

数据分析方法



1- 名词解释

1-1-随机向量

1-2-分布函数

1-3-边缘分布

1-4-条件分布

1-5-最小二乘法

1-6-可决系数

1-7-聚类分析相关距离公式

1-8-相似系数

1-9-相关系数

1-10-特征值&特征向量

1-11-因子载荷矩阵

1-12-回归方程的显著性检验

1-13-回归系数的显著性检验

2- 简答题

2-1-回归分析的基本思想

2-2-回归分析的检验流程

2-3-什么是欧式距离，什么是马氏距离，马氏距离和欧式距离相比有什么优势？距离判别法有什么缺点。

2-4-距离判别分析、贝叶斯判别分析和费希尔判别分析的基本思想以及其异同

2-5-比较判别分析和聚类分析的异同

2-6-简述K均值聚类法和系统聚类法的基本步骤以及二者的异同

2-7-比较主成分与因子分析的异同

2-8-主成分与原始变量的关系

2-9-主成分分析的性质

2-10-因子载荷矩阵的统计意义

3-计算题

掌握数学期望、协方差矩阵、相关矩阵、标准化变换、马氏距离的概念以及计算

◆ 例2.2.3 设 $x=(x_1, x_2, x_3)'$ 的数学期望和协差阵分别为

$$\mu = \begin{pmatrix} 5 \\ -2 \\ 7 \end{pmatrix} \text{ 和 } \Sigma = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{pmatrix}$$

令 $y_1=2x_1-x_2+4x_3, y_2=x_2-x_3, y_3=x_1+3x_2-2x_3$, 试求 $y=(y_1, y_2, y_3)'$ 的数

◆ 例2.2.5 设 $\mathbf{x}=(x_1, x_2, x_3)'$ 的数学期望和协差阵分别为

$$\boldsymbol{\mu} = \begin{pmatrix} 5 \\ -2 \\ 7 \end{pmatrix} \quad \text{和} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{pmatrix}$$

试求 \mathbf{x} 的相关矩阵。

二维随机向量 x 总体分布如下：

$X = [[-3, 2], [-2, 2], [-1, 2], [0, 2], [1, 2], [2, 2], [3, 2],$
 $[-3, 1], [-2, 1], [-1, 1], [0, 1], [1, 1], [2, 1], [3, 1],$
 $[-3, 0], [-2, 0], [-1, 0], [0, 0], [1, 0], [2, 0], [3, 0],$
 $[-3, -1], [-2, -1], [-1, -1], [0, -1], [1, -1], [2, -1], [3, -1],$
 $[-3, -2], [-2, -2], [-1, -2], [0, -2], [1, -2], [2, -2], [3, -2]]$

- (1) 求 X 中所有元素到总体 X 的欧式距离；
- (2) 求 X 中所有元素到总体 X 的标准化后的欧式距离；
- (3) 求 X 中所有元素到总体 X 的马氏距离。

主成分求解计算（根据协方差矩阵的特征值和特征向量）

1. 设随机变量 $X = (x_1, x_2, \dots, x_6)'$ 的相关矩阵为

$$R = \begin{pmatrix} 1.00000 & -0.31529 & -0.25061 & -0.65273 & 0.23696 & 0.43478 \\ -0.31529 & 1.00000 & -0.53312 & -0.03919 & 0.58914 & 0.28382 \\ -0.25061 & -0.53312 & 1.00000 & 0.65089 & -0.80595 & -0.66404 \\ -0.65273 & -0.03919 & 0.65089 & 1.00000 & -0.72416 & -0.89963 \\ 0.23696 & 0.58914 & -0.80595 & -0.72416 & 1.00000 & 0.85434 \\ 0.43478 & 0.28382 & -0.66404 & -0.89963 & 0.85434 & 1.0000 \end{pmatrix},$$

经计算 R 的特征值分别为 $\lambda_1 = 3.684508$, $\lambda_2 = 1.5702$, $\lambda_3 = 0.367204$, $\lambda_4 = 0.254582$, $\lambda_5 = 0.083233$,

