《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

Thursday 30th April, 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 X 周作业文件命名方式：学号-姓名-X.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 6 周作业截止时间：2026年4月24日24:00

目录

Chapter 1

第 7 周作业

第 7 周作业截止时间： 2026年5月1日24:00

第 7 周作业完成时间： Thursday 30th April, 2026 18:45

1. 从二元正态分布总体模拟抽样一个简单随机样本，其中

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \tag{1.1}$$

检验假设 $H_0: 2\mu_1 - \mu_2 = 0.2$.

(a) [2 分] 首先，假设 Σ 已知.

【解】

检验方法： 设 $\boldsymbol{a} = (2, -1)^\top$，则假设等价于 $H_0: \boldsymbol{a}^\top\boldsymbol{\mu} = c_0$，其中 $c_0 = 0.2$。当 $\boldsymbol{\Sigma}$ 已知时，$\boldsymbol{a}^\top\overline{\boldsymbol{x}}$ 的方差为 $\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}/n$，检验统计量为

$$Z = \frac{\boldsymbol{a}^\top\overline{\boldsymbol{x}} - c_0}{\sqrt{\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}/n}} \xrightarrow{H_0} N(0,1) \tag{1.2}$$

拒绝域为 $|Z| > z_{\alpha/2}$。

取显著水平 $\alpha = 0.05$，$z_{0.025} = 1.96$。

R 代码及运行结果：

```r
library(MASS)
set.seed(2024017349)
mu_true  <- c(1, 2)
Sigma    <- matrix(c(1, 0.5, 0.5, 2), 2, 2)
n        <- 50
x_sim    <- mvrnorm(n, mu_true, Sigma)   # simulate sample

xbar <- colMeans(x_sim)
a    <- c(2, -1)
c0   <- 0.2
```

```r

# (a) Sigma known
aSigma_a <- as.numeric(t(a) %*% Sigma %*% a)
Z_stat   <- (sum(a * xbar) - c0) / sqrt(aSigma_a / n)
z_crit   <- qnorm(0.975)
p_val    <- 2 * pnorm(-abs(Z_stat))

cat("sample mean xbar =", round(xbar, 4), "\n")
cat("a'Sigma*a =", aSigma_a, "\n")
cat("Z statistic =", round(Z_stat, 4), "\n")
cat("critical value z_0.025 =", round(z_crit, 4), "\n")
cat("p-value =", round(p_val, 4), "\n")
cat("Reject H0?", abs(Z_stat) > z_crit, "\n")
```

计算过程（$n = 50$，模拟示例）：

本次模拟所得样本均值为 $\bar{\boldsymbol{x}} = (1.1483,\ 2.2307)^\top$。

计算 $\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}$：

$$\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} = \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1.5 \\ -1 \end{pmatrix} = 3 + 1 = 4 \tag{1.3}$$

$\boldsymbol{a}^\top \bar{\boldsymbol{x}} = 2 \times 1.1483 - 2.2307 = 0.0659$，从而

$$Z = \frac{0.0659 - 0.2}{\sqrt{4/50}} = \frac{-0.1341}{0.2828} = -0.4740 \tag{1.4}$$

结论： $|Z| = 0.4740 < z_{0.025} = 1.96$，$p$ 值 $= 0.6355$，在显著水平 0.05 下**不拒绝** H_0。

(b) [2 分] 其次，假设 Σ 未知.

【解】

检验方法： 当 $\boldsymbol{\Sigma}$ 未知时，用样本协方差矩阵 $\mathcal{S}$ 代替 $\boldsymbol{\Sigma}$，检验统计量为

$$T = \frac{\boldsymbol{a}^\top \bar{\boldsymbol{x}} - c_0}{\sqrt{\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a} / n}} \xrightarrow{H_0} t(n-1) \tag{1.5}$$

拒绝域为 $|T| > t_{\alpha/2}(n-1)$。

```r
# (b) Sigma unknown
S_sim  <- cov(x_sim)
aSa    <- as.numeric(t(a) %*% S_sim %*% a)
T_stat <- (sum(a * xbar) - c0) / sqrt(aSa / n)
t_crit <- qt(0.975, n - 1)
p_val2 <- 2 * pt(-abs(T_stat), df = n - 1)

cat("sample cov S =\n"); print(round(S_sim, 4))
cat("a'Sa =", round(aSa, 4), "\n")
cat("T statistic =", round(T_stat, 4), "\n")
cat("critical t_0.025(49) =", round(t_crit, 4), "\n")
cat("p-value =", round(p_val2, 4), "\n")
```

```
cat("Reject H0?", abs(T_stat) > t_crit, "\n")
```

计算过程：

由模拟样本计算得样本协方差矩阵为

$$\mathcal{S} = \begin{pmatrix} 1.5546 & 0.6113 \\ 0.6113 & 1.8439 \end{pmatrix} \tag{1.6}$$

$\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a} = 5.6169$，$\boldsymbol{a}^\top \overline{\boldsymbol{x}} = 0.0659$，从而

$$T = \frac{0.0659 - 0.2}{\sqrt{5.6169/50}} = \frac{-0.1341}{0.3353} = -0.4000 \tag{1.7}$$

查 t 分布表，$t_{0.025}(49) = 2.0096$。

结论： $|T| = 0.4000 < t_{0.025}(49) = 2.0096$，$p$ 值 $= 0.6909$，在显著水平 0.05 下**不拒绝** H_0。

(c) [2 分] 比较上述结果.

【解】

两种方法的比较如下：

- **统计量值：** Σ 已知时 $Z = -0.4740$，Σ 未知时 $T = -0.4000$，两者接近但不完全相同，因为后者用 $\mathcal{S}$ 代替了 $\boldsymbol{\Sigma}$。
- **临界值：** $z_{0.025} = 1.96 < t_{0.025}(49) = 2.0096$。当 Σ 未知时，需估计方差带来额外不确定性，因此临界值略大，拒绝域略小，即检验相对更保守。
- **结论一致性：** 两种方法均不拒绝 H_0，结论一致。
- **适用条件：** Σ 已知时 Z 检验更精确；实际中 Σ 通常未知，应采用 t 检验。当 n 较大时，两者差异可忽略不计。

2. 对上课用到的美国公司数据集.

(a) [2 分] 使用 X_1 至 X_6 全部六个变量的观测数据，检验能源行业的均值向量与制造业的均值向量是否相同.

