可以将其看做是对因子进行线性组合时的系数。

（2）因子载荷矩阵 L 还可以看作是是 p 维空间的一组单位正交向量。

统计特征及其意义（ppt 上）：

1、因子载荷 aij 的统计意义

因子载荷a_{ij}是第 i 个变量与第 j 个公共因子的相关系数，

$$\operatorname{cov}(X_i,F_j)=\operatorname{cov}(\sum_{k=1}^{m}a_{ik}F_k+\varepsilon_i,F_j)=\operatorname{cov}(\sum_{k=1}^{m}a_{ik}F_k,F_j)+\operatorname{cov}(\varepsilon_i,F_j)$$

$$=a_{ij}$$

根据公共因子的模型性质，有 $\gamma_{x_iF_j}=a_{ij}$

模型为$X_i=a_{i1}F_1+\cdots+a_{im}F_m+\varepsilon_i$

（载荷矩阵中第 i 行，第 j 列的元素）反映了第 i 个变量与第 j 个公共因子的相关性。绝对值越大，相关的密切程度越高。

因子载荷不是惟一的

2、变量共同度的统计意义

统计意义：

$$X_i=a_{i1}F_1+\cdots+a_{im}F_m+\varepsilon_i$$ **两边求方差**

$$Var(X_i)=a_{i1}^2Var(F_1)+\cdots+a_{im}^2Var(F_m)+Var(\varepsilon_i)$$

$$1=\sum_{j=1}^{m}a_{ij}^2+\sigma_i^2$$

所有的公共因子和特殊因子对变量X_i的贡献为1。如果$\sum_{j=1}^{m}a_{ij}^2$非常靠近1，σ_i^2非常小，则因子分析的效果好，从原变量空间到公共因子空间的转化性质好。

定义：变量X_i的共同度是因子载荷矩阵的第 i 行的元素的平方和。记为$h_i^2=\sum_{j=1}^{m}a_{ij}^2$。

3、公共因子F_j方差贡献的统计意义

因子载荷矩阵中各列元素的平方和

$$g_j^2=\sum_{i=1}^{p}a_{ij}^2$$

称为$F_j\ (j=1,\cdots,m)$ 对所有的X_i的方差贡献和。衡量F_j的相对重要性。

3. 比较因子分析和主成分分析，说明它们的相似和不同之处。

(1) 相似:

①因子分析是主成分分析的推广和延伸,二者都是利用降维的思想;

②主成分(或因子)的个数远小于原始变量的个数;

③主成分(或因子)保留了原始变量的绝大部分信息;

④各主成分(或因子)之间互不相关。

(2) 不同:

1 原理不同

主成分分析基本原理：利用降维的思想，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,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，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；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。

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，实际应用时会根据碎石阁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，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：重点在于解释各变量的总方差，因子分析：则把重点放在解释各变量之问的协方差。

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）

4. 因子模型和回归模型相比较有何异同？

相同：两种模型的联系在于都是线性的。因子分析的过程就是一种线性变换。

不同：

因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义。

因子分析模型是一种通过显在变量测评潜在变量，通过具体指标测评抽象因子的统计分析方法的模型。而线性回归模型回归分析的目的是设法找出变量间的依存(数量)关系, 用函数关系式表达出来。

因子分析模型中每一个变量都可以表示成公共因子的线性函数与特殊因子之和。即

$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i$，（$i = 1,2,\cdots,p$） 该模型可用矩阵表示为：$\mathbf{X} = \mathbf{AF} + \boldsymbol{\varepsilon}$

而回归分析模型中多元线性回归方程模型为：$y_1 = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e_i$ 其中b_0是常数项，$b_1, b_2 \ldots b_n$是偏回归系数，e_i是残差。

因子模型满足：

（1） $m \le p$；　　（2） $Cov(\mathbf{F}, \boldsymbol{\varepsilon}) = 0$，即公共因子与特殊因子是不相关的；

（3） $\mathbf{D}_F = D(\mathbf{F}) = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & \ddots & \\ 0 & & 1 \end{bmatrix} = \mathbf{I}_m$，即各个公共因子不相关且方差为 1；

（4） $\mathbf{D}_\varepsilon = D(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & \ddots & \\ 0 & & \sigma_p^2 \end{bmatrix}$，即各个特殊因子不相关，方差不要求相等。

