Multivariate Analysis Homework 2

A49109720 Yi-Chen Zhang

March 25, 2018

5.1. (a) Evaluate T^2, for testing $H_0 : \boldsymbol{\mu}^T = (7,\ 11)$, using the data

$$\boldsymbol{X} = \begin{pmatrix} 2 & 12 \\ 8 & 9 \\ 6 & 9 \\ 8 & 10 \end{pmatrix}$$

(b) Specify the distribution of T^2 for the situation in (a).

(c) Using (a) and (b), test H_0 at the $\alpha = 0.05$ level. What conclusion do you reach?

Sol. (a) We have $n = 4$, $p = 2$, $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \begin{pmatrix} 6 \\ 10 \end{pmatrix}$, and

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)^T = \begin{pmatrix} 8 & -\frac{10}{3} \\ -\frac{10}{3} & 2 \end{pmatrix}. \ \Rightarrow\ \boldsymbol{S}^{-1} = \begin{pmatrix} \frac{9}{22} & \frac{15}{22} \\ \frac{15}{22} & \frac{18}{11} \end{pmatrix}$$

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) = 13.6364$$

(b) Under H_0, T^2 is distributed as $\dfrac{(n-1)p}{n-p}F_{p,n-p}$. That is,

$$T^2 \sim 3F_{2,2}.$$

(c) Using R we calculate $F_{2,2}(0.05) =$ `qf(1-0.05,df1=2,df2=2)` $= 19$.
Since $T^2 = 13.6364 < 3F_{2,2}(0.05) = 57$, we do not reject H_0 at significant level $\alpha = 0.05$.

5.2. Using the data in Example 5.1, verify that T^2 remains unchanged if each observations $\boldsymbol{x}_j$, $j = 1, 2, 3$, is replaced by $\boldsymbol{C}\boldsymbol{x}_j$, where

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

Note that the observations

$$\boldsymbol{C}\boldsymbol{x}_j = \begin{pmatrix} x_{j1} - x_{j2} \\ x_{j1} + x_{j2} \end{pmatrix}$$

yield the data matrix

$$\begin{pmatrix} (6-9) & (10-6) & (8-3) \\ (6+9) & (10+6) & (8+3) \end{pmatrix}^T$$

Sol. Let $\boldsymbol{z}_j = \boldsymbol{C}\boldsymbol{x}_j$ for $j = 1, 2, 3$ and $\boldsymbol{\mu}_{\boldsymbol{z},\mathbf{0}} = \boldsymbol{C}\boldsymbol{\mu}_{\mathbf{0}}$, then

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{C}\boldsymbol{x}_1 \\ \boldsymbol{C}\boldsymbol{x}_2 \\ \boldsymbol{C}\boldsymbol{x}_3 \end{pmatrix} = \begin{pmatrix} -3 & 15 \\ 4 & 16 \\ 5 & 11 \end{pmatrix} \text{ and } \boldsymbol{\mu}_{\boldsymbol{z},\mathbf{0}} = \boldsymbol{C}\boldsymbol{\mu}_0 = \begin{pmatrix} 4 \\ 14 \end{pmatrix}$$

$$\bar{\boldsymbol{z}} = \begin{pmatrix} 2 \\ 14 \end{pmatrix}, \quad \boldsymbol{S}_{\boldsymbol{z}} = \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^T = \begin{pmatrix} 19 & -5 \\ -5 & 7 \end{pmatrix} \text{ and } \boldsymbol{S}_{\boldsymbol{z}}^{-1} = \begin{pmatrix} \frac{7}{108} & \frac{5}{108} \\ \frac{5}{108} & \frac{19}{108} \end{pmatrix}$$

$$T^* = n(\bar{\boldsymbol{z}} - \boldsymbol{\mu}_{\boldsymbol{z},\mathbf{0}})^T \boldsymbol{S}_{\boldsymbol{z}}^{-1} (\bar{\boldsymbol{z}} - \boldsymbol{\mu}_{\boldsymbol{z},\mathbf{0}}) = \frac{7}{9} = T^2$$

Thus, T^2 remains unchanged.

5.9. Harry Roberts, a naturalist for the Alask Fish and Game departmen, studies grizzly bears with the goal of maintaining a healthy population. Measurements on $n = 61$ bears provided the following summary statistics:

Variable	Weight	Body	Neck	Girth	Head	Head
	(kg)	length	(cm)	(cm)	length	width
		(cm)			(cm)	(cm)
Sample mean $\bar{x}$	95.52	164.38	55.69	93.39	17.98	31.13

Covariance matrix

$$\boldsymbol{S} = \begin{pmatrix} 3266.46 & 1343.97 & 731.54 & 1175.50 & 162.68 & 238.37 \\ 1343.97 & 721.91 & 324.25 & 537.35 & 80.17 & 117.73 \\ 731.54 & 324.25 & 179.28 & 281.17 & 39.15 & 56.80 \\ 1175.50 & 537.35 & 281.17 & 474.98 & 63.73 & 94.85 \\ 162.68 & 80.17 & 39.15 & 63.73 & 9.95 & 13.88 \\ 238.37 & 117.73 & 56.80 & 94.85 & 13.88 & 21.26 \end{pmatrix}$$

(a) Obtain the large sample 95% simultaneous confidence intervals for the six population mean body measurements.

(b) Obtain the large sample 95% simultaneous ellipse for mean weight and mean girth.

(c) Obtain the 95% Bonferroni confidence intervals for the six means in Part (a).

(d) Refer to Part (b). Construct the 95% Bonferroni confidence rectangle for the mean weight and mean girth using $m = 6$. Compare this rectangle with the confidence ellipse in Part (b).

(e) Obtain the 95% Bonferroni confidence interval for

mean head width − mean head length

using $m = 6 + 1 = 7$ to allow for this statement as well as statements about each individual mean.

Sol. (a) One can use either Scheffe's (exact) or large sample (approximate) simultaneous confidence interval. Here we provide two solutions for this question. To construct Scheff's confidence interval we use

$$\boldsymbol{a}^T \bar{\boldsymbol{X}} \pm \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{\boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}}{n}}$$

and to construct large sample confidence interval we use

$$\boldsymbol{a}^T\bar{\boldsymbol{X}} \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}}{n}}.$$

The large sample result is from the Result 5.5 in the textbook at page 235. We note here these two confidence intervals will be very close because of the fact that $\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$ and $\chi_p^2(\alpha)$ are approximately equal for n large relative to p.
The above two intervals will contain $\boldsymbol{a}^T\boldsymbol{\mu}$, for every $\boldsymbol{a}$, with probability approximately $100(1-\alpha)\%$. Since $n = 61$, $p = 6$, and $\alpha = 0.05$, the value of $F_{p,n-p}(\alpha)$ is

$$F_{p,n-p}(\alpha) = \texttt{qf(1-0.05,df1=6,df2=61-6)} = 2.2687$$

and the value of $\chi_p^2(\alpha)$ is

$$\chi_p^2(\alpha) = \texttt{qchisq(1-0.05,df=6)} = 12.5916.$$

The critical value of Scheff's method is $\sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)} = 3.8535$ and that of large sample method is $\sqrt{\chi_p^2(\alpha)} = 3.5484$.
The $100(1-\alpha)\%$ Scheff's simultaneous confidence interval for the six population mean body measurements are:

$$\bar{x}_1 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} = 95.52 \pm 3.8535\sqrt{\frac{3266.46}{61}}$$

$$\Rightarrow\ 67.3210 \le \mu_1 \le 123.7190$$

$$\bar{x}_2 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} = 164.38 \pm 3.8535\sqrt{\frac{721.91}{61}}$$

$$\Rightarrow\ 151.1233 \le \mu_2 \le 177.6367$$

$$\bar{x}_3 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{33}}{n}} = 55.69 \pm 3.8535\sqrt{\frac{179.28}{61}}$$

$$\Rightarrow\ 49.0837 \le \mu_3 \le 62.2963$$

$$\bar{x}_4 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{44}}{n}} = 93.39 \pm 3.8535\sqrt{\frac{474.98}{61}}$$

$$\Rightarrow\ 82.6369 \le \mu_4 \le 104.1431$$

$$\bar{x}_5 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{55}}{n}} = 17.98 \pm 3.8535\sqrt{\frac{9.95}{61}}$$

$$\Rightarrow\ 16.4237 \le \mu_5 \le 19.5364$$

$$\bar{x}_6 \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{66}}{n}} = 31.13 \pm 3.8535\sqrt{\frac{21.26}{61}}$$

$$\Rightarrow\ 28.8550 \le \mu_6 \le 33.4050$$

The $100(1-\alpha)\%$ large sample simultaneous confidence for the six population mean body measurements are:

