**《多元统计分析》期末考试范围**

**试题分值：**

一、单项选择题（共5小题，每小题2分，共10分）

二、判断题（共10小题，每小题2分，共20分）

三、计算题（5大题，15、15、10、10、10）

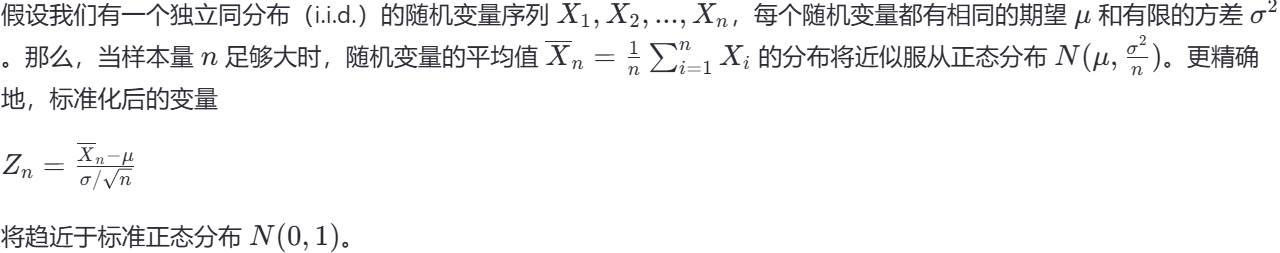
四、简答题（1题，10分）

**考点：**

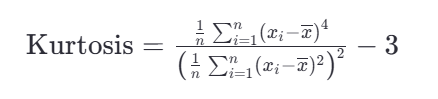
1. 随机变量的概念及相关知识点（比如还有：线性相关、相互独立、大数定律、中心极限定理、峰度、偏度、分位数）
   1. 随机变量是带有概率的变量。

随机变量分为两大类：离散随机变量、连续随机变量 。

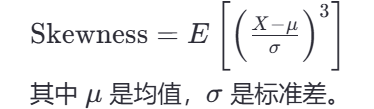
* 1. 线性相关：指两个或多个变量之间的一种关系，其中一个变量的变化可以大致通过另一个变量的线性变化来预测
  2. 相互独立：协方差等于0
  3. 大数定律：当试验次数足够大时，随机事件的平均结果将趋于稳定，并接近于其理论上的期望值
  4. 中心极限定理：在一定条件下，大量独立同分布随机变量的和的分布将趋向于正态分布（高斯分布）

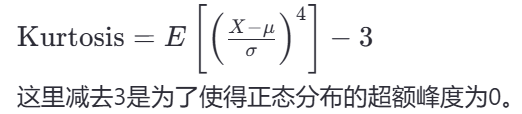


* 1. 峰度：峰度是统计学中用于描述概率分布形状的一个量度，具体来说，它衡量的是数据分布的尾部厚度和峰的尖锐程度



* 1. 偏度：用于衡量数据分布的不对称程度
  2. 分位数：用于描述数据分布特征，它将一组数值按照一定的比例分割成几个等份，最常见的分位数包括四分位数、百分位数和十分位数等。

1. 一元随机变量、样本的均值、随机变量的概念、方差、协方差、相关系数等概念及计算公式
   1. 期望（均值）
   2. 方差
   3. 协方差（单个变量与自己的协方差即方差）
   4. 标准差
   5. 偏度
   6. 峰度



1. 多元随机变量的基本概念、表示方法；多元随机变量的均值向量、协方差矩阵、相关系数矩阵的相关知识点；
   1. 多元随机变量的基本概念：是概率论和统计学中处理多个随机变量之间关系的数学模型。它们用于描述两个或更多随机变量的联合行为，以及这些变量之间的相互依赖性。
   2. 表示方法：向量表示法
   3. 均值向量
   4. 协方差矩阵
   5. 相关系数矩阵：对角线上的元素都是1，其他是两元素之间的相关系数
2. 多元随机变量的线性组合的基本概念与简单计算（比如给定多元随机变量，会计算其线性组合的均值、协方差矩阵以及判断是否线性相关等等）。

题型：

①给定随机向量（多个随机变量组成的集合）X，通过线性组合，用X把Y表达出来

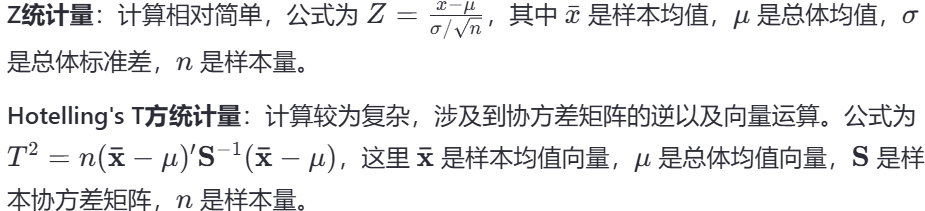
②计算线性组合的期望（均值）

③求线性组合的协方差矩阵S

④求线性组合的相关系数矩阵R

⑤广义方差和总方差：广义方差是协方差矩阵的行列式、总方差是指在一个多维数据集中，将各个维度（特征）的方差相加得到的总和

1. 统计假设检验的原理思想以及基本知识点（第一类错误、第二类错误、显著性水平、置信度）
   1. 单样本均值向量μ假设检验，方差已知，Z统计量；方差未知，t^2统计量

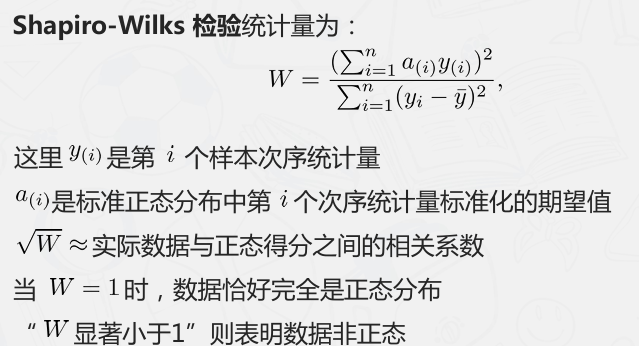


* 1. 一元两样本均值向量检验：
     1. 独立样本：群体1和群体2是独立的
     2. 成对样本：群体1和群体2是相关的
  2. 第一类错误：当原假设H0实际上是正确的，我们却错误地拒绝了它，即“弃真”。

