《多元统计分析》课后作业

姓名：李倩倩

学号：2024017349

班级：统计 24-1 班

中国石油大学（北京）克拉玛依校区文理学院数学与统计系

Friday 20th March, 2026

作业要求

1. 可以和其他同学讨论作业当中的问题，但应当自己独立完成作业
2. 计算、证明等要有过程，要有主要步骤的说明
3. 请将计算、绘图所用的 R 代码以及生成的结果和图像一并添加在作业文件当中
4. 请使用 LaTeX编辑并生成 PDF 格式的文件，第1周作业文件命名方式：学号-姓名-01.pdf
5. 评分标准：每一问得分 $\in \{2, 1, 0\}$
    - 2: 按时完成并上交作业，且答案基本正确
    - 1: 按时完成并上交作业，且答案部分正确
    - 0: 答案完全错误，或者迟交作业(规定时间72小时之后)
6. 请将完成的 PDF 格式的作业文件发送至邮箱：xiaolei@cup.edu.cn
7. 每位同学可以有一次迟交作业的机会，但不得晚于规定时间三日之后
8. 第1周作业截止时间：2026年3月20日24:00

目录

Chapter 1

第一周作业

第一周作业完成时间： Friday 20th March, 2026 22:04

1. [2 分] 最大值一定是异常值吗?

   【解】最大值不一定是异常值。而是用箱线图规则来判定，只有当最大值的位置超过四分位距1.5倍时才算异常值。

2. [2 分] 均值或中位数是否有可能位于四分位数之外?

   【解】均值则可能位于四分位数之外，均值受极端值影响较大，可能出现小于下四分位数或大于上四分位数的情况；而中位数一定位于四分位数之间，因为中位数是第50百分位数，而下四分位数是第25百分位数，上四分位数是第75百分位数，所以必然有下四分位数≤中位数≤上四分位数。

3. [2 分] 假设数据来自标准正态分布 $N(0,1)$. 你预计会有百分之多少的数据可能是异常值呢?

   【解】采用箱线图规则定义异常值：

   $$\text{下界} = Q_1 - 1.5 \times \text{IQR}, \quad \text{上界} = Q_3 + 1.5 \times \text{IQR}.$$

   标准正态分布的下四分位数Q1与上四分位数Q3分别为：

   $$Q_1 = \Phi^{-1}(0.25) \approx -0.67449, \quad Q_3 = \Phi^{-1}(0.75) \approx 0.67449.$$

   四分位距：

   $$\text{IQR} = Q_3 - Q_1 \approx 0.67449 - (-0.67449) = 1.34898.$$

   异常值边界：

   $$L = Q_1 - 1.5 \times \text{IQR} \approx -0.67449 - 1.5 \times 1.34898 = -2.69796,$$

   $$U = Q_3 + 1.5 \times \text{IQR} \approx 0.67449 + 2.02347 = 2.69796.$$

由对称性，可得异常值概率为

$$p = P(Z < L) + P(Z > U) = 2 \times \Phi(L) \approx 2 \times \Phi(-2.69796) \approx 2 \times 0.00349 = 0.00698 \approx 0.698\%.$$

综上：在标准正态分布中，基于箱线图规则，预计约有 0.7% 的数据会被判定为异常值。

4. 关于五数总括中的五个数字.

(a) [2 分] 有没有可能五个数字全部相等呢?

【解】有可能。

(b) [2 分] 如果可能的话，会在什么情况下发生呢?

【解】当且仅当数据集中所有观测值都相等时，五数总括中的五个数字（最小值、下四分位数、中位数、上四分位数、最大值）全部相等。

5. 对于瑞银纸币的对角线变量而言.

(a) [2 分] 使用带宽选择准则来计算对角线变量的最优选定带宽 h 并作核密度估计的图形.