$$\bar{x}_1 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{11}}{n}} = 95.52 \pm \sqrt{12.5916}\sqrt{\frac{3266.46}{61}} \quad \Rightarrow \quad 69.5535 \le \mu_1 \le 121.4865$$

$$\bar{x}_2 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{22}}{n}} = 164.38 \pm \sqrt{12.5916}\sqrt{\frac{721.91}{61}} \quad \Rightarrow \quad 152.1728 \le \mu_2 \le 176.5872$$

$$\bar{x}_3 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{33}}{n}} = 55.69 \pm \sqrt{12.5916}\sqrt{\frac{179.28}{61}} \quad \Rightarrow \quad 49.6067 \le \mu_3 \le 61.7733$$

$$\bar{x}_4 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{44}}{n}} = 93.39 \pm \sqrt{12.5916}\sqrt{\frac{474.98}{61}} \quad \Rightarrow \quad 83.4882 \le \mu_4 \le 103.2918$$

$$\bar{x}_5 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{55}}{n}} = 17.98 \pm \sqrt{12.5916}\sqrt{\frac{9.95}{61}} \quad \Rightarrow \quad 16.5469 \le \mu_5 \le 19.4131$$

$$\bar{x}_6 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{66}}{n}} = 31.13 \pm \sqrt{12.5916}\sqrt{\frac{21.26}{61}} \quad \Rightarrow \quad 29.0351 \le \mu_6 \le 33.2249$$

(b) From Result 5.3, for Scheff's method the $100(1-\alpha)\%$ simultaneous ellipse for (μ_i, μ_k) belongs to the sample mean-centered ellipses

$$n(\bar{x}_i - \mu_i, \bar{x}_k - \mu_k)\begin{pmatrix} s_{ii} & s_{ik} \\ s_{ki} & s_{kk} \end{pmatrix}^{-1}\begin{pmatrix} \bar{x}_i - \mu_i \\ \bar{x}_k - \mu_k \end{pmatrix} \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha),$$

and for large sample method the $100(1-\alpha)\%$ simultaneous ellipse for (μ_i, μ_k) belongs to the sample mean-centered ellipses

$$n(\bar{x}_i - \mu_i, \bar{x}_k - \mu_k)\begin{pmatrix} s_{ii} & s_{ik} \\ s_{ki} & s_{kk} \end{pmatrix}^{-1}\begin{pmatrix} \bar{x}_i - \mu_i \\ \bar{x}_k - \mu_k \end{pmatrix} \le \chi_p^2(\alpha).$$

We apply these two results and plot the 95% Scheff's simultaneous ellipse and large sample simultaneous ellipses for mean weight and mean girth in Figure 1.

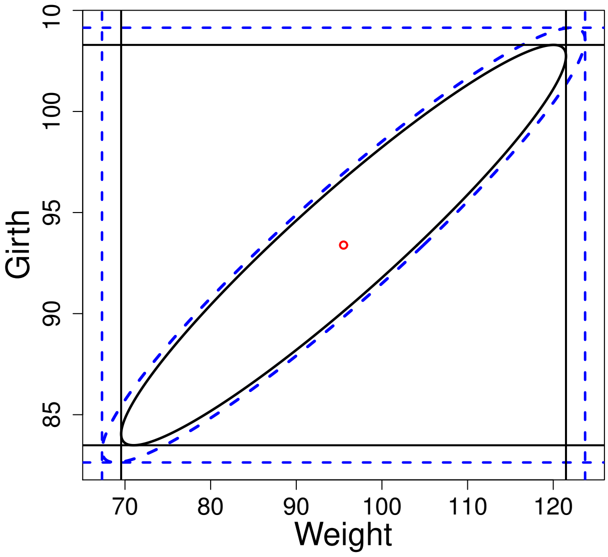


Figure 1: The 95% simultaneous ellipse with confidence rectangle. Scheff: dotted blue line; Large sample: solid black line.

(c) The Bonferroni $100(1-\alpha)\%$ confidence interval for μ_i (see page 232 for details) is

$$\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}, \quad \text{for} \quad i = 1, \ldots, p.$$

Since $n = 61$, $p = 6$, and $\alpha = 0.05$, the critical value of $t_{n-1}\left(\frac{\alpha}{2p}\right)$ is

$$t_{n-1}\left(\tfrac{\alpha}{2p}\right) = \texttt{qt(1-0.05/(2*6),df=61-1)} = 2.7286.$$

The Bonferroni confidence interval for the six population mean body measurements are:

$$\bar{x}_1 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} = 95.52 \pm 2.7286\sqrt{\frac{3266.46}{61}} \quad \Rightarrow \quad 75.5533 \le \mu_1 \le 115.4867$$

$$\bar{x}_2 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} = 164.38 \pm 2.7286\sqrt{\frac{721.91}{61}} \quad \Rightarrow \quad 154.9934 \le \mu_2 \le 173.7666$$

$$\bar{x}_3 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{33}}{n}} = 55.69 \pm 2.7286\sqrt{\frac{179.28}{61}} \quad \Rightarrow \quad 51.0123 \le \mu_3 \le 60.3677$$

$$\bar{x}_4 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{44}}{n}} = 93.39 \pm 2.7286\sqrt{\frac{474.98}{61}} \quad \Rightarrow \quad 85.7761 \le \mu_4 \le 101.0039$$

$$\bar{x}_5 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{55}}{n}} = 17.98 \pm 2.7286\sqrt{\frac{9.95}{61}} \quad \Rightarrow \quad 16.8780 \le \mu_5 \le 19.0820$$

$$\bar{x}_6 \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{66}}{n}} = 31.13 \pm 2.7286\sqrt{\frac{21.26}{61}} \quad \Rightarrow \quad 29.5192 \le \mu_6 \le 32.7408$$

(d) The 95% Bonferroni confidence rectangle for the mean weight and mean girth with the confidence ellipse in Part (b) are plotted in Figure 2.

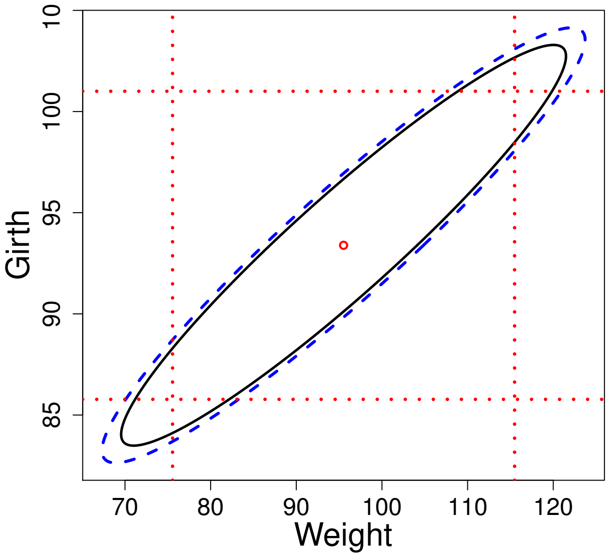


Figure 2: The 95% simultaneous ellipse with Bonferroni confidence rectangle.

From Figure 2 the Bonferroni confidence rectangle seems shorter than the Scheff's and large sample's simultaneous ellipse. This result is not surprising since we know the Bonferroni's method is more conservative.

(e) The 95% Bonferroni confidence interval for linear combinations $\boldsymbol{a}^T\boldsymbol{\mu}$ (see page 234 for details) is

$$\boldsymbol{a}^T\bar{\boldsymbol{X}} \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}}{n}}$$

The difference for mean head width − mean head length is $\mu_6 - \mu_5$. Let $\boldsymbol{a}^T = (0, 0, 0, , 0, -1, 1)$, then the Bonferroni confidence interval for $\mu_6 - \mu_5$ is

$$(\bar{x}_6 - \bar{x}_5) \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{55} - s_{56} - s_{65} + s_{66}}{n}}.$$

We note that the critical value is

$$t_{n-1}\left(\tfrac{\alpha}{2m}\right) = \texttt{qt(1-0.05/(2*7),df=n-1)} = 2.7855.$$

So the 95% Bonfferoni confidence interval for $\mu_6 - \mu_5$ is

$$(31.13 - 17.98) \pm 2.7855\sqrt{\frac{9.95 - 13.88 - 13.88 + 21.26}{61}}$$
$$\Rightarrow 12.4876 \le \mu_6 - \mu_5 \le 13.8124$$

6.8. Observations on two responses are collected for three treatments. The observation vectors $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ are

$$\begin{aligned}
\text{Treatment 1}: &\ \begin{pmatrix} 6 \\ 7 \end{pmatrix}, \begin{pmatrix} 5 \\ 9 \end{pmatrix}, \begin{pmatrix} 8 \\ 6 \end{pmatrix}, \begin{pmatrix} 4 \\ 9 \end{pmatrix}, \begin{pmatrix} 7 \\ 9 \end{pmatrix} \\
\text{Treatment 2}: &\ \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix} \\
\text{Treatment 3}: &\ \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}
\end{aligned}$$

(a) Break up the observations into mean, treatment, and residual components, as in (6-39). Construct the corresponding arrays for each variable.