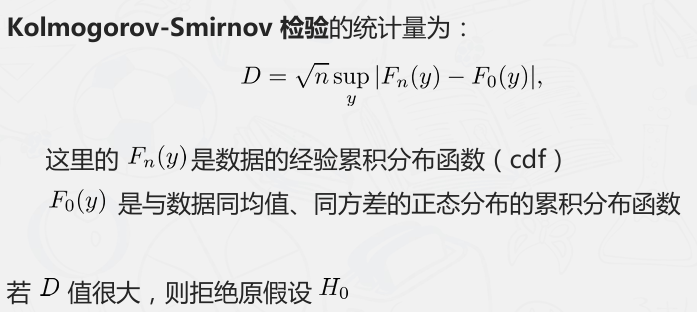
第二类错误：当原假设H0实际上是错误的，我们却没有拒绝它，即“取伪”。

* 1. 显著性水平：指的是在零假设（H0）为真时，错误地拒绝零假设的概率的最大可接受值，常见α=0.05
  2. 置信度：1-显著性水平，反映了我们对某个估计值或区间估计能够包含总体参数真实值的信心程度

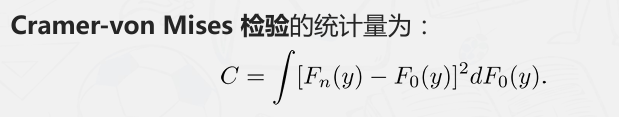
1. 基本了解一元正态性检验的方法和相关原理（SW检验、KS检验、CM检验、AD检验，PPT都有）
   1. SW检验



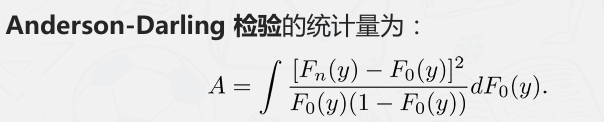
* 1. KS检验



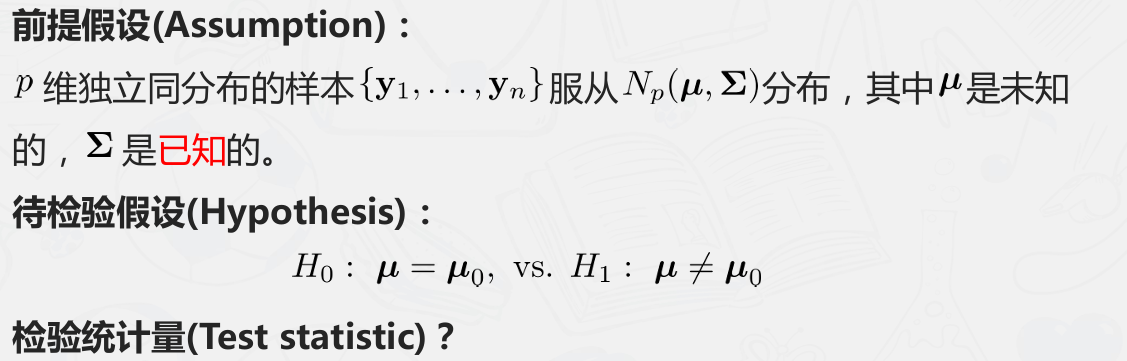
* 1. CM检验

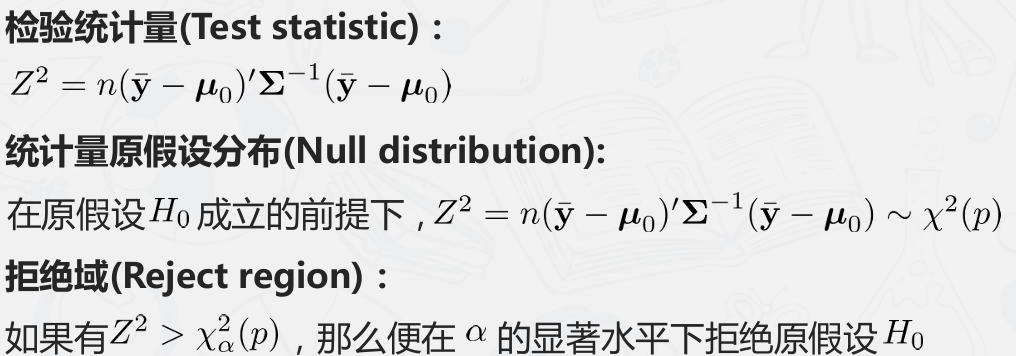


* 1. AD检验

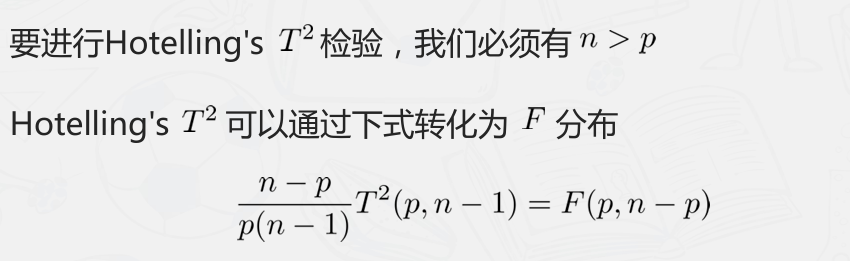
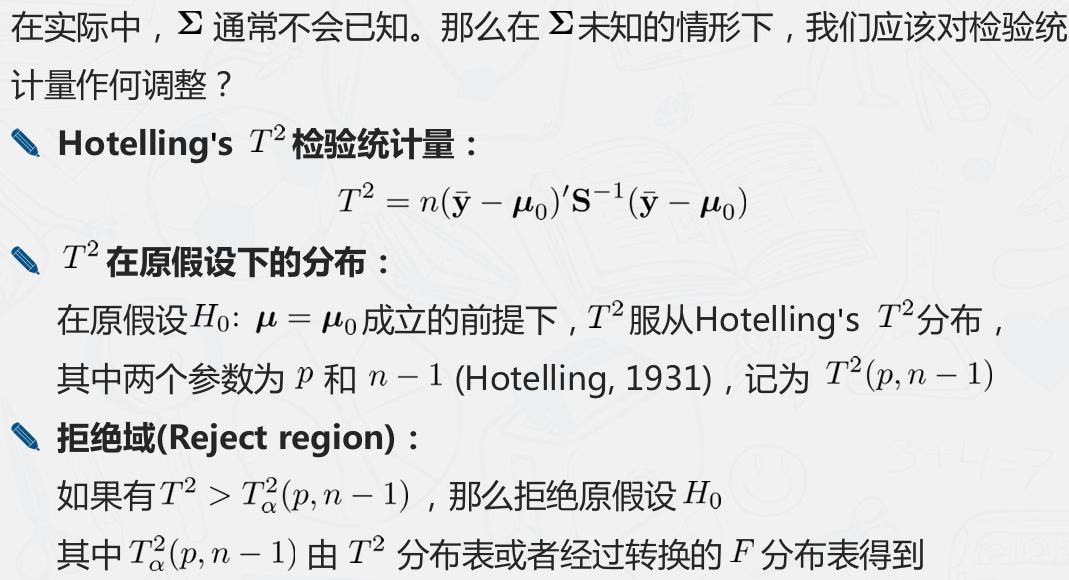