$\lambda_6 = 0.040273$, 相应的特征向量分别为

T_1	T_2	T_3	T_4	T_5	T_6
-0.250076	0.611687	0.619026	-0.397153	0.084256	0.123606
-0.223909	-0.668246	0.316087	-0.509386	-0.359721	-0.120534
0.446487	0.196485	-0.439811	-0.718962	0.062010	-0.218764
0.465850	-0.304856	0.229875	-0.107262	0.401958	0.681270
-0.485336	-0.202006	-0.151119	-0.133738	0.801457	-0.201460
-0.485191	0.083287	-0.497778	-0.191044	-0.236119	0.646236

试求 X 的前三个主成分及其对总变差的贡献率、前三个主成分对总变差的累积贡献率。

掌握因子分析中因子载荷矩阵的求解过程（根据约相关矩阵的特征值和特征向量）

例2 假定某地固定资产投资率 x_1 ，通货膨胀率 x_2 ，失业率 x_3 ，相关系数矩阵为

$$\begin{bmatrix} 1 & \frac{1}{5} & -\frac{1}{5} \\ \frac{1}{5} & 1 & -\frac{2}{5} \\ -\frac{1}{5} & -\frac{2}{5} & 1 \end{bmatrix}$$

建立因子分析模型
，求解因子载荷矩阵

4-应用题

掌握层次聚类分析和K-Means聚类分析的基本思想和聚类过程，并学会应用

2. 下面是6个样品两两间的距离矩阵：

$$D^{(0)} = \begin{bmatrix} 0 & & & & & \\ 2 & 0 & & & & \\ 7 & 5 & 0 & & & \\ 10 & 9 & 4 & 0 & & \\ 15 & 12 & 7 & 4 & 0 & \\ 18 & 15 & 9 & 5 & 2 & 0 \end{bmatrix}$$

试用最短距离法作系统聚类，并画出谱系聚类图。

5-选择&填空

1. 一元线形回归模型表示如下：

$$Y = \beta_0 + \beta_1 X + u \quad E(Y) = \beta_0 + \beta_1 X$$

已知 $n=62$, $\sum_{i=1}^{62} X_i = 620$, $\sum_{i=1}^{62} Y_i = 1240$, $L_{XX} = \sum_{i=1}^{62} (X_i - \bar{X})^2 = 40$, $L_{YY} = \sum_{i=1}^{62} (Y_i - \bar{Y})^2 = 250$,

$L_{XY} = \sum_{i=1}^{62} (X_i - \bar{X})(Y_i - \bar{Y}) = -80$, 则 Y 对 X 的直线回归方程为_____,

这种估计回归系数的方法称为_____。

2. 回归方程的显著性检验是检验_____是否显著；回归系数的显著性检验是检验_____是否显著。

3. 因子分析把每个原始变量分解为两部分因素：一部分为_____, 另一部分为_____。

4. 聚类分析就是分析如何对样品（或变量）进行量化分类的问题。通常聚类分析分为_____聚类和_____聚类。

5. 多元统计分析是运用_____方法来研究解决_____问题的理论和方法。

6. 变量之间的线性依存关系涉及“离差平方和”的分解公式为：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y})^2$$

其中： $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 是回归偏差； $SSE = \sum_{i=1}^n (y_i - \hat{y})^2$ 剩余偏差。则判定系数

$$R^2 = \frac{SSR}{SST}$$

主要内容

理论基础

掌握随机向量、分布函数、边缘分布、条件分布等概念的定义；掌握数学期望、协方差矩阵、相关矩阵、标准化变换、马氏距离的概念以及计算。

变量间的相互联系—回归分析

掌握回归分析的基本思想、最小二乘法求解回归系数、回归方程检验流程（拟合优度检验、可决系数、回归方程的显著性检验、回归系数的显著性检验）

归类问题—判别分析、聚类分析

掌握距离判别分析、贝叶斯判别分析和费希尔判别分析的基本思想以及其异同，掌握层次聚类分析和K-Means聚类分析的基本思想和聚类过程，并学会应用。

降维问题—主成分分析、因子分析

掌握二者的基本思想以及他们的相同和不同之处，理解主成分分析得到的主成分与原始变量的基本关系，主成分求解计算（根据协方差矩阵的特征值和特征向量），主成分分析的性质（方差贡献率等）；掌握因子分析中因子载荷矩阵的求解过程（根据约相关矩阵的特征值和特征向量），掌握因子载荷矩阵的统计意义（因子载荷、变量共同度、公因子的方差贡献）。