(b) Using the information in Part (a), construct the one-way MANOVA table.

(c) Evaluate Wilks' lambda, Λ^*, and use Table 6.3 to test for treatment effects. Set $\alpha = 0.01$. Repeat the test using the chi-square approximation with Bartlett's correction. Compare the conclusions.

Sol. (a) We calculate the mean of each treatment and the overall mean

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 6 \\ 8 \end{pmatrix}, \ \bar{\boldsymbol{x}}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \ \bar{\boldsymbol{x}}_3 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad \text{and} \quad \bar{\boldsymbol{x}} = \begin{pmatrix} 4 \\ 5 \end{pmatrix},$$

and break up the observations into mean, treatment, and residual components as $\boldsymbol{x}_{lj} = \bar{\boldsymbol{x}} + (\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}}) + (\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)$. The first variable can be decomposed as

$$\left(\begin{array}{ccccc} 6 & 5 & 8 & 4 & 7 \\ 3 & 1 & 2 & & \\ 2 & 5 & 3 & 2 & \end{array}\right) = \left(\begin{array}{ccccc} 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & & \\ 4 & 4 & 4 & 4 & \end{array}\right) + \left(\begin{array}{ccccc} 2 & 2 & 2 & 2 & 2 \\ -2 & -2 & -2 & & \\ -1 & -1 & -1 & -1 & \end{array}\right) + \left(\begin{array}{ccccc} 0 & -1 & 2 & -2 & 1 \\ 1 & -1 & 0 & & \\ -1 & 2 & 0 & -1 & \end{array}\right)$$

and the second variable can be decomposed as

$$\left(\begin{array}{ccccc} 7 & 9 & 6 & 9 & 9 \\ 3 & 6 & 3 & & \\ 3 & 1 & 1 & 3 & \end{array}\right) = \left(\begin{array}{ccccc} 5 & 5 & 5 & 5 & 5 \\ 5 & 5 & 5 & & \\ 5 & 5 & 5 & 5 & \end{array}\right) + \left(\begin{array}{ccccc} 3 & 3 & 3 & 3 & 3 \\ -1 & -1 & -1 & & \\ -3 & -3 & -3 & -3 & \end{array}\right) + \left(\begin{array}{ccccc} -1 & 1 & -2 & 1 & 1 \\ -1 & 2 & -1 & & \\ 1 & -1 & -1 & 1 & \end{array}\right)$$

(b) For first variable we have

$$\begin{aligned}
SS_{\text{obs}} &= 6^2+5^2+8^2+4^2+7^2+3^2+1^2+2^2+2^2+5^2+3^2+2^2 = 246\\
SS_{\text{mean}} &= 12\times 4^2 = 192\\
SS_{\text{trt}} &= 5\times 2^2 + 3\times(-2)^2 + 4\times(-1)^2 = 36\\
SS_{\text{res}} &= 0^2+(-1)^2+2^2+(-2)^2+1^2+1^2+(-1)^2+0^2+(-1)^2+2^2+0^2+(-1)^2 = 18
\end{aligned}$$

The corrected total sum of square is $SS_{\text{total}} = SS_{\text{obs}} - SS_{\text{mean}} = 246 - 192 = 54$. Repeating this process on the second variable, we have

$$\begin{aligned}
SS_{\text{obs}} &= 7^2+9^2+6^2+9^2+9^2+3^2+6^2+3^2+3^2+1^2+1^2+3^2 = 402\\
SS_{\text{mean}} &= 12\times 5^2 = 300\\
SS_{\text{trt}} &= 5\times 3^2 + 3\times(-1)^2 + 4\times(-3)^2 = 84\\
SS_{\text{res}} &= (-1)^2+1^2+(-2)^2+1^2+1^2+(-1)^2+2^2+(-1)^2+1^2+(-1)^2+(-1)^2+1^2 = 18
\end{aligned}$$

The corrected total sum of square is $SS_{\text{total}} = SS_{\text{obs}} - SS_{\text{mean}} = 402 - 300 = 102$. The cross product terms are:

$$\begin{aligned}
SS_{\text{obs}} &= 6\times 7+5\times 9+8\times 6+4\times 9+7\times 9+3\times 3+1\times 6+2\times 3+2\times 3+\\
&\quad 5\times 1+3\times 1+2\times 3 = 275\\
SS_{\text{mean}} &= 12\times 4\times 5 = 240\\
SS_{\text{trt}} &= 5\times 2\times 3+3\times(-2)\times(-1)+4\times(-1)\times(-3) = 48\\
SS_{\text{res}} &= 0\times(-1)+(-1)\times 1+2\times(-2)+(-2)\times 1+1\times 1+1\times(-1)+(-1)\times 2+\\
&\quad 0\times(-1)+(-1)\times 1+2\times(-1)+0\times(-1)+(-1)\times 1 = -13
\end{aligned}$$

The corrected total sum of square is $SS_{\text{total}} = SS_{\text{obs}} - SS_{\text{mean}} = 275 - 240 = 35$.

The one-way MANOVA table

Source of variation	Sum of square matrix	Degrees of freedom
Treatment	$\boldsymbol{B} = \begin{pmatrix} 36 & 48 \\ 48 & 84 \end{pmatrix}$	$3-1=2$
Residual	$\boldsymbol{W} = \begin{pmatrix} 18 & -13 \\ -13 & 18 \end{pmatrix}$	$5+3+4-3=9$
Total (correted)	$\boldsymbol{B}+\boldsymbol{W} = \begin{pmatrix} 54 & 35 \\ 35 & 102 \end{pmatrix}$	$12-1=11$

(c) To test the treatment effects, we make the following hypothesis

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3 = \mathbf{0} \;\text{ vs. }\; H_1 : \text{at least one } \boldsymbol{\tau}_l \neq \mathbf{0}$$

To carry out the test, we calculate Wilks' lambda

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{W}+\boldsymbol{B}|} = \frac{155}{4283} = 0.0362.$$

From Table 6.3 for $p=2$ and $g=3$ the test statistic is

$$\left(\frac{\sum_{l=1}^{g} n_l - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) = 17.0266$$

For $\alpha = 0.01$ the critical value of $F_{2(g-1),2(\sum_{l=1}^{g} n_l - g - 1)}(\alpha)$ is

$$F_{4,16}(0.01) = \texttt{qf(1-0.01,df1=4,df2=16)} = 4.7726$$

Since $17.0266 > 4.7726$, we reject H_0 at significant level $\alpha = 0.01$ and conclude that at least one $\boldsymbol{\tau}_l$ is not zero.

To carry out the chi-square approximation with Barlett's correction, we caluclate the test statistic

$$-\left(n - 1 - \frac{p+g}{2}\right)\log \Lambda^* = 28.2114$$

For $\alpha = 0.01$ the critical value of $\chi^2_{p(g-1)}(\alpha)$ is

$$\chi^2_4(0.01) = \texttt{qchisq(1-0.01,df=4)} = 13.2767$$

Since $28.2114 > 13.2767$, we reject H_0 at significant level $\alpha = 0.01$ and conclude that at least one $\boldsymbol{\tau}_l$ is not zero.

Both of the tests show that the treatment difference exists. We note here the Wilk's lambda can be applied for small sample size, however, the Bartlett's correction is an approximation for large sample size.

6.12. (Test for linear profiles, given that the profiles are parallel.) Let $\boldsymbol{\mu}_1^T = (\mu_{11}, \mu_{12}, \ldots, \mu_{1p})$ and $\boldsymbol{\mu}_2^T = (\mu_{21}, \mu_{22}, \ldots, \mu_{2p})$ be the mean responses to p treatments for populations 1 and 2, respectively. Assum that the profiles given by the two mean vectors are paralle.