1. 对于多元随机变量的均值向量的假设检验而言，分维度各自开展一元假设检验和开展多元假设检验会有什么结果。
   1. 一元检验完全忽略了变量之间的相关性
   2. 一元检验会使整体第一类错误的概率增大
   3. 一元检验有时会降低检验功效：小的个体效应可能累积成为显著的联合效应
2. 多元随机变量均值向量检验在协方差矩阵∑已知和未知情况下，独立样本、两样本情况下，分别应构造什么统计量并且其服从什么分布？
   1. 单样本均值向量μ的检验，∑已知情况下

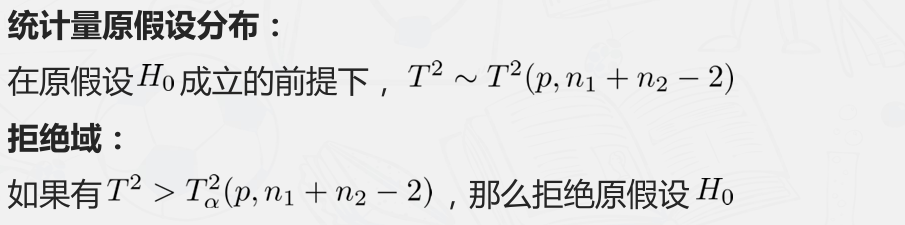
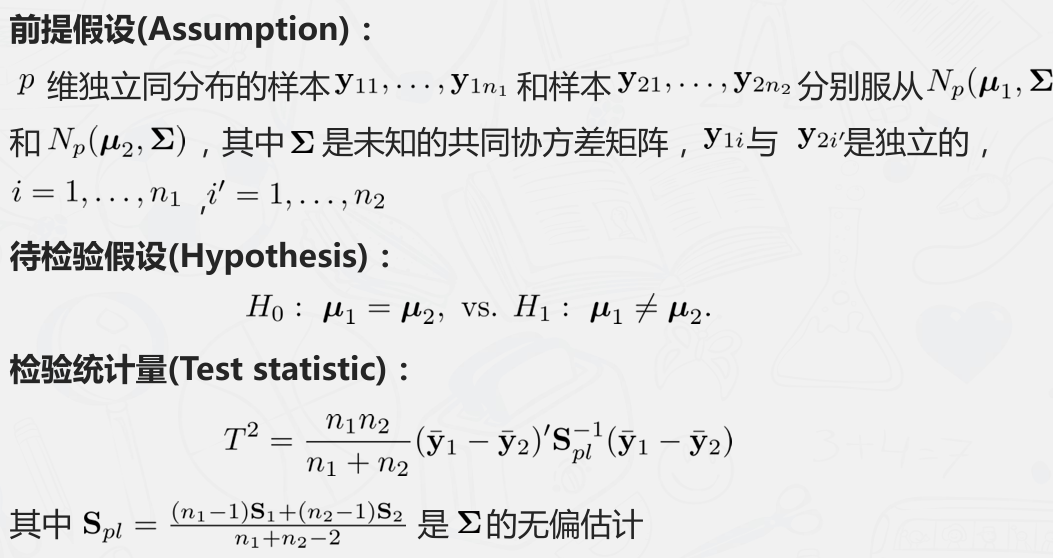