而回归分析模型满足

（1）正态性：随机误差（即残差）e 服从均值为 0，方差为σ^2的正态分布；

（2）等方差：对于所有的自变量 x，残差 e 的条件方差为σ^2 ，且σ为常数；

（3）独立性：在给定自变量 x 的条件下，残差 e 的条件期望值为 0（本假设又称零均值假设）；

（4）无自相关性：各随机误差项 e 互不相关。

5. 因子分析中对因子载荷矩阵进行旋转的目的是什么？常用的旋转方法有哪些？

正交旋转：由因子载荷矩阵 A 右乘一正交阵而得到，经过旋转后的新的公因子仍然保持彼此独立的性质。

1、方差最大法

方差最大法从简化因子载荷矩阵的每一列出发，使和每个因子有关的载荷值平方的方差最大。当只有少数几个变量在某个因子上有较高的载荷值时，对因子的解释最简单。方差最大的直观意义是希望通过因子旋转后，使每个因子上的载荷值尽量拉开距离，一部分的载荷趋于 1，另一部分趋于 0。

2、四次方最大旋转

四次方最大旋转是从简化载荷矩阵的行出发，通过旋转初始因子，使每个变量只在一个因子上有较高的载荷，而在其它的因子上尽可能低的载荷。如果每个变量只在一个因子上有非零的载荷，这时的因子解释是最简单的。四次方最大法通过使因子载荷矩阵中每一行的因子载荷平方的方差达到最大。

斜交旋转：放弃了因子之间彼此独立这个限制，可达到更简洁的形式，实际意义也更容易解释。

不论是正交旋转还是斜交旋转，都应该在因子旋转后，使每个因子上的载荷尽可能拉开

距离，一部分趋近 1，一部分趋近 0，使各个因子的实际意义能更清楚地表现出来。

6. 阐述运用因子分析进行综合评价时，综合评价函数的构造方法。

通过因子分析，取 m 个公共因子 $F_1, F_2, \cdots, F_m$，以每个公共因子 F_j 的方差贡献率 $\alpha_j = \frac{g_j^2}{p}$，为权，构造综合评价函数 $F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots \alpha_m F_m$。按 F 值的大小对样品进行排序比较或分类。

7. 阐述主成分分析和因子分析用于对变量降维时，两种方法在基本思想和做法上的差异。

基本思想：

主成分分析基本原理：利用降维的思想，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,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，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)

做法：

1 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。

2 假设条件不同

主成分分析不需要有假设；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，特殊因子之间也不相关，共同因子和特殊因子之间也不相关。

3 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。

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。

4 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，实际应用时会根据碎石阁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，指定的因子数量不同而结果也不同；

5 解释重点不同：

主成分分析：重点在于解释各变量的总方差，因子分析：则把重点放在解释各变量之问的协方差。

6 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）

典型相关分析

1. 典型相关分析的基本思想及其应用

（1）典型相关分析的概念：典型相关分析是对互协方差矩阵的一种理解，是利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法。

（2）典型相关分析的基本思想：首先在每组变量中找出变量的线性组合，使其具有最大相关性，然后再在每组变量中找出第二对线性组合，使其分别与第一对线性组合不相关，而第二对本身具有最大的相关性，如此继续下去，直到两组变量之间的相关性被提取完毕为止.有了这样线性组合的最大相关，则讨论两组变量之间的相关，就转化为只研究这些线性组合的最大相关，从而减少研究变量的个数。

（3）典型相关分析的应用：典型相关分析的用途很广。在实际分析问题中，当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。

又如，为了研究宏观经济走势与股票市场走势之间的关系，就需要考察各种宏观经济指标如经济增长率、失业率、物价指数、进出口增长率等与各种反映股票市场状况的指标如股票价格指数、股票市场融资金额等两组变量之间的相关关系。再如，工厂要考察所使用的原料的质量对所生产的产品的质量的影响，就需要对所生产产品的各种质量指标与所使用的原料的各种质量指标之间的相关关系进行测度。

又如，在分析评估某种经济投入与产出系统时，研究投入和产出情况之间的联系时，投入情况面可以从人力、物力等多个方面反映，产出情况也可以从产值、利税等方面反映 。

再如在分析影响居民消费因素时，我们可以将劳动者报酬、家庭经营收入、转移性收入等变量构成反映居民收入的变量组，而将食品支出、医疗保健支出、交通和通讯支出等变量构成反映居民支出情况的变量组，然后通过研究两变量组之间关系来分析影响居民消费因素情况。

2. 典型相关分析与相关分析有何异同点

（1）典型相关分析：典型相关分析是对互协方差矩阵的一种理解，是利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法。