(a) Show that the hypothesis that the profiles are linear can be written as $H_0 : (\mu_{1i} + \mu_{2i}) - (\mu_{1,i-1} + \mu_{2,i-1}) = (\mu_{1,i-1} + \mu_{2,i-1}) - (\mu_{1,i-2} + \mu_{2,i-2})$, $i = 3, \ldots, p$ or as $H_0 : \boldsymbol{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$, where the $(p-2) \times p$ matrix

$$\boldsymbol{C} = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \end{pmatrix}$$

(b) Following an argument similar to the one leading to (6-73), we reject $H_0 : \boldsymbol{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$ at level α if

$$T^2 = (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)^T \boldsymbol{C}^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{C}\boldsymbol{S}_{\text{pooled}}\boldsymbol{C}^T\right]^{-1} \boldsymbol{C}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) > c^2$$

where

$$c^2 = \frac{(n_1 + n_2 - 2)(p-2)}{n_1 + n_2 - p + 1} F_{p-2, n_1+n_2-p+1}(\alpha)$$

Let $n_1 = 30$, $n_2 = 30$, $\bar{\boldsymbol{x}}_1^T = (6.4, 6.8, 7.3, 7.0)^T$, $\bar{\boldsymbol{x}}_2^T = (4.3, 4.9, 5.3, 5.1)$, and

$$\boldsymbol{S}_{\text{pooled}} = \begin{pmatrix} 0.61 & 0.26 & 0.07 & 0.16 \\ 0.26 & 0.64 & 0.17 & 0.14 \\ 0.07 & 0.17 & 0.81 & 0.03 \\ 0.16 & 0.14 & 0.03 & 0.31 \end{pmatrix}$$

Test for linear profiles, assuming that the profiles are parallel. Use $\alpha = 0.05$.

Sol. (a) Given that the profiles are parallel, then one will be above the other for all $i = 1, \ldots, p$, that is

$$\mu_{1i} > \mu_{2i} \ \ (\text{or } \ \mu_{1i} < \mu_{2i}) \ \text{ for all } \ i = 1, \ldots, p$$

So, profiles will be linear only if the increment of the sum of two treatments from i to $i+1$ is the same as the increment of that from $i+1$ to $i+2$, for all $i = 1, \dots, p-2$. Thus, we might consider looking at the difference of these increment of the sum of two treatments:

$$(\mu_{1,i+2} + \mu_{2,i+2}) - (\mu_{1,i+1} + \mu_{2,i+1}) = (\mu_{1,i+1} + \mu_{2,i+1}) - (\mu_{1i} + \mu_{2,i})$$

for $i = 1, \dots, p-2$. This is equivalent to test

$$H_0 : (\mu_{1i} + \mu_{2i}) - (\mu_{1,i-1} + \mu_{2,i-1}) = (\mu_{1,i-1} + \mu_{2,i-1}) - (\mu_{1,i-2} + \mu_{2,i-2})$$

for $i = 3, \dots, p$. We can also rewrite the above hypothesis as

$$H_0 : \boldsymbol{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = 0$$

where $\boldsymbol{C}$ is a $(p-2) \times p$ constant matrix,

$$\boldsymbol{C} = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}$$

(b) The test statistic $T^2 = 16.8361$ is caluclated by R. The code is compiled in the Appendix. Moreover, the quantile $F_{p-2,n_1+n_2-p+1}(\alpha) =$ `qf(1-0.05,df1=2,df2=57)` $=$ 3.1588, and the critical vale c^2 is

$$c^2 = \frac{(n_1 + n_2 - 2)(p-2)}{n_1 + n_2 - p + 1} F_{p-2,n_1+n_2-p+1}(\alpha) = \frac{58 \times 2}{57} \times 3.1588 = 6.4285$$

Since $T^2 = 16.8361 > c^2 = 6.4285$, we reject the H_0 at significant level $\alpha = 0.05$ and conclude that the profiles are not linear, given that the profiles are parallel. The profile picture plotted in Figure 3 is also indicated that the profiles are not linear.

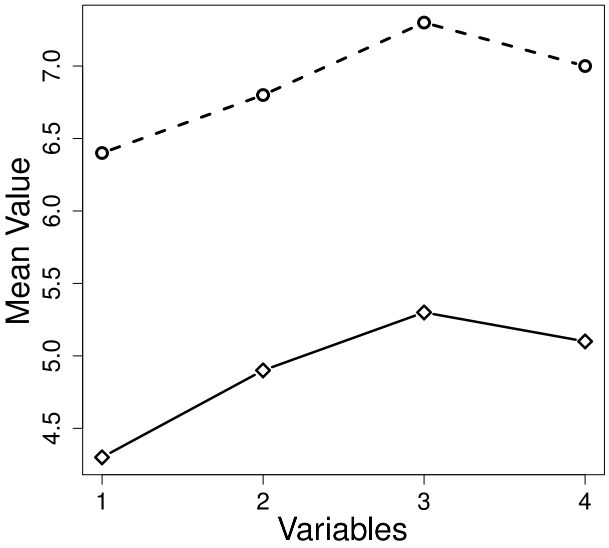


Figure 3: Profile analysis for two treatments.

6.20. The tail lengths in millimeters (x_1) and wing lengths in millimeters (x_2) for 45 male hook-billed kites are given in Table 6.11 on page 346. Similar measurements for female hook-billed kites were given in Table 5.12.

(a) Plot the male hook-billed kite data as a scatter diagram, and (visually) check for outliers. (Note, in particular, observation 31 with $x_1 = 284$.)

(b) Test for equality of mean vectors for the populations of male and female hook-billed kites. Set $\alpha = 0.05$. If $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ is rejected, find the linear combination most responsible for the rejection of H_0. (You may want to eliminate any outliers found in Part (a) for the male hook-billed kite data before conducting this test. Alternatively, you may want to interpret $x_1 = 284$ for observation 31 as a misprint and conduct the test with $x_1 = 184$ for this observation. Does it make any difference in this case how observation 31 for the male hook-billed kite data is treated?)

(c) Determine the 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and 95% simultaneous confidence intervals for the components of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

(d) Are male or female birds generally larger?

Sol. (a) The scatter plot for the male hook-billed kite is plotted in Figure 4. It is clear that the observation 31 with $x_1 = 281$ is an outlier.

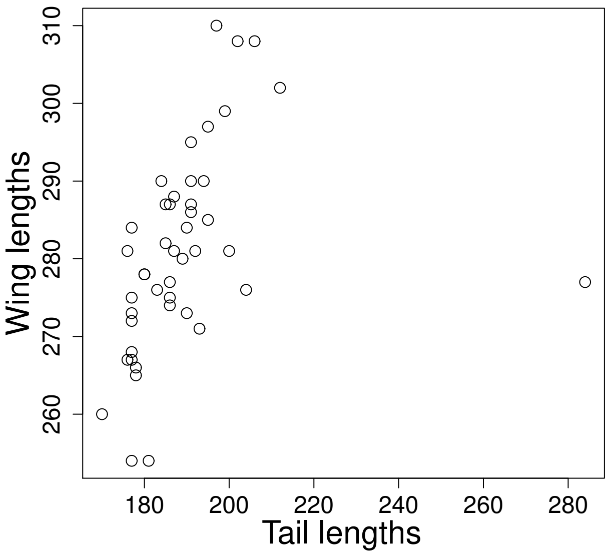


Figure 4: The scatter plot for the male hook-billed kite data

(b) We remove the observation 31 in male hook-billed kite data, since it is identifided as an outlier. We run Box's M-test for homogeneity of covariance matrices. The test statistic: $-2\rho \log \Lambda$ has an approximate χ^2_ν distribution with degrees of freedom ν, where

$$\rho = 1 - \left(\frac{2p^2 + 3p - 1}{6(p+1)(g-1)}\right)\left(\sum_{l=1}^{g} \frac{1}{(n_l - 1)} - \frac{1}{\sum_{l=1}^{g}(n_l - 1)}\right),$$

$$\Lambda = \prod_{l=1}^{g} \left(\frac{|\boldsymbol{S}_l|}{|\boldsymbol{S}_{\text{pooled}}|}\right)^{\frac{n_l - 1}{2}}, \quad \text{and} \quad \nu = \frac{1}{2}p(p+1)(g-1).$$

We found that $-2\rho \log \Lambda = 1.0431$ and $\nu = 3$, and the p-value is 0.7908. So we do not reject the null hypothesis and conclude the covariance matrices of male and female hook-billed kite are the same. It is reasonable to pool the covariance matrices of male and female hook-billed kite together and we denote it by

$$\boldsymbol{S}_{\text{pooled}} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$$

To test for equality of mean vectors for the populations of male and female hook-billed kites, we make the following hypothesis:

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq 0$$

From Result 6.2 at page 286, we calculate the test statistics

$$T^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{S}_{\text{pooled}}\right]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = 24.9649,$$

the quantile $F_{p,n_1+n_2-p-1}(\alpha) = \texttt{qf(1-0.05,df1=2,df2=86)} = 3.1026$, and the critical value

$$\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}(\alpha) = 6.2773$$

Since $T^2 = 24.9649 > 6.2773$, we reject H_0 and conclude that the male and female hook-billed kite population mean vectors are not equal.