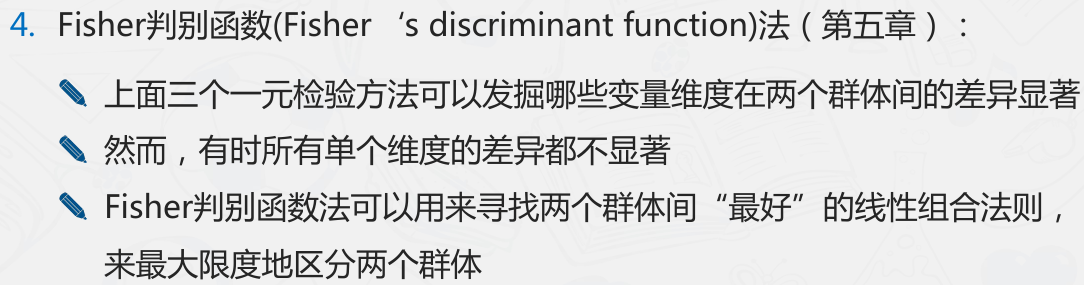


* 1. 单样本均值向量μ的检验，∑未知情况下

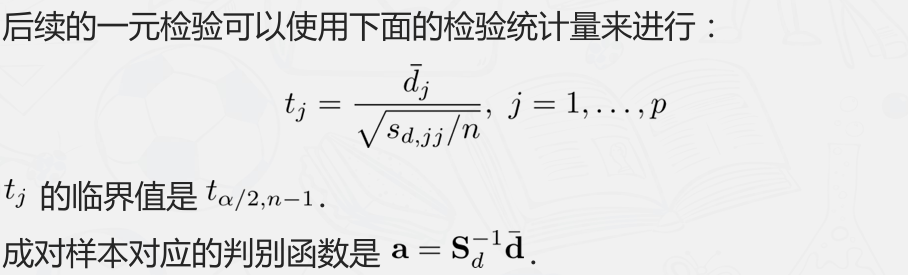
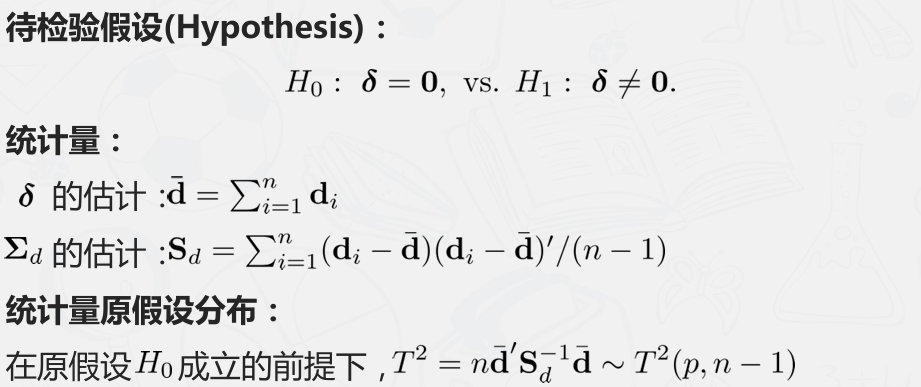
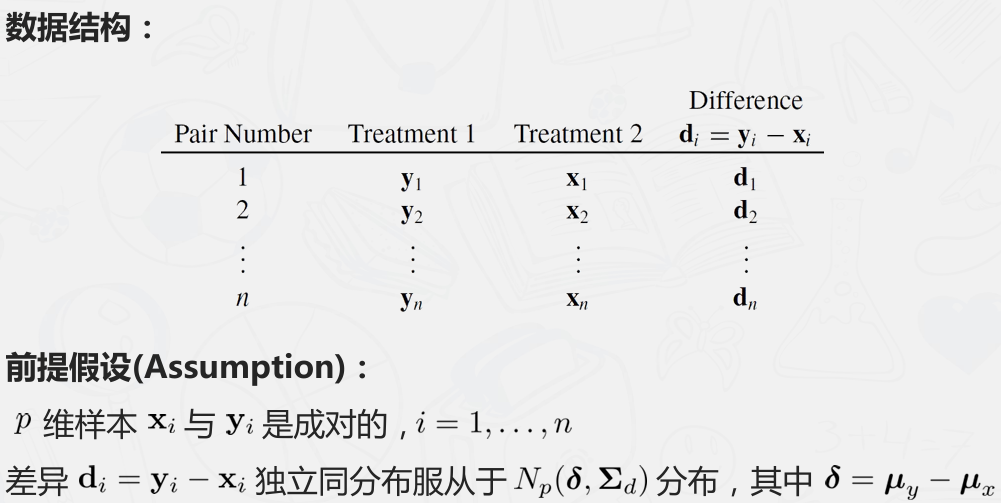


* 1. 多元两样本检验，独立样本

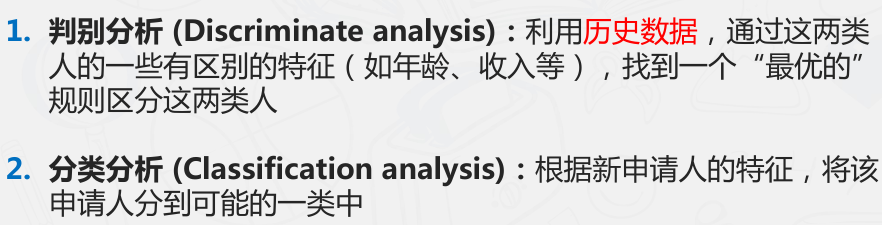




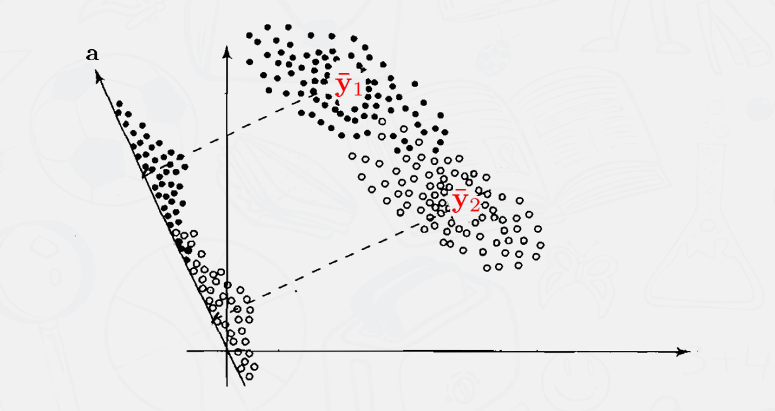
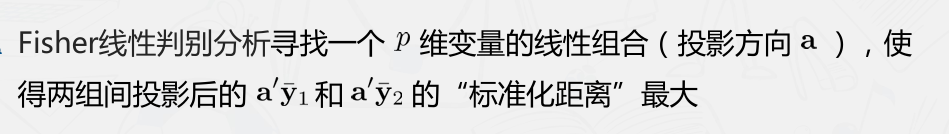
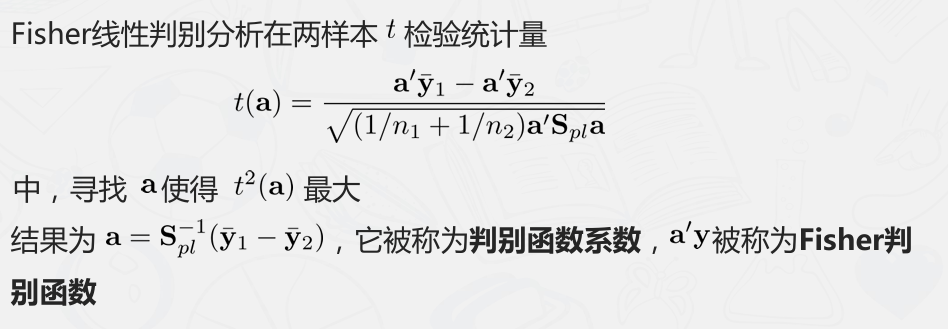
* 1. 多元两样本检验，成对样本