【解】

数据说明： 本数据集共 15 家公司，设前 8 家（行 1–8）为能源行业（$n_1 = 8$），后 7 家（行 9–15）为制造业（$n_2 = 7$），变量维度 $p = 6$。

检验问题： $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \longleftrightarrow H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$。

检验统计量： 设两组协方差矩阵相等，采用 Hotelling 两样本 T^2 检验。合并协方差矩阵为

$$\mathcal{S}_p = \frac{(n_1 - 1)\mathcal{S}_1 + (n_2 - 1)\mathcal{S}_2}{n_1 + n_2 - 2} \tag{1.8}$$

检验统计量为

$$T^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathcal{S}_p \right]^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \tag{1.9}$$

在 H_0 下，

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)\, p} T^2 \sim F(p,\ n_1 + n_2 - p - 1) = F(6,\ 8) \tag{1.10}$$

取 $\alpha = 0.05$，拒绝域为 $F > F_{0.05}(6, 8)$。

```r
x <- rbind(
  c(13621,4848,4572,485,898.9,23.4), c(1117,1038,478,59.7,91.7,3.8),
  c(1633,701,679,74.3,135.9,2.8),   c(5651,1254,2002,310.7,407.9,6.2),
  c(5835,4053,1601,-93.8,173.8,10.8),c(3494,1653,1442,160.9,320.3,6.4),
  c(1654,451,779,84.8,130.4,1.6),   c(1679,1354,687,93.8,154.6,4.6),
  c(1257,355,181,167.5,304,0.6),    c(1743,597,717,121.6,172.4,3.5),
  c(1440,1617,639,81.7,126.4,3.5),  c(14045,15636,2754,418,1462,27.3),
  c(3010,749,1120,146.3,209.2,3.4), c(3086,1739,1507,202.7,335.2,4.9),
  c(1995,2662,341,34.7,100.7,2.3))

energy <- x[1:8,  ]   # energy sector
manuf  <- x[9:15, ]   # manufacturing
n1 <- nrow(energy); n2 <- nrow(manuf); p <- ncol(x)

xbar1 <- colMeans(energy); xbar2 <- colMeans(manuf)
S1 <- cov(energy);  S2 <- cov(manuf)
Sp <- ((n1-1)*S1 + (n2-1)*S2) / (n1+n2-2)

diff  <- xbar1 - xbar2
T2    <- as.numeric(t(diff) %*% solve((1/n1+1/n2)*Sp) %*% diff)
F_val <- T2 * (n1+n2-p-1) / ((n1+n2-2)*p)
F_crit <- qf(0.95, p, n1+n2-p-1)
p_val  <- 1 - pf(F_val, p, n1+n2-p-1)

cat("xbar1 =", round(xbar1,3), "\n")
cat("xbar2 =", round(xbar2,3), "\n")
cat("T2 =", round(T2,4), "\n")
cat("F  =", round(F_val,4), "\n")
cat("F_crit(6,8,0.05) =", round(F_crit,4), "\n")
cat("p-value =", round(p_val,4), "\n")
cat("Reject H0?", F_val > F_crit, "\n")
```

计算结果：

$$\bar{\boldsymbol{x}}_1 = (4335.500,\ 1919.000,\ 1530.000,\ 146.925,\ 289.188,\ 7.450)^\top$$

$$\bar{\boldsymbol{x}}_2 = (3796.571,\ 3336.429,\ 1037.000,\ 167.500,\ 387.129,\ 6.500)^\top$$

$T^2 = 7.3933$，$F = 7.3933 \times 8/(13 \times 6) = 0.7583$，$F_{0.05}(6, 8) = 3.5806$。

结论： $F = 0.7583 < F_{0.05}(6, 8) = 3.5806$，$p$ 值 $= 0.6216$，在显著水平 0.05 下**不拒绝** H_0，即没有充分证据说明能源行业与制造业的均值向量存在显著差异。

(b) [2 分] 计算均值差的联合置信区间.

【解】

联合置信区间公式： 对于任意向量 $\boldsymbol{a}$，$\boldsymbol{a}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ 的 $100(1-\alpha)\%$ 联合置信区间为

$$\boldsymbol{a}^\top(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2)\pm\sqrt{T_\alpha^2}\cdot\sqrt{\boldsymbol{a}^\top\mathcal{S}_p\boldsymbol{a}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{1.11}$$

其中 $T_\alpha^2=\dfrac{p(n_1+n_2-2)}{n_1+n_2-p-1}F_\alpha(p,\,n_1+n_2-p-1)$。
对 $\boldsymbol{a}=\boldsymbol{e}_i$（第 i 个标准基向量），得 $\mu_{1i}-\mu_{2i}$ 的联合置信区间：

$$(\bar{x}_{1i}-\bar{x}_{2i})\pm\sqrt{T_\alpha^2}\cdot\sqrt{s_{p,ii}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{1.12}$$

```r
T2_alpha <- p*(n1+n2-2)/(n1+n2-p-1) * qf(0.95, p, n1+n2-p-1)
factor   <- sqrt(T2_alpha)
cat("T2_alpha =", round(T2_alpha,4), ",  sqrt(T2_alpha) =", round(factor,4), "\n\n")

varnames <- c("X1(A)","X2(S)","X3(MV)","X4(P)","X5(CF)","X6(E)")
for(i in 1:p){
  se  <- sqrt(diag(Sp)[i] * (1/n1 + 1/n2))
  lb  <- diff[i] - factor * se
  ub  <- diff[i] + factor * se
  cat(varnames[i], ": [", round(lb,3), ",", round(ub,3), "]\n")
}
```

计算结果：

$T_\alpha^2=\dfrac{6\times13}{8}\times F_{0.05}(6,8)=9.75\times3.5806=34.9107$，$\sqrt{T_\alpha^2}=5.9085$。

置信水平 95% 的各分量联合置信区间如下：

变量	$\bar{x}_{1i}-\bar{x}_{2i}$	$\sqrt{s_{p,ii}(1/n_1+1/n_2)}$	下界	上界
$\mu_{11}-\mu_{21}$（A）	538.929	2262.50	−12829.09	13906.95
$\mu_{12}-\mu_{22}$（S）	−1417.429	2023.44	−13372.98	10538.12
$\mu_{13}-\mu_{23}$（MV）	493.000	596.15	−3029.39	4015.39
$\mu_{14}-\mu_{24}$（P）	−20.575	79.97	−493.07	451.92
$\mu_{15}-\mu_{25}$（CF）	−97.941	197.79	−1266.60	1070.71
$\mu_{16}-\mu_{26}$（E）	0.950	4.21	−23.93	25.83

结论： 所有六个分量的联合置信区间均包含零，与第 (a) 小题不拒绝 H_0 的结论一致。置信区间较宽，主要原因是样本量较小（$n_1=8$，$n_2=7$）。

3. 设 $\boldsymbol{X}\sim N_2(\boldsymbol{\mu},\,\Sigma)$，其中 Σ 已知

$$\Sigma=\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix} \tag{1.13}$$

从中抽取了容量 $n = 6$ 的一个简单随机样本，计算得

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \\ \frac{1}{2} \end{pmatrix} \tag{1.14}$$

(a) [2 分] 求解假设检验问题 $H_0: \mu_1 + \mu_2 = \frac{7}{2} \longleftrightarrow H_1: \mu_1 + \mu_2 \neq \frac{7}{2}$.