For testing $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, the linear combination $\hat{\boldsymbol{a}}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, with coefficient vector $\hat{\boldsymbol{a}} \propto \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, quantifies the largest population difference. The linear combination of mean components most responsible for rejecting H_0 is given by the vector (see the remark on page 289, which builds on the argument at the top of page 225 in the textbook):

$$\hat{\boldsymbol{a}} \propto \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = \begin{pmatrix} -3.4842 \\ 2.0785 \end{pmatrix}.$$

Alternatively, we revise the observation 31 by changing it from $x_1 = 284$ to $x_1 = 184$ and repeat the whole processes. We found that for Box's M-test $-2\rho \log \Lambda = 1.2223$ and $\nu = 3$, and p-value is 0.7477, so the covariance matrices of male and female hook-billed kite are the same. To conduct the hypothesis for equality of mean vectors, we found the test statistic is $T^2 = 25.6625$ and the critical value is 6.2739. Hence, we still conclude that the male and female hook-billed kite population mean vectors are not equal.

There is no difference in this case how observation 31 for the male hook-billed kite data is treated. The R code for testing homogeneity of covariance matrices and for equality of mean vectors without (and with revised) observation 31 is compiled in the Appendix. In the following questions we only consider the data set with observation 31 being deleted.

(c) From Result 6.2 at page 286 in textbook, we have

$$\begin{aligned} T^2 &= \Big((\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Big)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{S}_{\text{pooled}}\right]^{-1} \Big((\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Big) \\ &\sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}. \end{aligned}$$

A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is

$$\left\{\Big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)\Big)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{\text{pooled}}\right]^{-1} \Big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)\Big) \le c^2\right\},$$

where

$$c^2 = \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}(\alpha)$$

The confidence region is an ellipse with center at $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ and the axes of the ellipse are

$$\pm\sqrt{\lambda_i}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2}\boldsymbol{e}_i$$

where λ_i and $\boldsymbol{e}_i$ are eigenvalues and eigenvectors of $\boldsymbol{S}_{\text{pooled}}$.

Here the 95% confidence ellipse for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is determined from the eigenvalue-eigenvector pairs $\lambda_1 = 262.5640$, $\boldsymbol{e}_1^T = (0.5588, 0.8293)$ and $\lambda_2 = 33.0469$, $\boldsymbol{e}_2^T = (-0.8293, 0.5588)$.

Since

$$\sqrt{\lambda_1}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2} = \sqrt{262.5640}\sqrt{\left(\frac{1}{44} + \frac{1}{45}\right) 6.2773} = 8.6072$$

and

$$\sqrt{\lambda_2}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2} = \sqrt{33.0469}\sqrt{\left(\frac{1}{44} + \frac{1}{45}\right) 6.2773} = 3.0536$$

we obtain the 95% confidence ellipse for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ sketched in Figure 5.

Moreover, from Result 6.3 a $100(1-\alpha)\%$ simultaneous confidence intervals for the components of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is

$$\boldsymbol{a}^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \pm c\sqrt{\boldsymbol{a}^T\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{\text{pooled}}\boldsymbol{a}}$$

will cover $\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all $\boldsymbol{a}$. In particular $\mu_{1i} - \mu_{2i}$ will be covered by

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{\text{pooled},ii}} \quad \text{for} \quad i = 1, \ldots, p$$

With $\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T = (\mu_{11} - \mu_{21}, \mu_{12} - \mu_{22})$, the 95% simultaneous confidence intervals for the population difference are

$$\mu_{11} - \mu_{21}: \ (187.1951 - 193.6222) \pm \sqrt{6.2773}\sqrt{\left(\frac{1}{44} + \frac{1}{45}\right) 104.7180}$$

or

$$-11.8989 \le \mu_{11} - \mu_{21} \le -1.0274$$

$$\mu_{12} - \mu_{22}: \ (280.9545 - 279.7778) \pm \sqrt{6.2773}\sqrt{\left(\frac{1}{44} + \frac{1}{45}\right) 109.8930}$$

or

$$-6.1623 \le \mu_{21} - \mu_{22} \le 8.5159$$

The 95% simultaneous confidence intervals are also sketched in Figure 5.

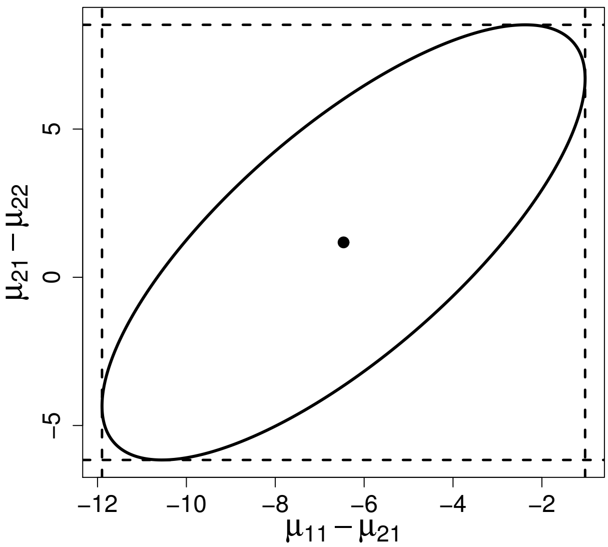


Figure 5: The 95% simultaneous ellipse with confidence rectangle.

(d) We conclude that there is a difference in tail lengths between male and female hook-billed kite. Female birds are general larger than male birds.

6.24. Researchers have suggested that a change in skull size over time is evidence of the interbreeding of a resident population with immigrant populations. Four measurements were made of male Egyptian skulls for three different time periods: period 1 is 4000 B.C., period 2 is 3300 B.C., and period 3 is 1850 B.C. The data are shown in Table 6.13 on page 349 (see the skull data on the website http://www.prenhall.com/statistics). The measured variables are

$$\begin{aligned} X_1 &= \text{maximum breadth of skull (mm)} \\ X_2 &= \text{basibregmatic height of skull (mm)} \\ X_3 &= \text{basialveolar length of skull (mm)} \\ X_4 &= \text{nasal height of skull (mm)} \end{aligned}$$

Construct a one-way MANOVA of the Egyptian slull data. Use $\alpha = 0.05$. Construct 95% simultaneous confidence intervals to determine which mean components differ among the populations represented by the three time periods. Are the usual MANOVA assumptions realistic for these data? Explain.

Sol. The one-way MANOVA model is specified by

$$\boldsymbol{X}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{e}_{lj}, \quad \text{for} \quad j = 1, \ldots, n_l \quad \text{and} \quad l = 1, \ldots, g$$

where the $\boldsymbol{e}_{lj}$ are independent $N(\mathbf{0}, \boldsymbol{\Sigma})$ variables. Here the parameter vector $\boldsymbol{\mu}$ is an over all mean (level), and $\boldsymbol{\tau}_l$ represents the l-th time period effect with $\sum_{l=1}^{g} n_l \boldsymbol{\tau}_l = \mathbf{0}$.

The hypothesis of no time period effects is tested by considering the relative size of the time effect and residual sums of squares and cross products.

$$H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \quad \text{vs.} \quad H_1 : \text{ at least one } \boldsymbol{\tau} \neq \mathbf{0}$$

We summarize the calculations leading to the test statistic in a MANOVA table:

Source	Matrix of sum of squares	Degrees of freedom
Time effect	$\boldsymbol{B} = \sum_{l=1}^{g} n_l(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})^T$	$g-1$
Residual	$\boldsymbol{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^T$	$\sum_{l=1}^{g} n_l - g$
Total	$\boldsymbol{B}+\boldsymbol{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})^T$	$\sum_{l=1}^{g} n_l - 1$

We reject H_0 if the ratio of generalized variances

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$$

is too small. The exact distribution of Λ^* can be derived for the special case listed in Table 6.3 in the textbook at page 303.

Since $p = 4$ and $g = 3$, the distribution of Wilks' Lambda Λ^* is

$$\left(\frac{\sum_{l=1}^{g} n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(\sum_{l=1}^{g} n_l - p - 2)}$$

Using `manova` function in R, we get $\Lambda^* = 0.8301$.

The test statistics is

$$\left(\frac{\sum_{l=1}^{g} n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) = \left(\frac{90-4-2}{4}\right)\left(\frac{1-\sqrt{0.8301}}{\sqrt{0.8301}}\right) = 2.0491$$

and the critical value is

$$F_{2p,2(\sum_{l=1}^{g} n_l - p - 2)}(\alpha) = \texttt{qf(1-0.05,df1=8,df2=168)} = 1.9939$$

Since the test statistics $2.0491 > 1.9939$, we reject H_0 and conclude that the time effect differences exist. There is a difference of male Egyptian skulls for three different time periods.