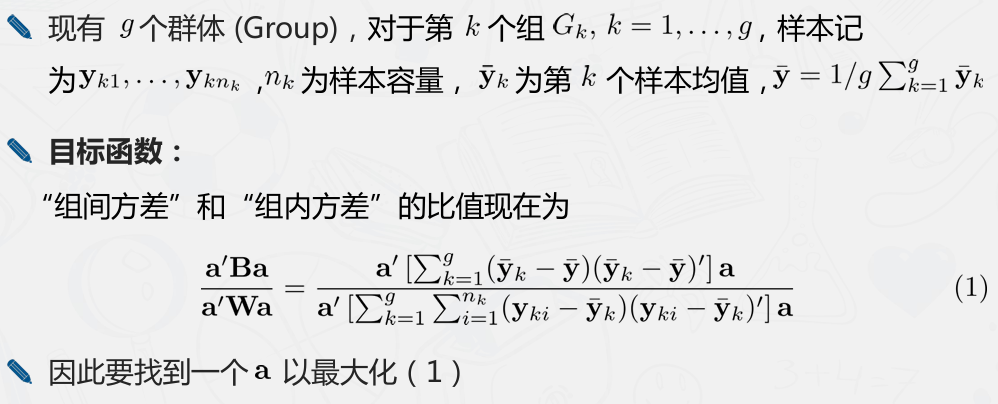
1. 判别分析和分类分析的基本概念和基本原理（两群体Fisher线性判别分析、多群体Fisher线性判别分析、两群体Fisher分类、两群体贝叶斯分类、多群体分类）。
   1. 判别分析和分类分析的基本概念和基本原理

判别分析在寻找一种“分类规则”：利用变量的函数（判别函数）来 描述或者解释两组或多组群体之间的区别。 分类分析则更偏向于给出“分类结果”：预测一个新观测对象的类别， 或者说是将其分配到一个类别，利用一些规则（分类函数）评估新观 测对象的测量值向量，找到该对象最有可能属于的类别。判别分析和分类分析中，所有个体都有既定标签，即有“标准答案” 可供监督，故其属于有监督学习