（2）相关分析：相关分析是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。例如，人的身高和体重之间；空气中的相对湿度与降雨量之间的相关关系都是相关分析研究的问题。

（3）联系：典型相关分析与相关分析都是分析变量之间相关性的分析方法，都是线性分析的范畴。

（4）区别：简单相关系数描述两组变量的相关关系的缺点:只是孤立考虑单个 X 与单个 Y 间的相关，没有考虑 X、Y 变量组内部各变量间的相关。两组间有许多简单相关系数，使问题显得复杂，难以从整体描述。典型相关是简单相关、多重相关的推广。典型相关是研究两组变量之间相关性的一种统计分析方法。也是一种降维技术。

3. 什么是典型变量？它具有哪些性质？

（1）典型变量的概念：在典型相关分析中，在一定条件下选取原始两组变量的系列线性组合配对以反映两组变量之间的线性关系，被选出的线性组合配对称为典型变量。

（2）典型变量的性质：

性质一：典型变量之间的相关性

设 $Vk=ak'X, Wk=bk'Y$，令 $V=(V1,\cdots,Vp)'$，$W=(W1,\cdots,Wp)'$，那么

D[VW]=[Ip Λ Λ Ip]
这说明：
Vi 和 Vj 不相关，Wi 和 Wj 不相关，在 i ≠ j 时
Vi 和 Wi 的方差都为 1
Vi 和 Wj 不相关，在 i ≠ j 时，这主要考虑考虑到 bj 可以由 aj 表示，所以 Wj 相当于 Vj
Vi 和 Wi 的相关系数为 λi
（2）性质二：原始变量和典型变量的相关性
回顾一下，典型变量可以表示为
V=(V1,⋯,Vp)′ =(a1′ X,⋯,ap′ X)′ =A′ XW=(W1,⋯,Wp)′ =(b1′ Y,⋯,bp′ Y)′ =B′ Y
由于期望都是 0，所以协方差为第一个变量乘上第二个变量的转置再求期望，可得
Cov(X,V)= Σ11A
Cov(X,W)= Σ12B
Cov(Y,V)= Σ21A
Cov(Y,W)= Σ22B
如果原始变量已经标准化过了，那么协方差阵就是相关系数矩阵。并且这四个矩阵将在之后用于冗余分析。
（3）性质三：线性变换不变性
对原始变量 X 和 Y 作线性变换得到
X*=C′ X+d,Y*=G′ Y+h
因为平移不会影响到相关性，其第 i 对典型相关变量的投影方向为

(ai)*=C−1ai,bi*=G−1bi(i=1,⋯,p)

也就是说，做线性变换后再最佳投影，与不做线性变换再最佳投影，得到相同的变量（无视掉常数部分）。
因为投影后的变量都相同了（无视常数），自然最大相关系数也不改变。

4. 简述典型相关分析中冗余分析的内容与作用

（1）冗余分析的内容：冗余分析是通过原始变量与典型变量之间的相关性。分析引起原始变量变异的原因。以原始变量为因变量，以典型变量为自变量，建立线性回归模型，则相应的确定系数（判定系数 R2）等于因变量与典型变量间的相关系数的平方，它描述了由于因变量与典型变量的线性关系引起的因变量变异在因变量的总变异中比例。
（2）冗余分析的作用：分析每组变量提取出的典型变量所能解释的该组样本总方差的比例，从而定量测度典型变量所包含的原始信息量

聚类分析

1. 如何测度样品和变量间的相似性？计算样品之间的距离有哪些公式？它们各有什么特点？

1. 测度样品之间的亲疏程度。将每一个样品看作 p 维空间的一个点，并用某种度量测量点与点之间的距离，距离较近的归为一类，距离较远的点应属于不同的类。
2. （1）明氏距离

设原始数据为 $\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$

令 d_{ij} 表示样品 x_i 与 x_j 的距离

$$d_{ij} = (\sum_{l=1}^{p} | x_{il} - x_{jl} |^{k})^{\frac{1}{k}}$$

当 k=1 时，为绝对值距离。

当 k=2 时，为欧氏距离　　特点：欧氏距离有明确的空间距离概念

当 k=∞时，为切比雪夫距离

$$d_{ij} = \max_{1 \le l \le p} |x_{il} - x_{jl}|$$

明氏距离缺点：①明氏距离的数值与指标的量纲有关 ②没有考虑各个变量之间相关性的影响

（2）马氏距离

$$d_{ij} = \left[\left(x_{i1} - x_{j1}, x_{i2} - x_{j2}, \cdots, x_{ip} - x_{jp} \right) S^{-1} \begin{pmatrix} x_{i1} - x_{j1} \\ x_{i2} - x_{j2} \\ \vdots \\ x_{ip} - x_{jp} \end{pmatrix} \right]^{\frac{1}{2}}$$

$$= \left[\left(x_i - x_j \right)' S^{-1} \left(x_i - x_j \right) \right]^{\frac{1}{2}}$$