【解】已知样本标准差的公式为

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

, 则使用R计算可得真钞与假钞对角线样本标准差分别为

$$\hat{\sigma}_{genuine} = 0.4470001, \quad \hat{\sigma}_{counterfeit} = 0.5578639$$

又已知基于 Silverman 经验法则的核密度估计带宽公式为：

$$h = 1.06 \cdot \hat{\sigma} \cdot n^{-1/5}.$$

真钞与假钞的样本量 n 均为100 则使用R构造函数计算可得真钞与假钞对角线变量的最优选定带宽分别为

$$h_{genuine} = 0.1886312, \quad h_{counterfeit} = 0.235415$$

图 ?? 展示了真钞与假钞对角线变量的核密度估计曲线。

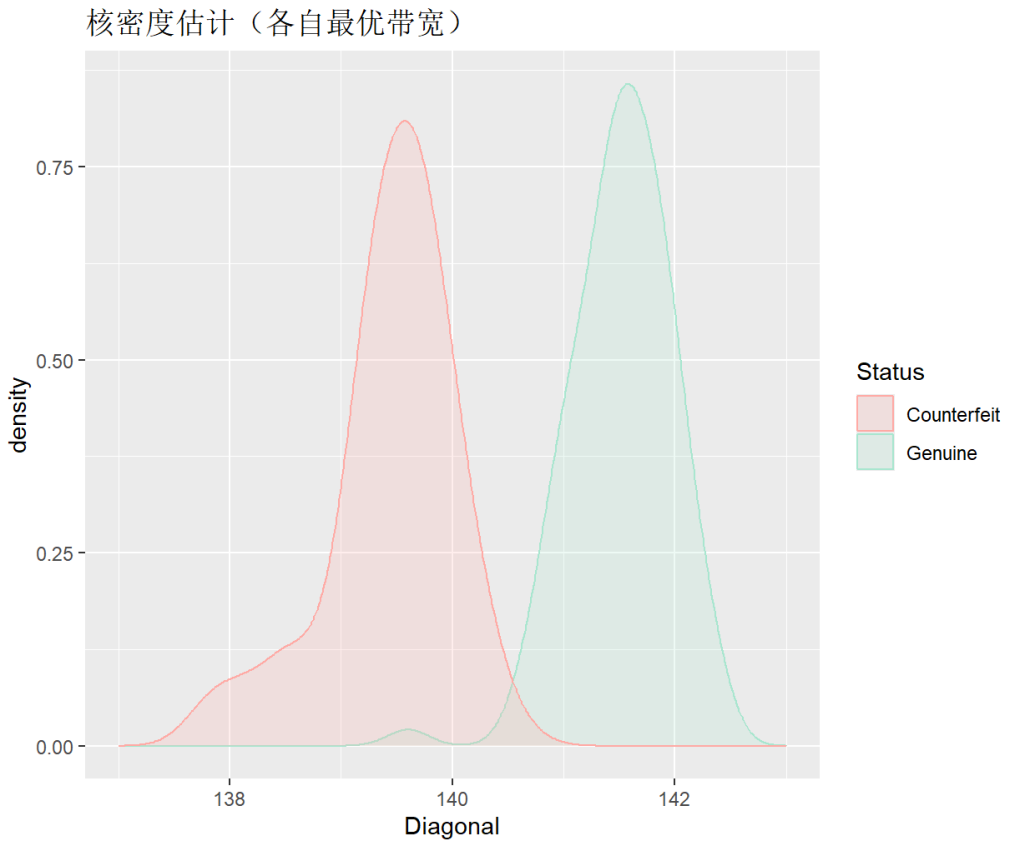

图 1.1: 核密度估计对比图

(b) [2 分] 为这两组 (真钞、假钞) 数据设置同一个带宽会更好吗?