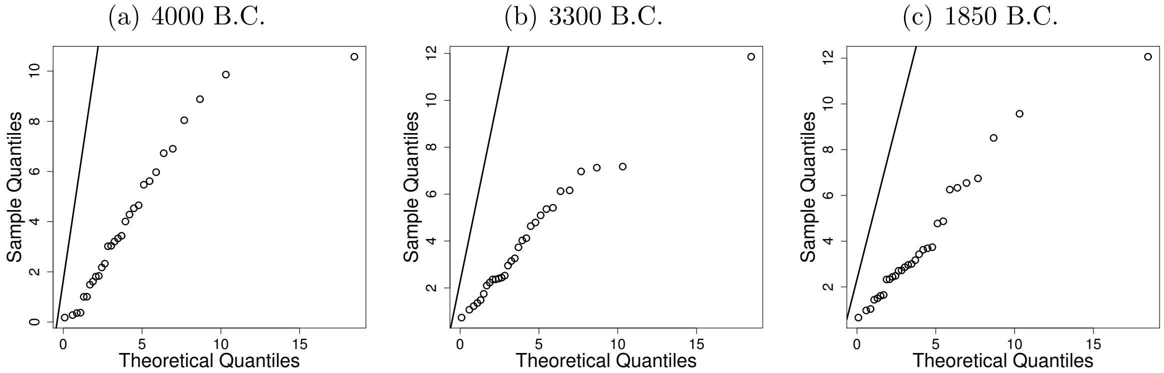
For pairwise comparisons, the Bonferroni approach can be used to construct simultaneous confidence intervals for the components of the differences $\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$. From Result 6.5, For the MANOVA model, with confidence level at $100(1-\alpha)\%$

$$\tau_{ki} - \tau_{li} \text{ belongs to } \bar{x}_{ki} - \bar{x}_{li} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}$$

The calculation of above equation for each pair is calculated by R and compiled in Appendix.

$\tau_{11} - \tau_{21} \in (-4.4423, 2.4423)$, $\tau_{11} - \tau_{31} \in (-6.5423, 0.3423)$, $\tau_{21} - \tau_{31} \in (-5.5423, 1.3423)$
$\tau_{12} - \tau_{22} \in (-2.6737, 4.4737)$, $\tau_{12} - \tau_{32} \in (-3.7737, 3.3737)$, $\tau_{22} - \tau_{32} \in (-4.6737, 2.4737)$
$\tau_{13} - \tau_{23} \in (-3.6801, 3.8801)$, $\tau_{13} - \tau_{33} \in (-0.6468, 6.9134)$, $\tau_{23} - \tau_{33} \in (-0.7468, 6.8134)$
$\tau_{14} - \tau_{24} \in (-2.0614, 2.6614)$, $\tau_{14} - \tau_{34} \in (-2.3948, 2.3281)$, $\tau_{24} - \tau_{34} \in (-2.6948, 2.0281)$

For $\alpha = 0.05$ we find all simultaneous confidence intervals cover zero, indicating that there is no significant difference between three different time periods. We further investigate the skull data and find that the normality assumption is violated for each period (see Figure 6). Hencn we conclude that the usual MANOVA assumptions are not realistic for these data.



(a) 4000 B.C.

(b) 3300 B.C.

(c) 1850 B.C.

Figure 6: QQ-plot for period 1, 2, and 3 from left to right.

6.37. Table 6.9 page 344 contains the carapace measurements for 24 female and 24 male turtles. Use Box's M-test to test $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}_1$ is the population covariance matrix for carapace measurements for female turtles, and $\boldsymbol{\Sigma}_2$ is the population covariance matrix for carapace measurements for male turtles. Set $\alpha = 0.05$.

Sol. We run Box's M-test for homogeneity of covariance matrices. The R code is compiled in the Appendix. We found that the test statistic is $-2\rho \log \Lambda = 23.405$ with degrees of freedom $v = 6$ and the p-value is 0.0007. So we reject H_0 at significant level $\alpha = 0.05$ and conclude that there is a difference of population covariance between male and female turtles.

6.39. Anacondas are some of the largest snakes in the world. Jesus Ravis and his fellow, researchers capture a snake and measure its (i) snout vent length (cm) or the length from the snout of the snake to its vent where it evacuaies waste and (ii) weight (kilograms). A sample of these measurements in shown in Table 6.19.

(a) Test for equality of means between males and females using $\alpha = 0.05$. Appiy the large sample statistic.

(b) Is it reasonable to pool variances in this case? Explain.

(c) Find the 95% Boneferroni confidence intervals for the mean differences between males and females on both length and weight.

Sol. (a) We first run Box M-test to test the equality of covariance matrices between male and female. The p-value is reported to be 0. Hence, we conclude that the covariance matrices between male and female are different.

To test for equality of means we make the following hypothesis:

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}$$

From Result 6.4 at page 292 in the textbook, under the large sample setting to test for equality of means we calculate the test statistics

$$T^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \left(\frac{1}{n_1} \boldsymbol{S}_1 + \frac{1}{n_2} \boldsymbol{S}_2 \right)^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = 76.9153,$$

and the critical value $\chi^2_p(\alpha) = 5.9915$.
Since $T^2 = 76.9153 > 5.9915$, we reject H_0 and conclude that the means for at least one of the variables between males and females are significantly different.

(b) Though the covariance matrices of female and male groups are significantly different, the sample sizes for the two groups are equal, so it is still reasonable to pool the variance. We note that under H_0, if $n_1 = n_2 = n$, then $(n-1)/(n+n-2) = 1/2$, so

$$\frac{1}{n_1}\boldsymbol{S}_1+\frac{1}{n_2}\boldsymbol{S}_2 = \frac{1}{n}(\boldsymbol{S}_1+\boldsymbol{S}_2) = \frac{(n-1)\boldsymbol{S}_1 + (n-1)\boldsymbol{S}_2}{n+n-2}\left(\frac{1}{n}+\frac{1}{n}\right) = \boldsymbol{S}_{\text{pooled}}\left(\frac{1}{n}+\frac{1}{n}\right).$$

With equal sample sizes, the large sample procedure is essentially the same as the procedure based on the pooled covariance matrix.

(c) The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for the p population mean differences (see page 291 in the textbook) are

$$\mu_{1i} - \mu_{2i} :\ (\bar{x}_{1i} - \bar{x}_{2i}) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{\text{pooled},ii}}$$

the critical value

$$t_{n_1+n_2-2}\left(\tfrac{\alpha}{2p}\right) = \texttt{qt(1-0.05/(2*2),df=54)} = 2.3056$$

The 95% Bonferroni confidence interval for the mean differences between males and females on length is $(-150.98, -88.06)$, and the 95% Bonferroni confidence interval for the mean differences between males and females on weight is $(-38.78, -21.17)$.

Appendix

R code for Problem 5.1.

```
> # (a)
> mu <- c(7,11)
> n <- 4
> p <- 2
> X <- matrix(c(2,8,6,8,12,9,9,10),nrow=n,ncol=p)
> xbar <- colMeans(X)
> S <- cov(X)
> Sinv <- solve(S)
> T2 <- n*(xbar-mu) %*% Sinv %*% (xbar-mu)
>
> # (c)
> alpha <- 0.05
> qf(1-alpha,df1=2,df2=2)
```

R code for Problem 5.2.

```
> n <- 3
> p <- 2
> mu <- c(9,5)
> X <- matrix(c(6,10,8,9,6,3),nrow=n,ncol=p)
> xbar <- colMeans(X)
> S <- cov(X)
> Sinv <- solve(S)
> T2 <- n*(xbar-mu) %*% Sinv %*% (xbar-mu)
>
```

```
> C <- matrix(c(1,1,-1,1),nrow=2,ncol=2)
> Z <- t(C%*%t(X))
> zbar <- colMeans(Z)
> muz <- C%*%mu
> Sz <- cov(Z)
> Szinv <- solve(Sz)
>
> Tstar <- n*t(zbar-muz) %*% Szinv %*% (zbar-muz)
>
> # (c)
> alpha <- 0.05
> qf(1-alpha,df1=2,df2=2)
```