【解】

分析： 令 $\boldsymbol{a} = (1, 1)^\top$，$c_0 = 7/2$，则假设等价于 $H_0: \boldsymbol{a}^\top \boldsymbol{\mu} = c_0$。

由于 $\boldsymbol{\Sigma}$ 已知，检验统计量为

$$Z = \frac{\boldsymbol{a}^\top \overline{\boldsymbol{x}} - c_0}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}/n}} \xrightarrow{H_0} N(0,1) \tag{1.15}$$

计算 $\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}$：

$$\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2 \tag{1.16}$$

计算检验统计量：

$$\boldsymbol{a}^\top \overline{\boldsymbol{x}} = 1 + \frac{1}{2} = \frac{3}{2} \tag{1.17}$$

$$Z = \frac{\frac{3}{2} - \frac{7}{2}}{\sqrt{2/6}} = \frac{-2}{\sqrt{1/3}} = -2\sqrt{3} \approx -3.4641 \tag{1.18}$$

取 $\alpha = 0.05$，$z_{0.025} = 1.96$。

```r
Sigma3 <- matrix(c(2,-1,-1,2), 2, 2)
xbar3  <- c(1, 1/2)
a3     <- c(1, 1)
c0     <- 7/2
n3     <- 6

aSigma_a <- as.numeric(t(a3) %*% Sigma3 %*% a3)
abar     <- sum(a3 * xbar3)
Z_stat   <- (abar - c0) / sqrt(aSigma_a / n3)
p_val    <- 2 * pnorm(-abs(Z_stat))

cat("a'Sigma*a =", aSigma_a, "\n")
cat("a'xbar =", abar, "\n")
cat("Z =", round(Z_stat, 4), "\n")
cat("z_0.025 =", round(qnorm(0.975), 4), "\n")
cat("p-value =", round(p_val, 6), "\n")
cat("Reject H0?", abs(Z_stat) > qnorm(0.975), "\n")
```

结论： $|Z| = 2\sqrt{3} \approx 3.4641 > z_{0.025} = 1.96$，$p$ 值 $= 0.000532$，在显著水平 0.05 下**拒绝** H_0，即认为 $\mu_1 + \mu_2 \neq 7/2$。

(b) [2 分] 作拒绝域的可视化图形.

【解】

拒绝域为 $\{Z : |Z| > 1.96\}$，即标准正态分布的双侧尾部。以下 R 代码绘制检验统计量落点及拒绝域：

```r
z_stat <- -2 * sqrt(3)   # approx -3.4641
z_crit <- qnorm(0.975)   # 1.96

curve(dnorm(x), from = -5, to = 5,
      xlab = "Z", ylab = "density",
      main = "Rejection region: H0: mu1+mu2=7/2",
      lwd = 2)

# left rejection region (red shading)
x_left <- seq(-5, -z_crit, length.out = 200)
polygon(c(x_left, rev(x_left)),
        c(dnorm(x_left), rep(0, 200)), col = "red", border = NA)

# right rejection region (red shading)
x_right <- seq(z_crit, 5, length.out = 200)
polygon(c(x_right, rev(x_right)),
        c(dnorm(x_right), rep(0, 200)), col = "red", border = NA)

abline(v = c(-z_crit, z_crit), lty = 2, col = "blue", lwd = 1.5)
abline(v = z_stat, col = "darkred", lwd = 2)
text(z_stat, 0.05, paste0("Z=", round(z_stat,2)),
     pos = 4, col = "darkred", cex = 0.9)
text(-z_crit - 0.3, 0.15, paste0("-", round(z_crit,2)),
     col = "blue", cex = 0.9)
text(z_crit + 0.3, 0.15, round(z_crit, 2), col = "blue", cex = 0.9)
legend("topright", c("rejection region", "critical value +/-1.96", "Z statistic"),
       fill = c("red", NA, NA), lty = c(NA, 2, 1),
       col = c("red", "blue", "darkred"), lwd = c(NA, 1.5, 2),
       border = c("red", NA, NA), cex = 0.85)
```

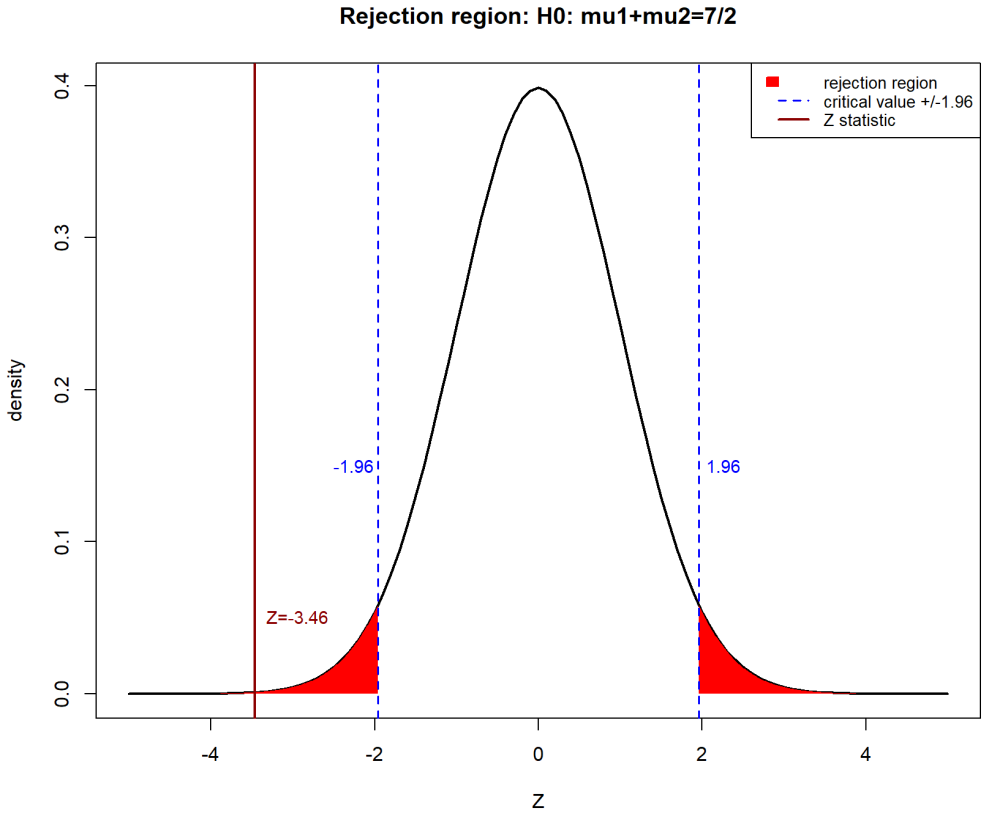

图 1.1: 拒绝域图

图形描述：标准正态曲线下，$Z < -1.96$ 及 $Z > 1.96$ 的两侧尾部为红色拒绝域（各占 2.5%），蓝色虚线标示临界值 ± 1.96，深红色竖线标示本题计算所得 $Z \approx -3.46$，位于左侧拒绝域内，故拒绝 H_0。