判别分析(Discriminatory Analysis)产生于 20 世纪 30 年代,是利用已知类别的样本建立判别模型,为未知类别的样本判别的一种统计方法。近年来,判别分析在自然科学、社会学及经济管理学科中都有广泛的应用。判别分析的特点是根据已掌握的、历史上每个类别的若干样本的数据信息,总结出客观事物分类的规律性,建立判别公式和判别准则。当遇到新的样本点时,只要根据总结出来的判别公式和判别准则,就能判别该样本点所属的类别。

判别分析与聚类分析都要求对样本进行分类,但两者的分析内容和要求是不一样的。聚类分析是给定了一定量的样品,但这些样品应该划分成怎样的类别还不清楚,需要通过聚类分析来决定。判别分析是已知样品应分为怎样的类别,即在类别已知的情况下,判别每一个样品应属于怎样的类别。判别分析与聚类分析都要求对样本进行分类,两者虽然不同,但也有一定的联系。判别分析中,在决定某一样本应属于哪一类型时,往往也使用聚类分析中的一些思想和方法。】

5-选择&填空

1. 设 X 和 Y 是两个随机向量, 则 X 和 Y 的协差阵与 Y 和 X 的协差阵 ()。
A. 互为转置 B. 不相等, 但阶数一定相同
C. 没有关系 D. 相等
2. 关于样本主成分的总样本方差与原始变量的总样本方差之间的关系, 如下哪一论述是正确的。 ()
A. 样本主成分的总样本方差大于原始变量的总样本方差
B. 样本主成分的总样本方差与原始变量的总样本方差没有必然的关系
C. 样本主成分的总样本方差等于原始变量的总样本方差
D. 样本主成分的总样本方差小于原始变量的总样本方差
3. 以下哪种系统聚类法的类与类之间的距离定义不止一种。 ()
A. 最短距离法 B. 离差平方和法
C. 类平均法 D. 最长距离法

一、随机向量基础

1. 随机向量 $X = (X_1, X_2)^T$ 的协方差矩阵 Σ ()。
A. 主对角线元素为方差, 非对角线元素为协方差
B. 一定是正定矩阵
C. 非对角线元素全为0
D. 与相关矩阵完全相同
2. 设 X 和 Y 为随机向量, 则 $\text{Cov}(X, Y)$ 与 $\text{Cov}(Y, X)$ 的关系是 ()。
A. 互为转置
B. 相等
C. 无必然联系
D. 阶数不同
3. 马氏距离用于度量 ()。
A. 欧氏空间中两点间的直线距离
B. 考虑变量相关性的标准化距离
C. 分类问题中的类别间距
D. 时间序列的相似性

二、回归分析

4. 最小二乘法估计线性回归系数的目标是（）。
A. 最大化可决系数 R^2
B. 最小化残差平方和
C. 最小化回归系数的方差
D. 最大化似然函数
5. 关于可决系数 R^2 ，错误的是（）。
A. $R^2 \in [0, 1]$ ，值越大模型拟合越好
B. 增加自变量一定提高 R^2
C. 衡量因变量与自变量的线性关系强度
D. 对非线性模型无意义
6. 回归方程的显著性检验（F检验）用于检验（）。
A. 单个回归系数是否为零
B. 所有回归系数是否同时为零
C. 模型是否存在异方差
D. 自变量间是否存在多重共线性
7. 若回归系数的t检验p值>0.05，表明（）。
A. 该自变量对因变量影响显著
B. 该自变量应从模型中剔除
C. 模型整体无效
D. 因变量与自变量无线性关系

