**云南大学数学与统计学院**

**实验报告**

**实验课名称： 应用多元统计分析实验**

**指导教师： 李会琼**

**专业（年级）： 统计学2021级**

**学生姓名： 枫叶 学号:**

**实验名称： 实验10**

**实验成绩：**

**《应用多元统计分析实验》实验报告 10**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验名称 | 多元正态分布均值向量的检验 | | 实验成绩 |  | |
| 学号 |  | | 姓名 | 枫叶 | |
| 实验时间 | 2024年6月4日 | 实验地点 | 格物楼3508 | 指导教师 | 李会琼 |
| 1. **实验目的**   学习使用R软件进行距离判别   1. **实验要求**   1. 对所使用的方法与所得到的结果进行适当的文字描述。  2. 在实验结果的相应部分附上完整的代码与适当的注释。  3. 采用一定的可视化方法体现出对应计算结果。   1. **实验内容**   6-10，6-11，6-12用啤酒数据案例步骤做   1. **实验软件**   R语言   1. **实验结果**   需要指出，狗熊会的啤酒数据案例中对系统聚类结果“画出你认为最适合的分割”意义不大且较为繁琐，其后续是基于k-means聚类的碎石图确定类别数，考虑到本实验中样本量都不大，这里采用先系统聚类，根据系统聚类的碎石图确定类别数，同时也基于这一类别数进行k-means聚类 加载包 library(readxl) library(MASS) library(dplyr) library(factoextra) 第一题Q聚类 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  apply(2,scale) #系统聚类选择最优聚类数 fviz\_nbclust(data,FUN = hcut, method = "wss",k.max = 5)    从肘形图来看，可以考虑分为4类  #动态聚类 kmeans(data,4)  ## K-means clustering with 4 clusters of sizes 1, 2, 2, 1 ##  ## Cluster means: ## X1 X2 X3 X4 X5 X6 ## 1 -0.4269830 -0.8139568 -1.8119419 -1.57958730 -1.64452187 -0.6063823 ## 2 -0.3793964 -0.7862805 0.1768136 0.08635119 0.86844009 0.6196029 ## 3 -0.4274072 1.1377887 0.8418505 1.01138623 -0.01007658 0.1445446 ## 4 2.0405901 0.1109406 -0.2253862 -0.61588755 -0.07220516 -0.9219127 ## X7 ## 1 0.7939407 ## 2 -0.1261971 ## 3 -0.9957984 ## 4 1.4500503 ##  ## Clustering vector: ## [1] 3 3 1 2 4 2 ##  ## Within cluster sum of squares by cluster: ## [1] 0.0000000 4.0882813 0.6691743 0.0000000 ## (between\_SS / total\_SS = 86.4 %) ##  ## Available components: ##  ## [1] "cluster" "centers" "totss" "withinss" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault"  K-means聚类结果如上，可以看到类间方差贡献了总方差的86.4%，聚类效果较好，但可能有过拟合之嫌 R聚类直接转置 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  t() #系统聚类选择最优聚类数 fviz\_nbclust(data, FUN = hcut, method = "wss",k.max = 6)    从肘形图来看，可以考虑分为3类  #动态聚类 m <- kmeans(data,3) m  ## K-means clustering with 3 clusters of sizes 2, 4, 1 ##  ## Cluster means: ## [,1] [,2] [,3] [,4] [,5] [,6] ## 1 4323.84500 5193.5000 6678.50000 9335.00000 10309.50000 6557.5000 ## 2 93.64575 92.7411 14.78179 81.09552 61.96375 63.7822 ## 3 1742.00000 2000.0000 1445.00000 1461.00000 1266.00000 2820.0000 ##  ## Clustering vector: ## X1 X2 X3 X4 X5 X6 X7  ## 2 2 2 2 1 3 1  ##  ## Within cluster sum of squares by cluster: ## [1] 106357022.5 322241.5 0.0 ## (between\_SS / total\_SS = 80.3 %) ##  ## Available components: ##  ## [1] "cluster" "centers" "totss" "withinss" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault"  K-means聚类结果如上，可以看到类间方差贡献了总方差的80.3%，聚类效果较好，与上一题相比，过拟合可能性略微降低 基于相关系数转化距离矩阵 为简便起见，用相关系数度量变量间相似关系  R <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  cor() #系统聚类选择最优聚类数 fviz\_nbclust(R, FUN = hcut, method = "wss",k.max = 6)    从肘形图来看，可以考虑分为3类  #动态聚类 kmeans(R,3)  ## K-means clustering with 3 clusters of sizes 2, 4, 1 ##  ## Cluster means: ## X1 X2 X3 X4 X5 X6 X7 ## 1 0.8563240 -0.2452065 -0.3886162 -0.5742737 -0.1485210 -0.4909403 0.8563240 ## 2 -0.0996406 0.6094268 0.8348498 0.8128152 0.5972361 0.1106660 -0.5786681 ## 3 -0.4361296 -0.2187961 0.2076847 0.3312584 0.1225170 1.0000000 -0.5457511 ##  ## Clustering vector: ## X1 X2 X3 X4 X5 X6 X7  ## 1 2 2 2 2 3 1  ##  ## Within cluster sum of squares by cluster: ## [1] 0.5821408 1.5172319 0.0000000 ## (between\_SS / total\_SS = 85.7 %) ##  ## Available components: ##  ## [1] "cluster" "centers" "totss" "withinss" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault" 第二题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题5-9数据.xlsx") %>%  select(-2) %>%  mutate(across(1,factor,labels=c(1,2)))  ## Warning: There was 1 warning in `mutate()`. ## ℹ In argument: `across(1, factor, labels = c(1, 2))`. ## Caused by warning: ## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0. ## Supply arguments directly to `.fns` through an anonymous function instead. ##  ## # Previously ## across(a:b, mean, na.rm = TRUE) ##  ## # Now ## across(a:b, \(x) mean(x, na.rm = TRUE))  #系统聚类选择最优聚类数 fviz\_nbclust(data[2:4], FUN = hcut, method = "wss")    从肘形图来看，可以考虑分为4类，不过我们已经知道样本只有2类，故下面选择2类进行k-means聚类  #动态聚类 m <- kmeans(data,2) m  ## K-means clustering with 2 clusters of sizes 7, 7 ##  ## Cluster means: ## 类型 Cu Ag Bi ## 1 1 2.902857 1.458571 0.9014286 ## 2 2 2.345714 1.628571 1.1685714 ##  ## Clustering vector: ## [1] 1 1 1 1 1 1 1 2 2 2 2 2 2 2 ##  ## Within cluster sum of squares by cluster: ## [1] 3.061314 1.605743 ## (between\_SS / total\_SS = 51.4 %) ##  ## Available components: ##  ## [1] "cluster" "centers" "totss" "withinss" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault"  prop.table(table(m$cluster==as.numeric(data$类型)))  ##  ## TRUE  ## 1  虽然类间方差只占总方差的51.4%，但聚类结果与样本的真实类别完全一致 第三题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题5-11数据.xlsx") %>%  select(-1) #系统聚类选择最优聚类数 fviz\_nbclust(data[-4], FUN = hcut, method = "wss")    从肘形图来看，选择3类是不错的选择  #动态聚类 m <- kmeans(data,3) m  ## K-means clustering with 3 clusters of sizes 6, 4, 4 ##  ## Cluster means: ## X1 X2 X3 类别 ## 1 0.0720 0.05550 0.2368333 2 ## 2 0.1995 0.07200 0.2647500 1 ## 3 0.0265 0.01025 0.1225000 3 ##  ## Clustering vector: ## [1] 1 1 1 3 3 2 1 2 2 1 3 3 2 1 ##  ## Within cluster sum of squares by cluster: ## [1] 0.008818333 0.005037750 0.004290750 ## (between\_SS / total\_SS = 99.8 %) ##  ## Available components: ##  ## [1] "cluster" "centers" "totss" "withinss" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault"  prop.table(table(m$cluster==as.numeric(data$类别)))  ##  ## TRUE  ## 1  类间方差占比达到了99.8%，且聚类结果与真实类别完全一致 | | | | | |