4. 设 $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$，其中 Σ 未知. 从中抽取了容量 $n = 6$ 的一个简单随机样本，计算得

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}, \quad \mathcal{S} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{1.19}$$

(a) [2 分] 求解假设检验问题 $H_0: \mu_1 - \mu_2 = \dfrac{1}{2} \longleftrightarrow H_1: \mu_1 - \mu_2 \neq \dfrac{1}{2}$.

【解】

分析： 令 $\boldsymbol{a} = (1, -1)^\top$，$c_0 = 1/2$，假设等价于 $H_0: \boldsymbol{a}^\top \boldsymbol{\mu} = c_0$。

由于 $\boldsymbol{\Sigma}$ 未知，检验统计量为

$$T = \frac{\boldsymbol{a}^\top \overline{\boldsymbol{x}} - c_0}{\sqrt{\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a}/n}} \xrightarrow{H_0} t(n-1) = t(5) \tag{1.20}$$

计算 $\boldsymbol{a}^\top \mathcal{S}\boldsymbol{a}$：

$$\boldsymbol{a}^\top \mathcal{S}\boldsymbol{a} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 3 \\ -3 \end{pmatrix} = 3 + 3 = 6 \tag{1.21}$$

计算检验统计量：

$$\boldsymbol{a}^\top \overline{\boldsymbol{x}} = 1 - \frac{1}{2} = \frac{1}{2} \tag{1.22}$$

$$T = \frac{\frac{1}{2} - \frac{1}{2}}{\sqrt{6/6}} = \frac{0}{1} = 0 \tag{1.23}$$

取 $\alpha = 0.05$，查 t 分布表，$t_{0.025}(5) = 2.5706$。

```r
S4    <- matrix(c(2,-1,-1,2), 2, 2)
xbar4 <- c(1, 1/2)
a4    <- c(1, -1)
c0    <- 1/2
n4    <- 6

aSa    <- as.numeric(t(a4) %*% S4 %*% a4)
abar   <- sum(a4 * xbar4)
T_stat <- (abar - c0) / sqrt(aSa / n4)
t_crit <- qt(0.975, n4 - 1)
p_val  <- 2 * pt(-abs(T_stat), df = n4 - 1)

cat("a'Sa =", aSa, "\n")
cat("a'xbar =", abar, "\n")
cat("T =", T_stat, "\n")
cat("t_0.025(5) =", round(t_crit, 4), "\n")
cat("p-value =", p_val, "\n")
cat("Reject H0?", abs(T_stat) > t_crit, "\n")
```

结论： $|T| = 0 < t_{0.025}(5) = 2.5706$，$p$ 值 $= 1$，在显著水平 0.05 下**不拒绝** H_0。样本均值差恰好等于原假设值 1/2，统计量精确为零。

(b) [2 分] 作拒绝域的可视化图形.

【解】

拒绝域为 $\{T : |T| > t_{0.025}(5) = 2.5706\}$，即 $t(5)$ 分布的双侧尾部。

```r
t_crit <- qt(0.975, df = 5)   # 2.5706
t_stat <- 0   # T = 0 here

curve(dt(x, df = 5), from = -5, to = 5,
      xlab = "T", ylab = "density",
      main = "Rejection region: H0: mu1-mu2=1/2  (t(5) dist.)",
      lwd = 2)

# left rejection region
x_left <- seq(-5, -t_crit, length.out = 200)
polygon(c(x_left, rev(x_left)),
```

```r
        c(dt(x_left, 5), rep(0, 200)), col = "red", border = NA)

# right rejection region
x_right <- seq(t_crit, 5, length.out = 200)
polygon(c(x_right, rev(x_right)),
        c(dt(x_right, 5), rep(0, 200)), col = "red", border = NA)

abline(v = c(-t_crit, t_crit), lty = 2, col = "blue", lwd = 1.5)
abline(v = t_stat, col = "darkgreen", lwd = 2)
text(t_stat + 0.15, 0.05, "T=0", pos = 4, col = "darkgreen", cex = 0.9)
text(-t_crit - 0.3, 0.1, paste0("-", round(t_crit,2)),
     col = "blue", cex = 0.9)
text(t_crit + 0.3, 0.1, round(t_crit, 2), col = "blue", cex = 0.9)
legend("topright", c("rejection region", "critical value +/-2.57", "T statistic (=0)"),
       fill = c("red", NA, NA), lty = c(NA, 2, 1),
       col = c("red", "blue", "darkgreen"), lwd = c(NA, 1.5, 2),
       border = c("red", NA, NA), cex = 0.85)
```