三、判别分析

8. 距离判别法的核心是（）。
A. 计算样本到各类别中心的马氏距离
B. 基于贝叶斯定理计算后验概率
C. 最大化类间方差与类内方差的比值
D. 最小化误判代价
9. 费希尔判别分析的目标是（）。
A. 找到投影方向使类内离散度最小、类间离散度最大
B. 直接计算样本属于各类的概率
C. 利用先验概率调整判别边界
D. 通过距离阈值分类
10. 贝叶斯判别分析与距离判别法的主要区别在于（）。
A. 是否考虑变量的协方差结构
B. 是否使用先验概率和误判代价
C. 是否需要假设总体分布
D. 是否适用于多类别问题

四、聚类分析

11. 系统聚类法中，类间距离定义方式不唯一的是（）。

- A. 最短距离法
- B. 离差平方和法 (Ward法)
- C. 类平均法
- D. 最长距离法

12. K-Means聚类与层次聚类的本质区别是（）。

- A. K-Means需预先指定类别数，层次聚类不需要
- B. K-Means基于距离，层次聚类基于相似度
- C. K-Means只能处理数值型数据
- D. 层次聚类结果具有唯一性

13. 离差平方和法 (Ward法) 的类间距离定义为（）。

- A. 两类最近样本间的距离
- B. 两类中心点的欧氏距离
- C. 两类合并后总离差平方和的增量
- D. 两类最远样本间的距离

五、主成分分析 (PCA)

14. 设原始变量协方差矩阵的特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ ，则第 k 主成分的方差为（）。

- A. λ_k
- B. $\sum_{i=1}^k \lambda_i$
- C. $\lambda_k / \sum_{i=1}^p \lambda_i$
- D. $1 / \lambda_k$

15. 主成分分析的总样本方差与原始变量的总样本方差的关系是（）。

- A. 主成分的总方差更大
- B. 两者相等
- C. 主成分的总方差更小
- D. 无必然联系

16. 关于主成分的贡献率，正确的是（）。

- A. 是主成分方差占原始变量总方差的比例
- B. 反映主成分与原始变量的相关系数
- C. 累计算贡献率随主成分数量增加而减小
- D. 第一主成分贡献率一定超过50%

六、因子分析

17. 因子载荷矩阵 A 的元素 a_{ij} 表示（）。
- A. 第 i 个变量与第 j 个公因子的相关系数
 - B. 第 j 个公因子对第 i 个变量的方差贡献
 - C. 第 i 个变量被公因子解释的比例
 - D. 公因子间的协方差

18. 变量共同度 (Communality) 是指（）。
- A. 所有公因子对某一变量方差的解释比例
 - B. 某一公因子对所有变量方差的解释比例
 - C. 公因子间的相关性
 - D. 特殊因子的方差

19. 因子旋转的目的是（）。
- A. 提高公因子的方差贡献率
 - B. 使因子载荷矩阵结构简化，易于解释
 - C. 减少公因子数量
 - D. 消除特殊因子的影响

七、综合应用

20. 对标准化后的数据计算马氏距离，等价于（）。
- A. 欧氏距离
 - B. 余弦相似度
 - C. 相关系数
 - D. 曼哈顿距离

21. 若回归模型中存在多重共线性，会导致（）。
- A. 回归系数估计值不稳定，标准差增大
 - B. 可决系数 R^2 降低
 - C. 残差平方和减小
 - D. F检验失效

22. 在主成分分析中，第一主成分的方向对应于（）。
- A. 协方差矩阵的最大特征值对应的特征向量
 - B. 协方差矩阵的最小特征值对应的特征向量
 - C. 原始变量的均值向量
 - D. 样本方差最大的坐标轴

23. 因子分析中，公因子方差 (SS loadings) 反映（）。
- A. 每个公因子对所有原始变量方差的解释量
 - B. 每个变量被公因子解释的方差
 - C. 特殊因子的方差和
 - D. 因子间的相关性

24. 下列哪种方法不属于监督学习？（）
- A. 判别分析
 - B. K-Means聚类
 - C. 逻辑回归
 - D. 决策树

25. 标准化变换的作用是（）。
- A. 使变量均值为0，方差为1
 - B. 消除变量间的相关性
 - C. 将数据转换为正态分布
 - D. 减少异常值影响