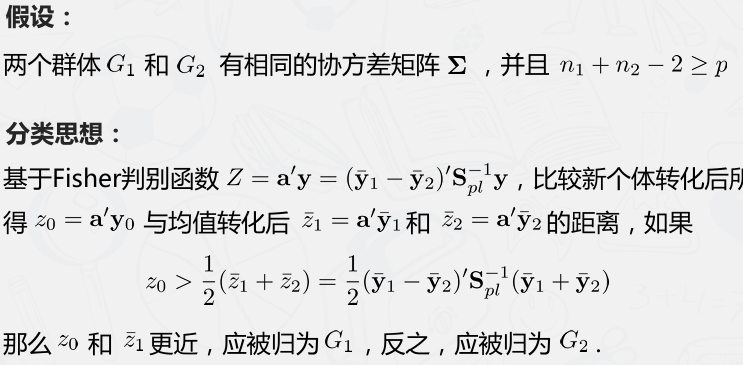
* 1. 两群体Fisher线性判别分析

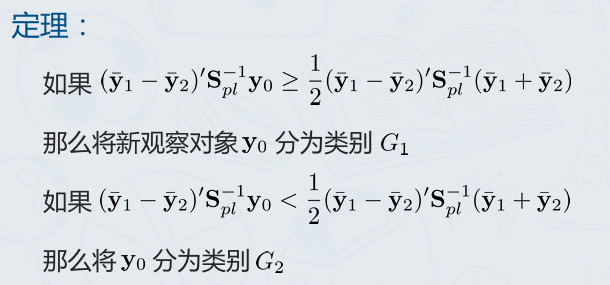
   

* 1. 多群体Fisher线性判别分析

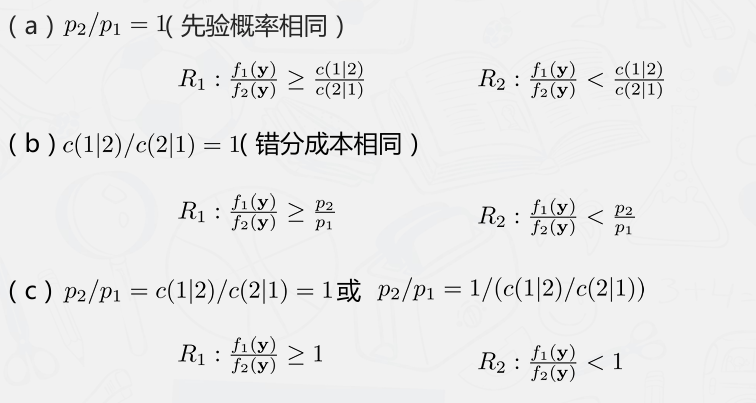
* 1. 两群体Fisher分类



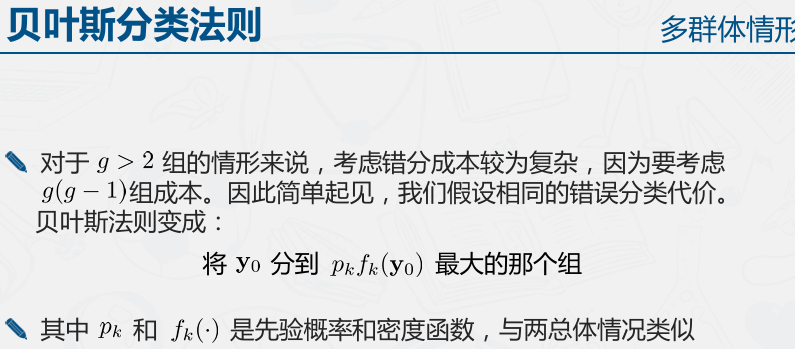


* 1. 两群体贝叶斯分类

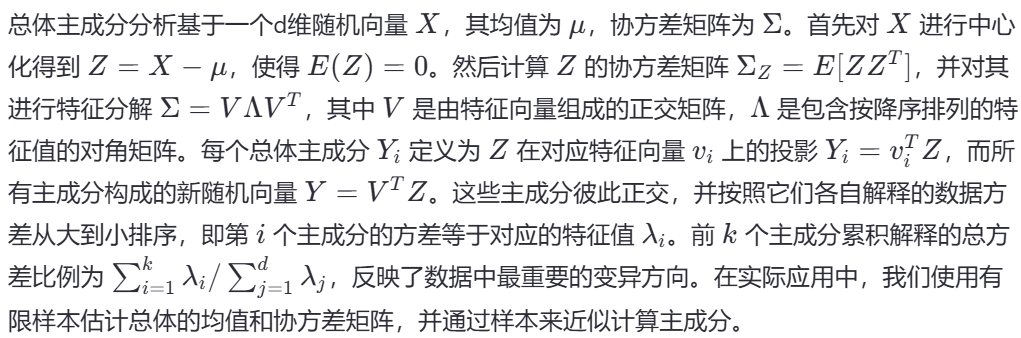




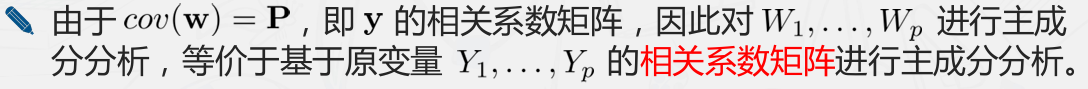
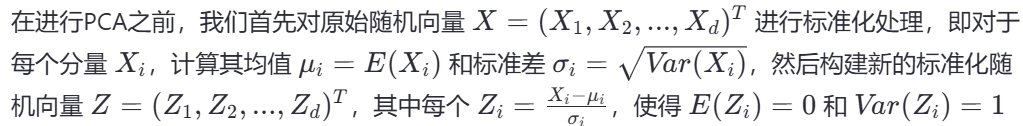
* 1. 贝叶斯多群体分类



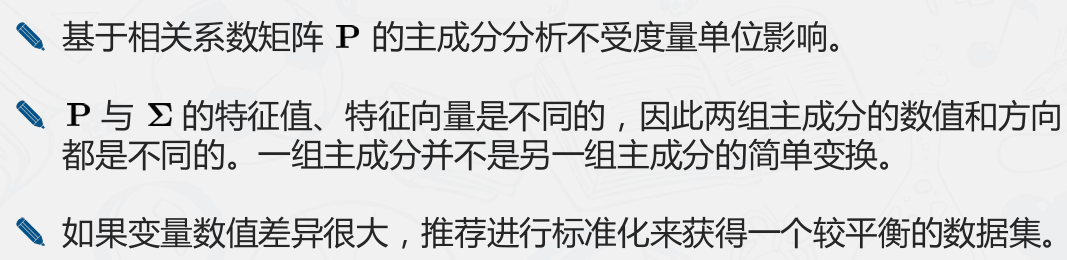
1. 主成分分析的目的、基本原理和概念（方差贡献度、得到的每个主成分怎么解释）、总体主成分推导、基于标准化变量的总体主成分分析（用协方差矩阵和相关系数矩阵得到的主成分有什么差别）、样本主成分推导、主成分个数的选择。
   1. 主成分分析的目的：数据降维、特征提取
   2. 基本原理：主成分分析（PCA）的基本原理是通过线性变换将原始数据转换到一个新的坐标系统中，使得在新的坐标系统中的每一维都是彼此正交的，并且这些维度按照其能够解释的数据方差大小排序。
   3. 方差贡献度：每个主成分的方差贡献度是指该主成分解释的总方差的比例。所有主成分的方差贡献度之和等于100%，因为它们一起解释了原始数据的全部方差。通常，我们会选择那些累积贡献度达到一定阈值（如85%-95%）的主成分作为分析结果，以确保大部分信息被保留。
   4. 总体主成分推导：



* 1. 基于标准化变量的主成分推导



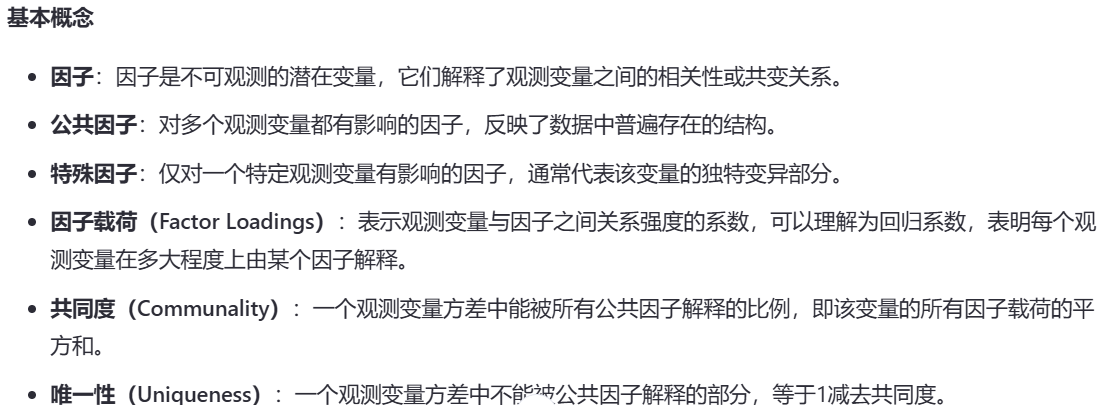
* 1. 用协方差矩阵和相关系数矩阵得到的主成分有什么差别



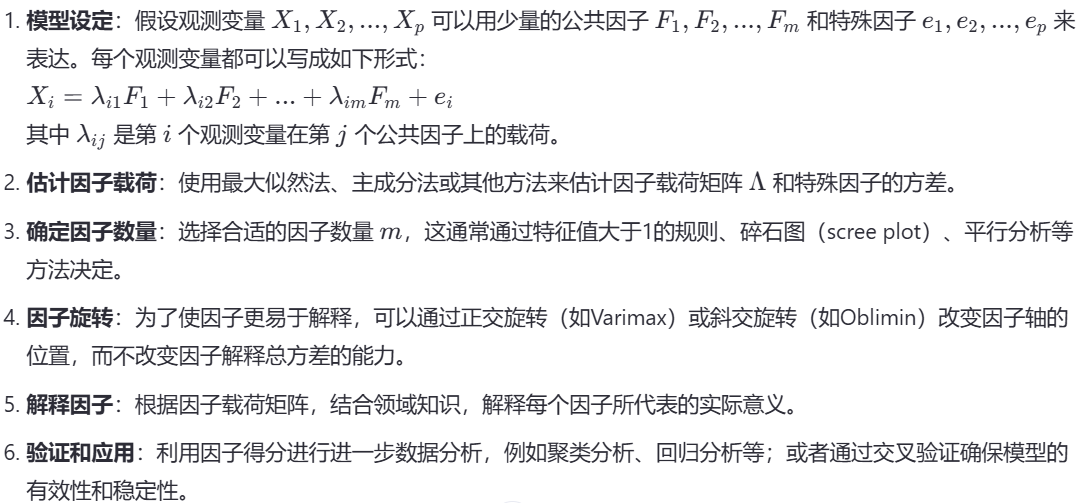
使用协方差矩阵进行主成分分析（PCA）时，结果会受到变量尺度的影响，具有较大数值范围或方差的变量会在主成分中占据更大权重，反映那些高方差特征的信息。而使用相关系数矩阵进行PCA前，每个变量会被标准化（均值为0，标准差为1），因此PCA结果不会受变量尺度影响，而是专注于变量间的线性相关性，更均衡地捕捉不同变量间的关系模式。