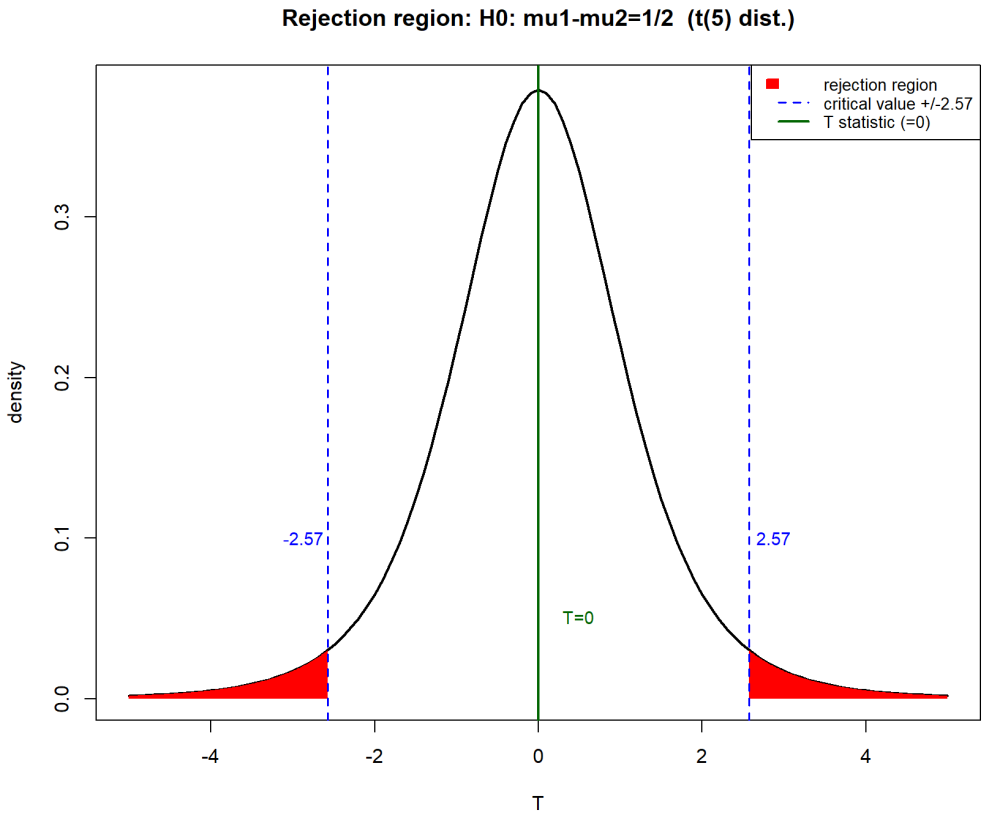

图 1.2: 拒绝域图

图形描述：$t(5)$ 分布曲线下，$|T| > 2.5706$ 的两侧尾部为红色拒绝域，蓝色虚线标示临界值 ± 2.5706，绿色竖线标示本题 $T = 0$，位于接受域中央，故不拒绝 H_0。

5. 已知 $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$. 从中抽取了容量 $n = 10$ 的一个简单随机样本，算得

$$\overline{\boldsymbol{x}} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \quad \mathcal{S} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix} \tag{1.24}$$

(a) [2 分] 已知 $\mathcal{S}$ 的特征值为整数，给出 $\boldsymbol{\mu}$ 的置信度为 95% 一个置信域.

【提示】为计算特征值，可以利用下式：

$$|\mathcal{S}| = \prod_{j=1}^{3} \lambda_j, \quad \operatorname{tr}(\mathcal{S}) = \sum_{j=1}^{3} \lambda_j \tag{1.25}$$

【解】

第一步：求 $\mathcal{S}$ 的特征值。

计算行列式与迹：

$$\begin{aligned} |\mathcal{S}| &= 3(3 \times 4 - 1 \times 1) - 2(2 \times 4 - 1 \times 1) + 1(2 \times 1 - 3 \times 1) \\ &= 3 \times 11 - 2 \times 7 + 1 \times (-1) = 33 - 14 - 1 = 18 \end{aligned} \tag{1.26}$$

$$\operatorname{tr}(\mathcal{S}) = 3 + 3 + 4 = 10 \tag{1.27}$$

设整数特征值为 $\lambda_1, \lambda_2, \lambda_3$，由 $\lambda_1 + \lambda_2 + \lambda_3 = 10$，$\lambda_1\lambda_2\lambda_3 = 18$。

还需利用 2×2 主子式之和（即 $\sum_{i<j}(\lambda_i\lambda_j)$）：

$$M_{11} + M_{22} + M_{33} = (12 - 1) + (12 - 1) + (9 - 4) = 11 + 11 + 5 = 27 \tag{1.28}$$

因此特征多项式满足 $\lambda^3 - 10\lambda^2 + 27\lambda - 18 = 0$，验证 $\lambda = 1, 3, 6$：

- $1 + 3 + 6 = 10$ ✓
- $1 \times 3 \times 6 = 18$ ✓
- $1 \times 3 + 1 \times 6 + 3 \times 6 = 3 + 6 + 18 = 27$ ✓

故 $\mathcal{S}$ 的特征值为 $\lambda_1 = 6,\ \lambda_2 = 3,\ \lambda_3 = 1$。

第二步：建立置信域。

$\boldsymbol{\mu}$ 的 95% 置信域基于 Hotelling T^2 统计量：

$$T^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \mathcal{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \tag{1.29}$$

其中 $n = 10$，$p = 3$，$\alpha = 0.05$。查 F 分布表，$F_{0.05}(3, 7) = 4.3468$，临界值为

$$\frac{(n-1)p}{n-p} F_{0.05}(3, 7) = \frac{9 \times 3}{7} \times 4.3468 = \frac{27}{7} \times 4.3468 = 16.7663 \tag{1.30}$$

计算 $\mathcal{S}^{-1}$：$|\mathcal{S}| = 18$，余子式矩阵为

$$\mathcal{S}^{-1} = \frac{1}{18}\begin{pmatrix} 11 & -7 & -1 \\ -7 & 11 & -1 \\ -1 & -1 & 5 \end{pmatrix} \tag{1.31}$$

```r
n5 <- 10; p5 <- 3
xbar5 <- c(1, 0, 2)
S5    <- matrix(c(3,2,1,2,3,1,1,1,4), 3, 3, byrow = TRUE)

cat("eigenvalues:", sort(eigen(S5)$values, decreasing=TRUE), "\n")
cat("|S| =", det(S5), "  tr(S) =", sum(diag(S5)), "\n")

S5_inv  <- solve(S5)
F_crit5 <- qf(0.95, p5, n5-p5)
T2_crit <- (n5-1)*p5/(n5-p5) * F_crit5

cat("S^{-1} =\n"); print(round(S5_inv, 6))
cat("F_crit(3,7,0.05) =", round(F_crit5, 4), "\n")
cat("T2_crit =", round(T2_crit, 4), "\n")
```

结论： $\boldsymbol{\mu}$ 的 95% 置信域为

$$\left\{\boldsymbol{\mu} \in \mathbb{R}^3 \;\middle|\; 10(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \mathcal{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \le 16.7663\right\} \tag{1.32}$$

即以 $\overline{\boldsymbol{x}} = (1, 0, 2)^\top$ 为中心、由 $\mathcal{S}^{-1}$ 决定形状的椭球体。

(b) [2 分] 计算 μ_1，μ_2，以及 μ_3 的联合置信区间.