R code for Problem 5.9.

```
> xbar <- c(95.52, 164.38, 55.69, 93.39, 17.98, 31.13)
> S <- matrix(c(3266.46, 1343.97, 731.54, 1175.50, 162.68, 238.37,
+              1343.97,  721.91, 324.25,  537.35,  80.17, 117.73,
+               731.54,  324.25, 179.28,  281.17,  39.15,  56.80,
+              1175.50,  537.35, 281.17,  474.98,  63.73,  94.85,
+               162.68,   80.17,  39.15,   63.73,   9.95,  13.88,
+               238.37,  117.73,  56.80,   94.85,  13.88,  21.26),
+             nrow=6, ncol=6, byrow=TRUE)
> p <- 6
> n <- 61
> alpha <- 0.05
>
> # (a) Scheff's
> SH.qlevel <- qf(1-alpha,df1=p,df2=n-p)
> for ( i in 1:p ){
+   SH.LCI <- xbar[i]-sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[i,i]/n)
+   SH.UCI <- xbar[i]+sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[i,i]/n)
+   print(c(SH.LCI, SH.UCI))
+ }
> # (a) Large sample
> qlevel <- qchisq(1-alpha,df=p)
> for ( i in 1:p ){
+   LCI <- xbar[i]-sqrt(qlevel)*sqrt(S[i,i]/n)
+   UCI <- xbar[i]+sqrt(qlevel)*sqrt(S[i,i]/n)
+   print(c(LCI, UCI))
+ }
>
> # (b)
> center <- xbar[c(1,4)]
> Sn2 <- S[c(1,4),c(1,4)]
> npoints <- 1000
> theta <- seq(0, 2*pi, length = npoints)
>
> # transform for points on ellipse for Scheff's
> SH.r <- sqrt((n-1)*p/(n-p)*SH.qlevel/n)
> SH.v <- rbind(SH.r*cos(theta), SH.r*sin(theta))
> SH.z <- backsolve(chol(solve(Sn2)),SH.v)+center
>
> # calculate the 95% simultaneous confidence interval
> SH.LCIx <- xbar[1]-sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[1,1]/n)
> SH.UCIx <- xbar[1]+sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[1,1]/n)
> SH.LCIy <- xbar[4]-sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[4,4]/n)
> SH.UCIy <- xbar[4]+sqrt((n-1)*p/(n-p)*SH.qlevel)*sqrt(S[4,4]/n)
>
> # transform for points on ellipse for large sample
> r <- sqrt(qlevel/n)
> v <- rbind(r*cos(theta), r*sin(theta))
> z <- backsolve(chol(solve(Sn2)),v)+center
>
> # calculate the 95% simultaneous confidence interval
> LCIx <- xbar[1]-sqrt(qlevel)*sqrt(S[1,1]/n)
> UCIx <- xbar[1]+sqrt(qlevel)*sqrt(S[1,1]/n)
> LCIy <- xbar[4]-sqrt(qlevel)*sqrt(S[4,4]/n)
> UCIy <- xbar[4]+sqrt(qlevel)*sqrt(S[4,4]/n)
>
> # plot the ellipse for Scheff's
> plot(t(SH.z), type='l', xlab='Weight', ylab='Girth', lty=2, col='blue')
>
> # plot 95% simultaneous confidence interval
> abline(v=SH.LCIx, col='blue')
> abline(v=SH.UCIx, col='blue')
> abline(h=SH.LCIy, col='blue')
> abline(h=SH.UCIy, col='blue')
>
> # plot the ellipse for large sample
> lines(t(z))
>
> # plot 95% simultaneous confidence interval
> abline(v=LCIx)
> abline(v=UCIx)
> abline(h=LCIy)
> abline(h=UCIy)
>
> # plot center of ellipse
> points(center[1], center[2], col='red')
>
> # (c)
> BF.qlevel <- qt(1-alpha/(2*p),df=n-1)
> for ( i in 1:p ){
+   BF.LCI <- xbar[i]-BF.qlevel*sqrt(S[i,i]/n)
+   BF.UCI <- xbar[i]+BF.qlevel*sqrt(S[i,i]/n)
+   print(c(BF.LCI,BF.UCI))
+ }
>
> # (d)
> # calculate the 95% simultaneous confidence interval for Bonferroni method
> BF.LCIx <- xbar[1]-BF.qlevel*sqrt(S[1,1]/n)
> BF.UCIx <- xbar[1]+BF.qlevel*sqrt(S[1,1]/n)
> BF.LCIy <- xbar[4]-BF.qlevel*sqrt(S[4,4]/n)
> BF.UCIy <- xbar[4]+BF.qlevel*sqrt(S[4,4]/n)
> plot(t(SH.z), type='l', xlab='Weight', ylab='Girth', lty=2, col='blue')
> lines(t(z))
> points(center[1], center[2], col='red')
>
> # plot 95% simultaneous confidence interval
> abline(v=BF.LCIx, lty=3, col='red')
> abline(v=BF.UCIx, lty=3, col='red')
> abline(h=BF.LCIy, lty=3, col='red')
> abline(h=BF.UCIy, lty=3, col='red')
>
> # (e)
> a <- c(0,0,0,0,-1,1)
> m <- 7
> dqlevel <- qt(1-alpha/(2*m),df=n-1)
> dLCI <- a%*%xbar-dqlevel*sqrt(a%*%S%*%a/n)
> dUCI <- a%*%xbar+dqlevel*sqrt(a%*%S%*%a/n)
```

R code for Problem 6.8.

```
> x1 <- c(6,5,8,4,7,3,1,2,2,5,3,2)
> x2 <- c(7,9,6,9,9,3,6,3,3,1,1,3)
> trt <- as.factor(c(1,1,1,1,1,2,2,2,3,3,3,3))
> mfit <- lm(cbind(x1,x2)~trt)
> mvafit <- manova(mfit)
> summary(mva, test='Wilk')
```

R code for Problem 6.12.

```
> alpha <- 0.05
> p <- 4
> n1 <- 30
> n2 <- 30
> xbar1 <- c(6.4,6.8,7.3,7.0)
> xbar2 <- c(4.3,4.9,5.3,5.1)
> Sp <- matrix(c(0.61,0.26,0.07,0.16,0.26,0.64,0.17,0.14,0.07,
+                0.17,0.81,0.03,0.16,0.14,0.03,0.31),nrow=4,ncol=4)
> C <- matrix( c(1,0,-2,1,1,-2,0,1), nrow=2, ncol=4)
>
> mudiff <- C%*%(xbar1+xbar2)
> Sinv <- solve((1/n1+1/n2)*C%*%Sp%*%t(C))
> T2 <- t(mudiff) %*% Sinv %*% mudiff
> c2 <- (n1+n2-2)*(p-2)/(n1+n2-p+1)*qf(1-alpha,df1=p-2,df2=n1+n2-p+1)
>
> # plot the profile
> ymax <- max(xbar1,xbar2)
> ymin <- min(xbar1,xbar2)
> plot(1:4,xbar1,type='b',lty=2,ylim=c(ymin,ymax),xlab='Variables',
+      ylab='Mean Value',xaxt='n')
```

```
> lines(1:4,xbar2,type=’b’,pch=23)
> axis(side=1,at=1:4)
```

R code for Problem 6.20.

```
> # (a) load the data and plot it
> male <- read.table(’./T6-11.dat’)
> colnames(male) <- c(’Tail lengths’,’Wing lengths’)
> female <- read.table(’./T5-12.dat’)
> colnames(female) <- c(’Tail lengths’,’Wing lengths’)
> plot(male)
>
> # (b) test equality of mean
> library(’biotools’)
> # number of variables and alpha
> p <- 2
> alpha <- 0.05
>
> # combine two dataset
> n1 <- dim(male)[1]
> n2 <- dim(female)[1]
> bird <- rbind(male,female)
> bird$gender <- c(rep(’male’,n1),rep(’female’,n2))
>
> # remove obs. 31
> bird <- bird[-31,]
> n1 <- n1-1
> box <- boxM(bird[,-3], bird[,3])
>
> # test for equality of mean vector
> xbar1 <- colMeans(bird[bird$gender==’male’,-3])
> xbar2 <- colMeans(bird[bird$gender==’female’,-3])
> S1 <- cov(bird[bird$gender==’male’,-3])
> S2 <- cov(bird[bird$gender==’female’,-3])
> Sp <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> Spinv <- solve((1/n1+1/n2)*Sp)
>
> # test statistic
> T2 <- (xbar1-xbar2)%*%Spinv%*%(xbar1-xbar2)
>
> # critical value
> c2 <- (n1+n2-2)*p/(n1+n2-p-1)*qf(1-alpha,df1=p,df2=n1+n2-p-1)
>
> # Alternative method combine two data set
> n1 <- dim(male)[1]
> n2 <- dim(female)[1]
> bird <- rbind(male,female)
> bird$gender <- c(rep(’male’,n1),rep(’female’,n2))
>
> # change observation 31 from x1=284 to x1=184
> bird[31,1] <- 184
```

```
>
> # Run Box’s M-test
> box <- boxM(bird[,-3], bird[,3])
>
> # test for equality of mean vector
> xbar1 <- colMeans(bird[bird$gender==’male’,-3])
> xbar2 <- colMeans(bird[bird$gender==’female’,-3])
> S1 <- cov(bird[bird$gender==’male’,-3])
> S2 <- cov(bird[bird$gender==’female’,-3])
> Sp <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> Spinv <- solve((1/n1+1/n2)*Sp)
>
> # test statistic
> T2 <- (xbar1-xbar2)%*%Spinv%*%(xbar1-xbar2)
>
> # critical value
> c2 <- (n1+n2-2)*p/(n1+n2-p-1)*qf(1-alpha,df1=p,df2=n1+n2-p-1)
>
> # (c) Determine the 95% confidence region for mu1-mu2 and
> # simultaneous confidence intervals for the components of mu1-mu2
> center <- xbar1-xbar2
> npoints <- 1000
> theta <- seq(0, 2*pi, length = npoints)
> qlevel <- qf(1-alpha,df1=p,df2=n1+n2-p-1)
>
> r <- sqrt((1/n1+1/n2)*(n1+n2-2)*p/(n1+n2-p-1)*qlevel)
> v <- rbind(r*cos(theta), r*sin(theta))
> z <- backsolve(chol(solve(Sp)),v)+center
>
> # calculate the 95% simultaneous confidence interval
> LCIx <- center[1]-r*sqrt(Sp[1,1])
> UCIx <- center[1]+r*sqrt(Sp[1,1])
> LCIy <- center[2]-r*sqrt(Sp[2,2])
> UCIy <- center[2]+r*sqrt(Sp[2,2])
>
> # plot the ellipse for Scheff’s
> plot(t(z),type=’l’,xlab=expression(mu[11]-mu[21]),ylab=expression(mu[21]-mu[22]))
> points(center[1], center[2],pch=19)
>
> # plot 95% simultaneous confidence interval
> abline(v=LCIx, lty=2)
> abline(v=UCIx, lty=2)
> abline(h=LCIy, lty=2)
> abline(h=UCIy, lty=2)
```