1. 因子分析的基本概念和原理、什么是正交因子模型（基本假设条件和相关推导）、载荷估计方法（主成分法和其他估计方法）、什么是估计因子得分、因子旋转的目的和原理（正交、斜交旋转）；主成分法和主成分分析叫法相似但有什么不同？主成分分析和因子分析有什么差异？因子的载荷所表示的意义、因子的方差贡献等。
   1. 因子分析的基本概念和原理：

基本概念：

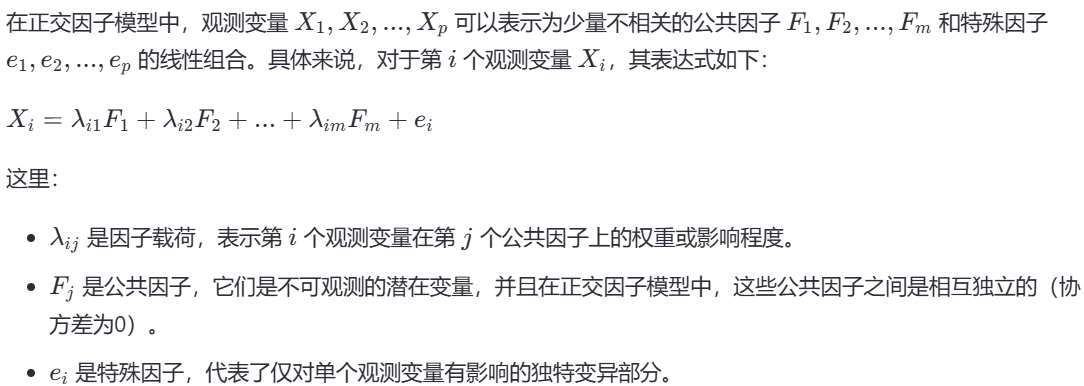


原理：



* 1. 什么是正交因子模型？

正交因子模型（Orthogonal Factor Model）是因子分析中的一种简化模型，它假设公共因子之间是彼此独立的，即它们之间的相关性为零。



* 1. 载荷估计方法：

常用的因子载荷估计方法包括主成分法（Principal Component Method）、主轴因子法（Principal Axis Factoring, PAF）、最大似然法（Maximum Likelihood Estimation, MLE）和加权最小二乘法（Weighted Least Squares, WLS）。主成分法通过主成分分析初始化因子载荷，适合初步探索；主轴因子法专注于公共因子对变量共同度的解释，常用于探索性分析；最大似然法基于概率模型并假设数据正态分布，提供最优参数估计及统计推断；加权最小二乘法则适用于非正态分布的数据，通过对残差平方和赋予不同权重来优化估计。

* 1. 什么是估计因子得分

估计因子得分（Factor Score Estimation）是指在因子分析中，根据观测变量的数据和已估计的因子载荷矩阵来推算每个样本在各个公共因子上的得分。

* 1. 因子旋转的目的和原理（正交、斜交旋转）

目的：简化解释：未经旋转的初始因子解往往难以解释，因为观测变量可能在多个因子上都有中等程度的载荷，这使得很难明确区分哪些变量与哪个因子相关联。旋转后，理想情况下，每个变量应该在少数几个（最好是仅一个）因子上有高载荷，而在其他因子上的载荷接近零。

增强模式识别：旋转可以帮助揭示数据中的潜在结构或模式，使研究者能够更清楚地看到哪些变量组合在一起形成特定的因子。

优化模型拟合：虽然旋转不会改变因子解释总方差的能力，但它可以改善模型的整体拟合度，使得结果更加直观和易于理解。

原理：因子旋转基于两个基本假设：公共因子之间关系不变：旋转不改变因子之间的协方差结构，即公共因子间的相关性保持恒定。

总方差解释量不变：旋转前后，所有因子共同解释的数据总方差保持一致。

正交旋转：正交旋转假设公共因子之间相互独立，彼此无相关性。

斜交旋转：斜交旋转允许公共因子之间存在一定相关性，这在实际应用中更为常见，因为它更贴近现实世界中因子之间的复杂关系。

* 1. 主成分法和主成分分析叫法相似但有什么不同

主成分法和主成分分析（PCA）虽然名称相似，但应用背景和目的不同。主成分法是因子分析中的一种估计方法，主要用于初始化因子载荷矩阵，通过最大化观测变量的方差来提取公共因子，不直接用于数据降维。而PCA是一种独立的数据分析技术，旨在通过线性变换将高维数据映射到低维空间，同时尽可能保留数据中的信息（即最大化的方差），广泛应用于数据降维、可视化和特征提取。简而言之，主成分法侧重于因子分析的初步阶段，而PCA则专注于数据简化和结构揭示。

* 1. 主成分分析和因子分析有什么差异？

主成分分析（PCA）和因子分析（FA）虽然都是用于降维和数据结构简化的统计方法，但它们的目的和假设不同。PCA旨在通过线性组合原始变量来创建新的正交变量（主成分），以最大化数据方差并实现数据降维和可视化；而因子分析则试图识别潜在的、不可观测的因子，这些因子解释了观测变量之间的相关性或共变关系，更多地用于探索数据内部结构和理论构建。

* 1. 因子的载荷所表示的意义、因子的方差贡献等。

因子载荷（Factor Loadings）表示观测变量与潜在因子之间的线性关系强度。

因子的方差贡献是指某个因子能够解释的数据总变异的比例，它衡量了因子的重要性或影响力。

1. 聚类分析的基本概念、思路和相关原理；什么是有监督学习、无监督学习；聚类分析和分类分析有什么区别？如何刻画相似度？什么是层次聚类？系统树图的基本概念和绘制方法、层次聚类的方法（简单连接、完全连接、平均连接、质心连接）和相关计算比如如何求解距离矩阵、如果基于距离矩阵进行层次聚类和得到相关系统树图。
   1. 基本概念：