【解】

μ_i 的联合（同时）置信区间公式（取 $\boldsymbol{a} = \boldsymbol{e}_i$）：

$$\bar{x}_i \pm \sqrt{T^2_\alpha} \cdot \sqrt{\frac{s_{ii}}{n}} \tag{1.33}$$

其中 $\sqrt{T^2_\alpha} = \sqrt{16.7663} = 4.0947$。

```r
factor5 <- sqrt(T2_crit)
cat("sqrt(T2_crit) =", round(factor5, 4), "\n\n")

for(i in 1:p5){
  se <- sqrt(S5[i,i] / n5)
  lb <- xbar5[i] - factor5 * se
  ub <- xbar5[i] + factor5 * se
  cat(paste0("mu_", i, ": [", round(lb,4), ", ", round(ub,4), "]\n"))
}
```

计算各分量：

$$\mu_1: \quad 1 \pm 4.0947 \times \sqrt{3/10} = 1 \pm 4.0947 \times 0.5477 = 1 \pm 2.2427 \Rightarrow [-1.2427,\ 3.2427]$$

$$\mu_2: \quad 0 \pm 4.0947 \times \sqrt{3/10} = 0 \pm 2.2427 \Rightarrow [-2.2427,\ 2.2427]$$

$$\mu_3: \quad 2 \pm 4.0947 \times \sqrt{4/10} = 2 \pm 4.0947 \times 0.6325 = 2 \pm 2.5898 \Rightarrow [-0.5898,\ 4.5898]$$

结论：三个均值分量的 95% 联合置信区间分别为 $\mu_1 \in [-1.2427,\ 3.2427]$，$\mu_2 \in [-2.2427,\ 2.2427]$，$\mu_3 \in [-0.5898,\ 4.5898]$。

(c) [2 分] 可否认为 μ_1 等于 μ_2 与 μ_3 的平均?

【解】

命题 "$\mu_1 = \dfrac{\mu_2 + \mu_3}{2}$" 等价于 $2\mu_1 - \mu_2 - \mu_3 = 0$，即检验

$$H_0: \boldsymbol{a}^\top \boldsymbol{\mu} = 0, \quad \boldsymbol{a} = (2,\ -1,\ -1)^\top \tag{1.34}$$

由于 $\boldsymbol{\Sigma}$ 未知，检验统计量为

$$T = \frac{\boldsymbol{a}^\top \overline{\boldsymbol{x}} - 0}{\sqrt{\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a}/n}} \sim t(n-1) = t(9) \tag{1.35}$$

计算 $\boldsymbol{a}^\top \overline{\boldsymbol{x}} = 2 \times 1 - 0 - 2 = 0$，从而 $T = 0$。

计算 $\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a}$：

$$\mathcal{S}\boldsymbol{a} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 6-2-1 \\ 4-3-1 \\ 2-1-4 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ -3 \end{pmatrix} \tag{1.36}$$

$$\boldsymbol{a}^\top \mathcal{S} \boldsymbol{a} = \begin{pmatrix} 2 & -1 & -1 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ -3 \end{pmatrix} = 6 + 0 + 3 = 9 \tag{1.37}$$

```r
a5c  <- c(2, -1, -1)
abar <- sum(a5c * xbar5)
aSa  <- as.numeric(t(a5c) %*% S5 %*% a5c)
T5c  <- abar / sqrt(aSa / n5)
t_crit5c <- qt(0.975, n5-1)
p_val5c  <- 2 * pt(-abs(T5c), df = n5-1)

cat("a'xbar =", abar, "\n")
cat("a'Sa =", aSa, "\n")
cat("T =", T5c, "\n")
cat("t_0.025(9) =", round(t_crit5c,4), "\n")
cat("p-value =", p_val5c, "\n")
cat("Reject H0?", abs(T5c) > t_crit5c, "\n")
```

结论： $T=0$，p 值 $=1$，$t_{0.025}(9)=2.2622$，在显著水平 0.05 下**不拒绝** H_0。样本数据与 $\mu_1=(\mu_2+\mu_3)/2$ 完全吻合（$\bar{x}_1=1=(0+2)/2$），因此**可以认为** μ_1 等于 μ_2 与 μ_3 的平均。

6. 对取自两个二元正态分布总体、容量均为 10 的两个独立样本，计算得

$$\bar{\boldsymbol{x}}_1=\begin{pmatrix}3\\1\end{pmatrix},\quad \mathcal{S}_1=\begin{pmatrix}4&-1\\-1&2\end{pmatrix},\quad \bar{\boldsymbol{x}}_2=\begin{pmatrix}1\\1\end{pmatrix},\quad \mathcal{S}_2=\begin{pmatrix}2&-2\\-2&4\end{pmatrix} \tag{1.38}$$

求解以下假设检验问题：

(a) [2 分] $H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \longleftrightarrow H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$.

【解】

方法： 设两总体协方差矩阵相等，采用 Hotelling 两样本 T^2 检验。合并协方差矩阵为

$$\mathcal{S}_p=\frac{(n_1-1)\mathcal{S}_1+(n_2-1)\mathcal{S}_2}{n_1+n_2-2}=\frac{\mathcal{S}_1+\mathcal{S}_2}{2}=\begin{pmatrix}3&-3/2\\-3/2&3\end{pmatrix} \tag{1.39}$$

$|\mathcal{S}_p|=9-9/4=27/4$，故

$$\mathcal{S}_p^{-1}=\frac{4}{27}\begin{pmatrix}3&3/2\\3/2&3\end{pmatrix}=\begin{pmatrix}4/9&2/9\\2/9&4/9\end{pmatrix} \tag{1.40}$$

$\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2=(2,0)^\top$，检验统计量为

$$\begin{aligned}T^2&=(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top\left[\left(\frac{1}{10}+\frac{1}{10}\right)\mathcal{S}_p\right]^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)\\&=(2,0)\cdot 5\mathcal{S}_p^{-1}\cdot\begin{pmatrix}2\\0\end{pmatrix}=(2,0)\cdot 5\begin{pmatrix}4/9\\2/9\end{pmatrix}\cdot 2=5\times\frac{4}{9}\times 4=\frac{80}{9}\approx 8.8889\end{aligned} \tag{1.41}$$

转化为 F 统计量（$p=2$，$n_1+n_2-p-1=17$）：

$$F=\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2=\frac{17}{18\times 2}\times\frac{80}{9}=\frac{17\times 80}{324}\approx 4.1975 \tag{1.42}$$

```r
n6_1 <- n6_2 <- 10; p6 <- 2
xbar6_1 <- c(3,1); xbar6_2 <- c(1,1)
S6_1 <- matrix(c(4,-1,-1,2), 2, 2)
S6_2 <- matrix(c(2,-2,-2,4), 2, 2)

Sp6   <- ((n6_1-1)*S6_1 + (n6_2-1)*S6_2) / (n6_1+n6_2-2)
diff6 <- xbar6_1 - xbar6_2
T2_6  <- as.numeric(t(diff6) %*% solve((1/n6_1+1/n6_2)*Sp6) %*% diff6)
F6    <- T2_6 * (n6_1+n6_2-p6-1) / ((n6_1+n6_2-2)*p6)
F_c6  <- qf(0.95, p6, n6_1+n6_2-p6-1)
p6a   <- 1 - pf(F6, p6, n6_1+n6_2-p6-1)