特点：1.马氏距离又称为广义欧氏距离。 2.马氏距离考虑了观测变量之间的相关性。如果假定各变量之间相互独立，即观测变量的协方差矩阵是对角矩阵，此时马氏距离就是标准化的欧氏距离。 3.马氏距离不受指标量纲及指标间相关性的影响

（3）兰氏 Canberra 距离

特点：1.不受量纲影响 2.对大的奇异值不敏感，特别适合于高度偏倚的数据 3.没有考虑指标之间的相关性

$$d_{ij} = \frac{1}{p} \sum_{l=1}^{p} \frac{|x_{il} - x_{jl}|}{x_{il} + x_{jl}}$$

2. Q 型聚类法和 R 型聚类法有什么异同？

Q 型聚类法：是对样品进行分类，即从实际问题中观测得到了 n 个样品，根据某相似性原则，对 n 个样品进行归类。通过 Q 型分析可以综合利用多个变量的信息 对样品进行分类，分类结果直观，相比其他传统分类方法更细致全面合理。

R 型聚类法：是对变量进行分类，可以了解变量之间的亲疏程度，根据变量的分

类结果以及它们之间的关系，选择主要变量进行回归分析或 Q 型聚类。

3. 简述系统聚类法的基本思想及主要步骤。

1. 基本思想：先将 n 个样品各自看成一类，然后规定样品之间的“距离”和类与类之间的距离（开始时，由于每个样品自成一类，所以类与类之间的距离和样品之 间的距离是“相等”的）。选择距离最小的两类合并成一个新类，计算新类和其他类（各当前类）的距离，再将距离最近的两类合并。这样，每次合并少一个类， 直至所有的样品都归为一类为止。

2. 具体步骤：

（1）计算 n 个样品两两的距离；

（2）构造 n 个类，每个类只含有一个样品。

（3）合并距离最近的两类为一个新类。

（4）计算新类与各当前类的距离。

（5）重复（3）、（4），合并距离最近的两类为新类，直到所有的类并为一类为止。 （6）画聚类谱系图。

（7）决定类的个数和类。

4. 简述系统聚类分析的优缺点。

优点：通过谱系图直接指出由粗到细的多种分类情况，分类方法选择后，分类结果稳定

缺点：计算量大，尤其遇到研究样品多时，计算距离矩阵和绘制谱系图十分复杂

5. Q 型系统聚类法包括哪几种方法，有什么特点。

一共六种方法，归类的基本步骤一致，只是类与类之间的距离有不同的定义

(1)最短距离法：类与类之间的距离等于两类最近样品之间的距离；

(2)最长距离法：类与类之间的距离等于两类最远样品之间的距离：

(3)中间距离法：最长距离法夸大了类间距离，最短距离法低估了类间距离介于两者问的距离法即为中间距离法，类与类之问的距离既不采用两类之间最近距离。也不采用最远距离，而是采用介于最远和最近之间的距离；

(4) 重心法：类与类之间的距离定义为对应这两类重心之间的距离对样品分类来说，每一类的类重心就是该类样品的均值；

(5) 类平均法：类与类之问的距离等于各类元素两两之间的平方距离的平均；

(6)离差平方和法(Ward 法)：基于方差分析的思想，如果分类正确，同类样品之间的离差平方和应当较小，类与类之间的离差平方和应当较大。

6. 简述动态聚类法的基本思想与步骤。

基本思想：选取若干个样品作为凝聚点，计算每个样品和凝聚点的距离，进行初始分类，再根据分类计算其重心，再进行第二次分类，直到所有样品不再调整为止

基本步骤：

（1）选择凝聚点（主管判别法、重心法、均差法、密度法）

（2）初始分类。即对于取定的凝聚点，视每个凝聚点为一类，将每个样品根据定义的距离向最近的凝聚点归类

（3）修改分类 得到初始分类，计算重心 以重心作为新的凝聚点，再分类，直到不再调整为止，说明分类已经合理

优点：计算量小，方法简便，可以根据经验，先做主观分类。

缺点：结果受选择凝聚点好坏的影响，分类结果不稳定。