簇（Cluster）：由一组具有相似特征的数据点组成的集合。

相似性/距离度量：用于衡量数据点之间相似程度的指标，常见的包括欧几里得距离、曼哈顿距离、余弦相似度等。

质心（Centroid）：在某些聚类算法中，每个簇有一个代表性的中心点，即该簇所有成员坐标的平均值。

链接方法（Linkage Method）：层次聚类中用来定义簇间距离的方法，如单链接法（Single Linkage）、全链接法（Complete Linkage）、平均链接法（Average Linkage）等。

* 1. 思路：

1选择合适的数据表示：确保数据已经被适当地预处理和标准化，以便于比较。

2确定相似性或距离度量：选择合适的度量标准来评估数据点之间的相似性或差异。

3应用聚类算法：根据数据特性和需求选择适当的聚类算法，并调整参数以获得最佳结果。

4解释和验证聚类结果：分析生成的簇，检查它们是否有实际意义，并通过内部或外部指标评估聚类的质量。

* 1. 相关原理：

1聚类算法分类

2簇的有效性评估

3确定最优簇数

* 1. 监督学习与无监督学习：

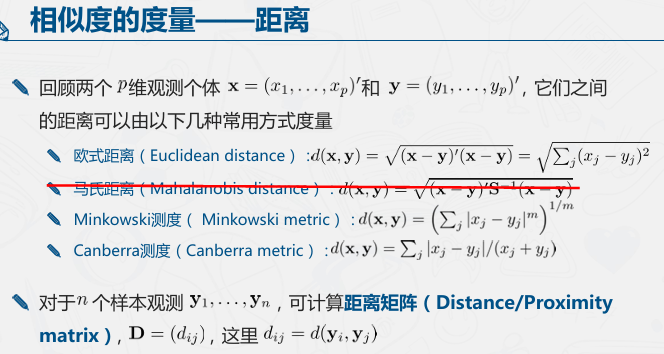


* 1. 聚类分析和分类分析的区别：

聚类分析是无监督学习；分类分析是监督学习

聚类分析评估标准模糊；分类分析评估标准清晰

* 1. 如何刻画相似度？



* 1. 什么是层次聚类？



我们主要讨论凝聚法

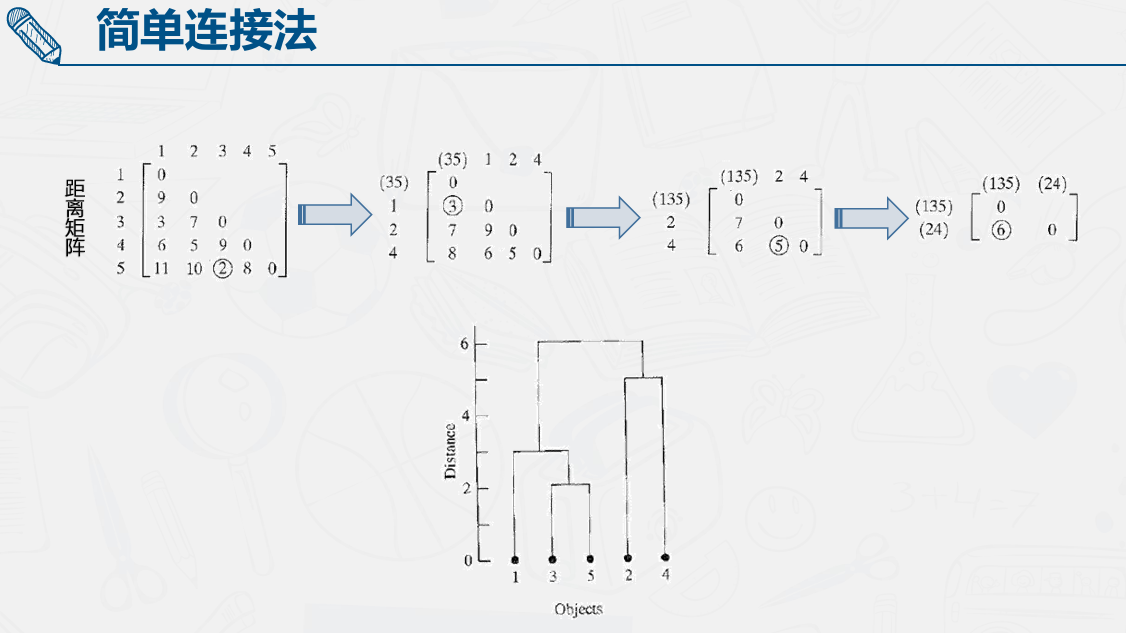
* 1. 系统树图的基本概念和绘制方法

系统树图（Dendrogram）是层次聚类分析中常用的可视化工具，它以树状结构展示了数据点之间的层级关系。系统树图能够直观地展示样本或特征间的相似度以及它们是如何逐步合并成更大的簇的。

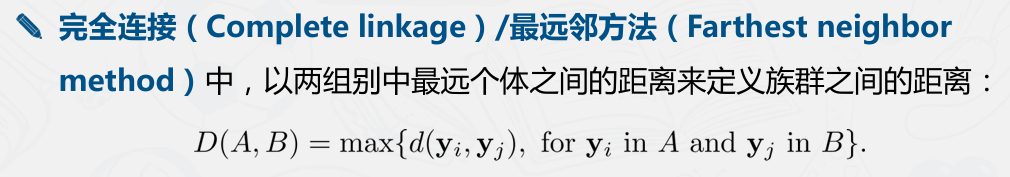
绘制方法：

1. 选择合适的聚类算法
2. 计算距离矩阵
3. 选择链接法则
4. 迭代合并
5. 绘制系统树图
6. 解释和应用
   1. 层次聚类的方法（简单连接、完全连接、平均连接、质心连接）和相关计算比如如何求解距离矩阵、如何基于距离矩阵进行层次聚类和得到相关系统树图

1.简单连接法



2.完全连接法（注意是先选取距离最短的两个群体合并，再选取两个群体中与其他群体最长的距离构成新的距离矩阵）



3.平均连接法（把两个族群内所有点做笛卡尔积，再将每组点求欧式距离，最后将所有距离做平均值）



4.质心连接法（先找簇内样本点均值，再算两均值的欧式距离）