```

```
cat("Sp =\n"); print(Sp6)
cat("T2 =", round(T2_6,4), "\n")
cat("F =",  round(F6,4), "\n")
cat("F_crit(2,17,0.05) =", round(F_c6,4), "\n")
cat("p-value =", round(p6a,4), "\n")
cat("Reject H0?", F6 > F_c6, "\n")
```

结论： $F = 4.1975 > F_{0.05}(2,17) = 3.5915$，$p$ 值 $= 0.0330$，在显著水平 0.05 下**拒绝** H_0，即两总体均值向量存在显著差异。

(b) [2 分] $H_0: \mu_{11} = \mu_{21} \longleftrightarrow H_1: \mu_{11} \neq \mu_{21}$.

【解】

方法： 对第一分量进行单变量两样本 t 检验。合并方差为

$$s_p^2 = \frac{(n_1-1)s_{1,11} + (n_2-1)s_{2,11}}{n_1+n_2-2} = \frac{9\times 4 + 9\times 2}{18} = \frac{54}{18} = 3 \tag{1.43}$$

检验统计量为

$$T = \frac{\bar{x}_{11} - \bar{x}_{21}}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{3-1}{\sqrt{3}\times\sqrt{2/10}} = \frac{2}{\sqrt{3/5}} = \frac{2\sqrt{5}}{\sqrt{3}} = \sqrt{20/3} \approx 2.5820 \tag{1.44}$$

```
sp2_b <- ((n6_1-1)*S6_1[1,1] + (n6_2-1)*S6_2[1,1]) / (n6_1+n6_2-2)
T6b   <- (xbar6_1[1]-xbar6_2[1]) / sqrt(sp2_b*(1/n6_1+1/n6_2))
t_c6b <- qt(0.975, n6_1+n6_2-2)
p6b   <- 2*pt(-abs(T6b), df=n6_1+n6_2-2)

cat("sp^2 =", sp2_b, ",  sp =", round(sqrt(sp2_b),4), "\n")
cat("T =", round(T6b,4), "\n")
cat("t_0.025(18) =", round(t_c6b,4), "\n")
cat("p-value =", round(p6b,4), "\n")
cat("Reject H0?", abs(T6b) > t_c6b, "\n")
```

结论： $|T| = 2.5820 > t_{0.025}(18) = 2.1009$，$p$ 值 $= 0.0188$，在显著水平 0.05 下**拒绝** H_0，即第一分量均值存在显著差异。

(c) [2 分] $H_0: \mu_{12} = \mu_{22} \longleftrightarrow H_1: \mu_{12} \neq \mu_{22}$.

【解】

方法： 对第二分量进行单变量两样本 t 检验。合并方差为

$$s_p^2 = \frac{(n_1-1)s_{1,22} + (n_2-1)s_{2,22}}{n_1+n_2-2} = \frac{9\times 2 + 9\times 4}{18} = \frac{54}{18} = 3 \tag{1.45}$$

检验统计量为

$$T = \frac{\bar{x}_{12} - \bar{x}_{22}}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{1-1}{\sqrt{3}\times\sqrt{2/10}} = 0 \tag{1.46}$$

```
sp2_c <- ((n6_1-1)*S6_1[2,2] + (n6_2-1)*S6_2[2,2]) / (n6_1+n6_2-2)
T6c   <- (xbar6_1[2]-xbar6_2[2]) / sqrt(sp2_c*(1/n6_1+1/n6_2))
```

```r
t_c6c <- qt(0.975, n6_1+n6_2-2)
p6c   <- 2*pt(-abs(T6c), df=n6_1+n6_2-2)

cat("sp^2 =", sp2_c, "\n")
cat("T =", T6c, "\n")
cat("t_0.025(18) =", round(t_c6c,4), "\n")
cat("p-value =", p6c, "\n")
cat("Reject H0?", abs(T6c) > t_c6c, "\n")
```

结论： $T = 0$，p 值 $= 1$，在显著水平 0.05 下**不拒绝** H_0，第二分量均值无显著差异。

(d) [2 分] 比较上述结果并作简要分析.

【解】

三种检验的结论汇总如下：

检验	统计量	p 值	结论（$\alpha = 0.05$）
(a) Hotelling T^2（$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$）	$F = 4.1975$	0.0330	拒绝 H_0
(b) 一元 t（$\mu_{11} = \mu_{21}$）	$T = 2.5820$	0.0188	拒绝 H_0
(c) 一元 t（$\mu_{12} = \mu_{22}$）	$T = 0$	1.0000	不拒绝 H_0

分析：

- 两组的均值差异**完全来自第一分量**（$\bar{x}_{11} - \bar{x}_{21} = 2 \neq 0$），第二分量完全相同（$\bar{x}_{12} = \bar{x}_{22} = 1$）。
- 联合检验 (a) 正确地检测到了总体均值向量的差异，与分量检验 (b) 结论一致。
- 若只进行两次单变量 t 检验（不加调整），I 类错误概率会膨胀。Hotelling T^2 检验在保持整体显著性水平的同时，兼顾了变量间的相关结构（$\mathcal{S}_p$ 中含有协方差项），比分别检验更为严格。
- 本例中，联合检验 (a) 的 p 值（0.033）大于单变量检验 (b) 的 p 值（0.019），反映了多变量检验在保持整体犯错概率时的保守性。

7. [2 分] 对于课堂中讨论过的美国公司数据集，利用 $X_1 \sim X_6$ 的全部六个变量的观测数据，检验能源行业和制造业的协方差矩阵是否相等.