R code for Problem 6.24.

```
> skull <- read.table(’./T6-13.dat’)
> colnames(skull) <- c(’MB’,’BH’,’BL’,’NH’,’Period’)
> skull$Period <- as.factor(skull$Period)
> n <- dim(skull)[1]
> p <- dim(skull[,-5])[2]
> mfit <- lm(cbind(MB,BH,BL,NH) ~ Period, data=skull)
> mvafit <- manova(mfit)
> mfitsummary <- summary(mvafit)
>
> B <- mfitsummary$SS$Period
> W <- mfitsummary$SS$Residuals
>
> # Wilk's Lambda
> Lambda <- det(W)/(det(B+W))
> FV <- ((n-p-2)/p)*((1-sqrt(Lambda))/sqrt(Lambda))
> alpha <- 0.05
> qf(1-alpha,df1=2*p,df2=2*(n-p-2))
>
> # or one can do this
> summary(mvafit,test='Wilk')
>
> # pair comparison
> g <- 3
> n1 <- length(which((skull$Period==1)))
> n2 <- length(which((skull$Period==2)))
> n3 <- length(which((skull$Period==3)))
> n <- n1+n2+n3
> xbar1 <- colMeans(skull[skull$Period==1,-5])
> xbar2 <- colMeans(skull[skull$Period==2,-5])
> xbar3 <- colMeans(skull[skull$Period==3,-5])
> xbar <- (n1*xbar1+n2*xbar2+n3*xbar3)/(n1+n2+n3)
> S1 <- cov(skull[skull$Period==1,-5])
> S2 <- cov(skull[skull$Period==2,-5])
> S3 <- cov(skull[skull$Period==3,-5])
> W <- (n1-1)*S1+(n2-1)*S2+(n3-1)*S3
> qtlevel <- qt(1-alpha/(p*g*(g-1)),df=n-g)
> for ( i in 1:p ){
+   # \tau_{11}-\tau_{21}
+   LCI12 <- (xbar1[i]-xbar2[i])-qtlevel*sqrt(W[i,i]/(n-g)*(1/n1+1/n2))
+   UCI12 <- (xbar1[i]-xbar2[i])+qtlevel*sqrt(W[i,i]/(n-g)*(1/n1+1/n2))
+   cat("tau1[",i,"]-tau2[",i,"] belongs to (",LCI12,",",UCI12,")\n",sep="")
+
+   # \tau_{11}-\tau_{31}
+   LCI13 <- (xbar1[i]-xbar3[i])-qtlevel*sqrt(W[i,i]/(n-g)*(1/n1+1/n3))
+   UCI13 <- (xbar1[i]-xbar3[i])+qtlevel*sqrt(W[i,i]/(n-g)*(1/n1+1/n3))
+   cat("tau1[",i,"]-tau3[",i,"] belongs to (",LCI13,",",UCI13,")\n",sep="")
+
+   # \tau_{21}-\tau_{31}
+   LCI23 <- (xbar2[i]-xbar3[i])-qtlevel*sqrt(W[i,i]/(n-g)*(1/n2+1/n3))
+   UCI23 <- (xbar2[i]-xbar3[i])+qtlevel*sqrt(W[i,i]/(n-g)*(1/n2+1/n3))
+   cat("tau2[",i,"]-tau3[",i,"] belongs to (",LCI23,",",UCI23,")\n",sep="")
+ }
>
> # Check normality for each group
> S1inv <- solve(S1)
> skull1 <- skull[skull$Period==1,-5]
> datachisq <- diag(t(t(skull1)-xbar1) %*% S1inv %*% (t(skull1)-xbar1))
> qqplot(qchisq(ppoints(500),df=p),datachisq, main="",
+        xlab="Theoretical Quantiles",ylab="Sample Quantiles")
> qqline(datachisq,distribution=function(p) qchisq(p, df = p))
>
> S2inv <- solve(S2)
> skull2 <- skull[skull$Period==2,-5]
> datachisq <- diag(t(t(skull2)-xbar2) %*% S2inv %*% (t(skull2)-xbar2))
> qqplot(qchisq(ppoints(500),df=p),datachisq,main="",
+        xlab="Theoretical Quantiles",ylab="Sample Quantiles")
> qqline(datachisq,distribution=function(p) qchisq(p, df = p))
>
> S3inv <- solve(S3)
> skull3 <- skull[skull$Period==3,-5]
> datachisq <- diag(t(t(skull3)-xbar3) %*% S3inv %*% (t(skull3)-xbar3))
> qqplot(qchisq(ppoints(500),df=p),datachisq,main="",
+        xlab="Theoretical Quantiles",ylab="Sample Quantiles")
> qqline(datachisq,distribution=function(p) qchisq(p, df = p))
```

R code for Problem 6.37.

```
> turtle <- read.table('./T6-9.dat')
>
> library('biotools')
> box <- boxM(turtle[,-4], turtle[,4])
```

R code for Problem 6.39.

```
> anacondas <- read.table('./T6-19.dat')
> colnames(anacondas) <- c('length', 'weight', 'sex')
> # test equal variance
> library('biotools')
> box <- boxM(anacondas[,-3], anacondas[,3])
>
> # (a)
> p <- 2
> alpha <- 0.05
> n1 <- dim(anacondas[anacondas$sex=='M',])[1]
> n2 <- dim(anacondas[anacondas$sex=='F',])[1]
> xbar1 <- colMeans(anacondas[anacondas$sex=='M',-3])
> xbar2 <- colMeans(anacondas[anacondas$sex=='F',-3])
> S1 <- cov(anacondas[anacondas$sex=='M',-3])
> S2 <- cov(anacondas[anacondas$sex=='F',-3])
> Sp <- S1/n1+S2/n2
> Spinv <- solve(Sp)
>
> # test statistic
> T2 <- (xbar1-xbar2) %*% Spinv %*% (xbar1-xbar2)
>
```

```
> # critical value
> qchisq(1-alpha,df=2)
>
> # or one can simply apply MANOVA model
> fit.lm <- lm(cbind(length,weight)˜sex, data=anacondas)
> fit.manova <- manova(fit.lm)
> summary(fit.manova, test="Wilks")
>
> # (c)
> Sp <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> qlevel <- qt(1-alpha/(2*p),df=n1+n2-2)
> LCIx <- (xbar1[1]-xbar2[1])-qlevel*sqrt((1/n1+1/n2)*Sp[1,1])
> UCIx <- (xbar1[1]-xbar2[1])+qlevel*sqrt((1/n1+1/n2)*Sp[1,1])
> LCIy <- (xbar1[2]-xbar2[2])-qlevel*sqrt((1/n1+1/n2)*Sp[2,2])
> UCIy <- (xbar1[2]-xbar2[2])+qlevel*sqrt((1/n1+1/n2)*Sp[2,2])
```