【解】由上一题计算得出这两组数据的最优带宽不同。由于真钞与假钞样本量相同，我们对这两组带宽取均值得

$$\overline{h} = \frac{1}{2} \times (h_genuine = 0.1886312 + h_counterfeit = 0.235415) = 0.2120231 \approx 0.212$$

以这一带宽画出对角线变量的核密度估计的图形

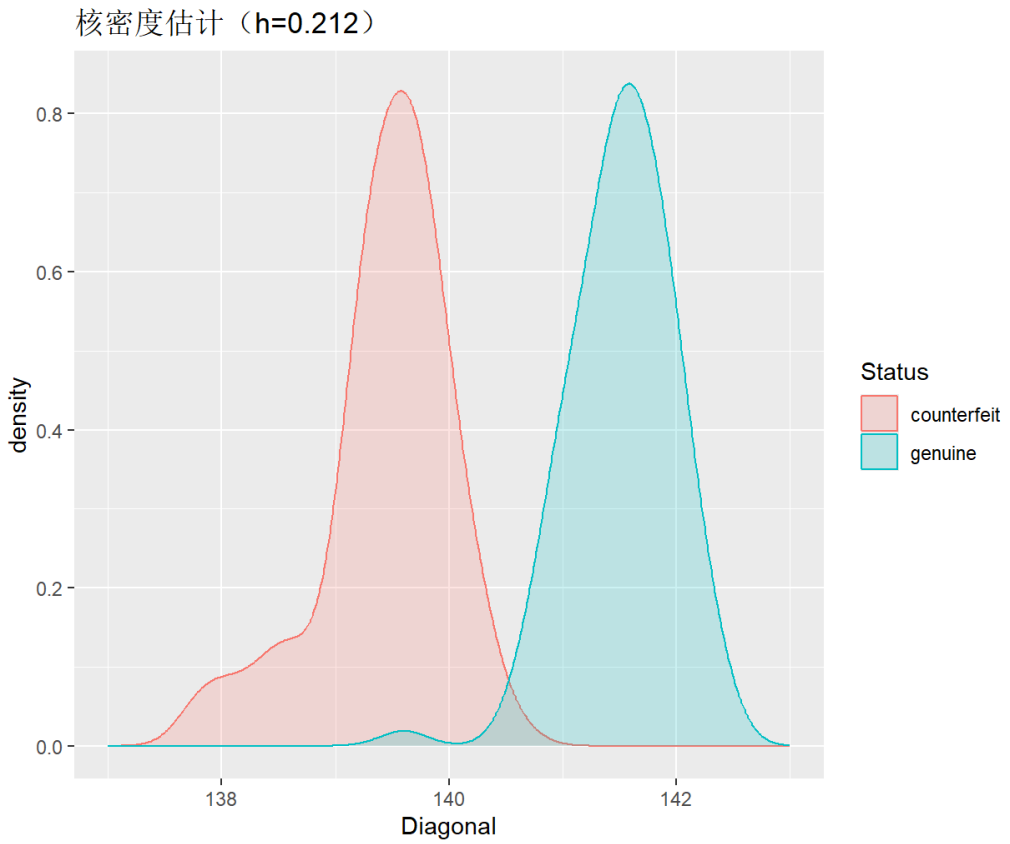

图 1.2: 核密度估计图（使用同一带宽）

再将图1.1与图1.2重叠画出对比图

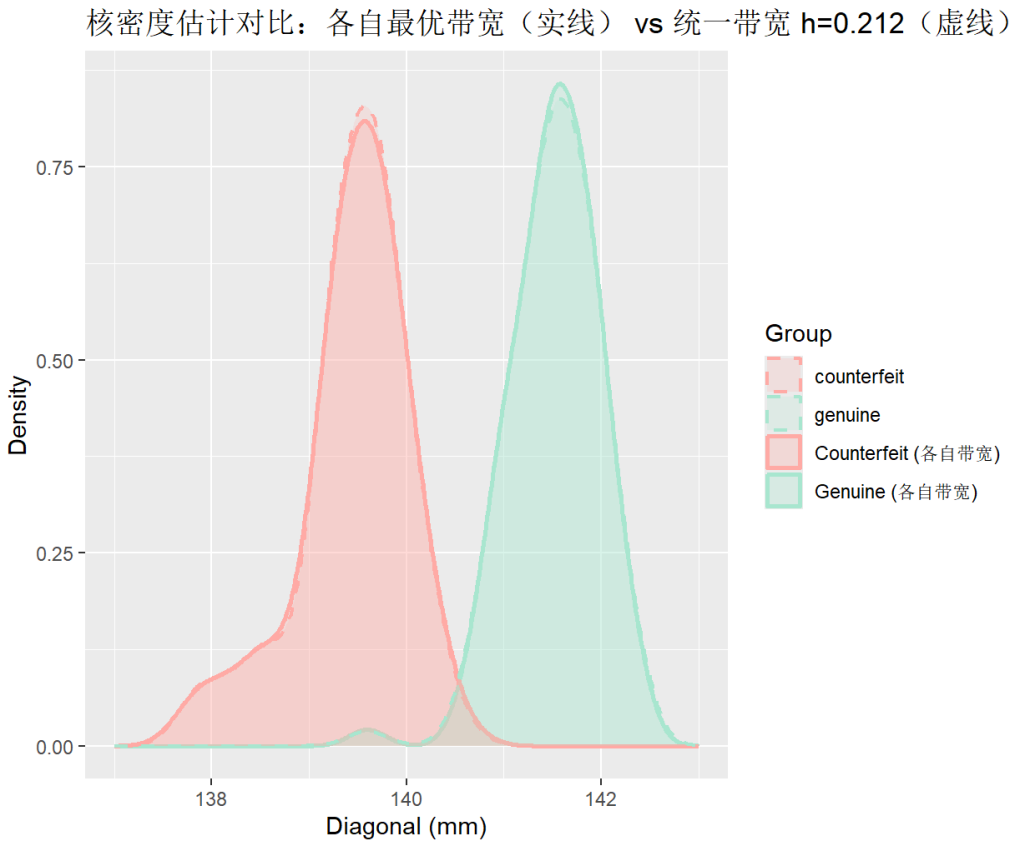

图 1.3: 核密度估计图（各自最优带宽VS同一带宽）

可知设置同一的带宽变化不大，不会更好。

6. [2 分] 设 $|\mathcal{A}| = 0$. 问矩阵 $\mathcal{A}$ 的所有特征值都有可能是正数吗?

【解】 不可能。若 $|\mathcal{A}| = 0$，则 $\mathcal{A}$ 必有一个特征值为 0，因此所有特征值不可能都是正数。

7. [2 分] 设矩阵 $\mathcal{A}$ (方阵) 的所有特征值都不为零. 问矩阵 $\mathcal{A}$ 是否一定可逆?

【解】 一定可逆。若所有特征值都不为零，则 $\det(\mathcal{A}) = \prod \lambda_i \neq 0$，故 $\mathcal{A}$ 可逆。

8. 设有矩阵 $\mathcal{A}$ 如下:

$$\mathcal{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} \tag{1.1}$$

(a) [2 分] 利用 R 计算矩阵 $\mathcal{A}$ 的行列式 $|\mathcal{A}|$.

【解】计算出来 `det_A = -8`

```
A <- matrix(c(1, 2, 3, 2, 1, 2, 3, 2, 1), nrow = 3, byrow = TRUE)
det_A <- det(A)
print(det_A)
```

(b) [2 分] 利用 R 求矩阵 $\mathcal{A}$ 的特征值与特征向量.

【解】特征值: 5.372281, -0.372281, -2.000000; 特征向量: 对应列向量。

```
eig <- eigen(A)
values <- eig$values
vectors <- eig$vectors
print(values)
print(vectors)
```

(c) [2 分] 利用 R 验证矩阵 $\mathcal{A}$ 的 Jordan 分解 (定理 2.1).

【解】代码运行结果表明 A_recovered 与原始矩阵 A 在数值精度内相等，说明 $\mathcal{A}$ 可对角化，Jordan 分解即为其谱分解。

```
P <- vectors
D <- diag(values)
P_inv <- solve(P)
A_recovered <- P %*% D %*% P_inv
# 检查是否与原矩阵相等（允许数值误差）
all.equal(A, A_recovered)
```