【解】

检验问题： $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \longleftrightarrow H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$。

方法： 采用 Box 的 M 检验。对 $g = 2$ 组，Box M 统计量为

$$M = \left(\sum_{i=1}^{g} \nu_i \right) \ln |\mathcal{S}_p| - \sum_{i=1}^{g} \nu_i \ln |\mathcal{S}_i|, \qquad \nu_i = n_i - 1 \tag{1.47}$$

修正统计量 $u=(1-c_1)M$ 近似服从 $\chi^2\left(\dfrac{p(p+1)(g-1)}{2}\right)$，其中

$$c_1=\left(\sum_{i=1}^{g}\frac{1}{\nu_i}-\frac{1}{\sum_i \nu_i}\right)\cdot\frac{2p^2+3p-1}{6(p+1)(g-1)} \tag{1.48}$$

自由度 $df=p(p+1)(g-1)/2=6\times7\times1/2=21$。

```r
# data from Q2 (energy, manuf, n1=8, n2=7, p=6)
S1 <- cov(energy);  S2 <- cov(manuf)
Sp <- ((n1-1)*S1 + (n2-1)*S2) / (n1+n2-2)
g  <- 2

M <- (n1+n2-2)*log(det(Sp)) - ((n1-1)*log(det(S1)) + (n2-1)*log(det(S2)))
cat("ln|Sp| =", round(log(det(Sp)),4), "\n")
cat("ln|S1| =", round(log(det(S1)),4), "\n")
cat("ln|S2| =", round(log(det(S2)),4), "\n")
cat("M =", round(M,4), "\n")

c1 <- (1/(n1-1) + 1/(n2-1) - 1/(n1+n2-2)) *
       (2*p^2 + 3*p - 1) / (6*(p+1)*(g-1))
u  <- (1 - c1) * M
df_box <- p*(p+1)*(g-1)/2
chi2_c <- qchisq(0.95, df_box)
p_val7 <- 1 - pchisq(u, df_box)

cat("c1 =", round(c1,6), "\n")
cat("u = (1-c1)*M =", round(u,4), "\n")
cat("df =", df_box, "\n")
cat("chi2_crit(0.05,21) =", round(chi2_c,4), "\n")
cat("p-value =", round(p_val7,6), "\n")
cat("Reject H0?", u > chi2_c, "\n")
```

计算结果：

$$\ln|\mathcal{S}_p|=59.1867,\quad \ln|\mathcal{S}_1|=50.0977,\quad \ln|\mathcal{S}_2|=44.9488$$

$$M=13\times59.1867-(7\times50.0977+6\times44.9488)=149.0500$$

$$c_1=\left(\frac{1}{7}+\frac{1}{6}-\frac{1}{13}\right)\cdot\frac{2\times36+18-1}{6\times7\times1}=0.23261\times2.11905=0.4929$$

$$u=(1-0.4929)\times149.05=75.5845$$

查 χ^2 分布表，$\chi^2_{0.05}(21)=32.6706$。

结论： $u=75.58\gg\chi^2_{0.05}(21)=32.67$，$p$ 值 ≈0，在显著水平 0.05 下**拒绝** H_0，即能源行业与制造业的协方差矩阵**不相等**。这与两行业内部公司规模差异显著（能源行业大公司与小公司之间的离散程度远大于制造业）的直觉相符。此结果表明，在第 2 题中采用合并协方差矩阵的 Hotelling T^2 检验的前提条件实际上并不满足。

8. 对于瑞士银行钞票数据集 (mclust 包中的 banknote 数据集) 当中的伪钞数据，我们想知道钞票对角线的长度 X_6 是否可以由 $X_1 \sim X_5$ 的一个线性模型来预测.

(a) [2 分] 拟合线性模型，给出拟合结果.

【解】

模型设定： 对伪钞子集（$n = 100$）拟合线性回归模型

$$X_6 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{1.49}$$

```r
library(mclust)
data(banknote)

# extract counterfeit subset
fake <- banknote[banknote$Status == "counterfeit", ]
cat("counterfeit n =", nrow(fake), "\n")
cat("variables:", names(fake), "\n\n")

# fit linear model
fit <- lm(Diagonal ~ Length + Left + Right + Bottom + Top, data = fake)
summary(fit)
```

运行上述代码，`summary(fit)` 输出包括：

- **回归系数表**（Coefficients）：各 $\hat{\beta}_j$、标准误 $\mathrm{se}(\hat{\beta}_j)$、$t$ 统计量及 p 值；
- **残差标准差** $\hat{\sigma}$（Residual standard error）；
- **决定系数** R^2 及调整 R^2（Multiple R-squared / Adjusted R-squared）；
- **整体** F **检验**（F-statistic）及其 p 值。

拟合结果的解读：R^2 衡量模型对 X_6 变异的解释比例；若 R^2 较高（接近 1），说明 $X_1 \sim X_5$ 对对角线长度有较强的线性预测能力。

(b) [2 分] 检验回归系数是否显著不等于零 (取显著水平 $\alpha = 0.05$).

【解】

整体显著性检验：

H_0：$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ $\longleftrightarrow$ H_1：至少一个 $\beta_j \neq 0$。

F 统计量为

$$F = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F(p,\, n-p-1) = F(5,\, 94) \quad (\text{在 } H_0 \text{ 下}) \tag{1.50}$$

$F_{0.05}(5, 94) = 2.3113$。若 $F > 2.3113$ 则整体显著。

各系数的个别检验：

对每个 β_j（$j = 1, \ldots, 5$）：H_0：$\beta_j = 0$ $\longleftrightarrow$ H_1：$\beta_j \neq 0$。

$$t_j = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)} \sim t(n-p-1) = t(94) \quad (\text{在 } H_0 \text{ 下}) \tag{1.51}$$

$t_{0.025}(94) = 1.9855$。若 $|t_j| > 1.9855$（或 p 值 < 0.05），则 β_j 显著不等于零。

```r
# overall F test
summary_fit <- summary(fit)
cat("F-statistic:", summary_fit$fstatistic, "\n")
cat("F_crit(5, 94, 0.05) =", round(qf(0.95, 5, 94), 4), "\n")
cat("overall p-value:", pf(summary_fit$fstatistic[1],
    summary_fit$fstatistic[2],
    summary_fit$fstatistic[3], lower.tail=FALSE), "\n\n")

# individual coefficient tests
cat("regression coefficients:\n")
print(summary_fit$coefficients)
cat("\nt_crit(0.025, 94) =", round(qt(0.975, 94), 4), "\n")

# check significance of each coefficient
coef_table <- summary_fit$coefficients
sig <- abs(coef_table[,3]) > qt(0.975, 94)
cat("significant (|t|>t_crit):\n")
print(sig)
```

判断准则（基于 $\alpha = 0.05$）：

- 若 `summary(fit)` 中 `F-statistic` 对应的 p 值 < 0.05，则整体回归显著，$X_1 \sim X_5$ 对 X_6 有显著线性预测作用；
- 对每个预测变量，`Pr(>t—)—` < 0.05 说明该系数显著不等于零，即对应变量在控制其他变量后对 X_6 有显著贡献；
- `Pr(>t—)—` ≥ 0.05 的变量系数在该显著水平下不显著，可考虑从模型中剔除。