Formative Assessment 3

Baybayon, Darlyn Antoinette B.

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(knitr)
  library(kableExtra)
  library(readxl)
  library(MVN)
  library(Hotelling)
  library(dplyr)
  library(heplots)
  library(MASS)
})
```

Flea Beetles Dataset

```
df1 <- suppressMessages(read_excel("Flea Beetles.xlsx",skip=1))
beetle1 <- df1[-20, 2:5]
beetle2 <- df1[, 7:10]
names(beetle1) <- c("Y1","Y2","Y3","Y4")
names(beetle2) <- c("Y1","Y2","Y3","Y4")
beetle1$Species <- "H. oleracea"
beetle2$Species <-  "H. carduorum"

beetles<- bind_rows(beetle1, beetle2)
beetles$Species <- factor(beetles$Species)
head(beetles)
```

```
## # A tibble: 6 x 5
##      Y1    Y2    Y3    Y4 Species
##   <dbl> <dbl> <dbl> <dbl> <fct>
## 1   189   245   137   163 H. oleracea
## 2   192   260   132   217 H. oleracea
## 3   217   276   141   192 H. oleracea
## 4   221   299   142   213 H. oleracea
## 5   171   239   128   158 H. oleracea
## 6   192   262   147   173 H. oleracea
```

(a) **Find the discriminant function coefficient vector and test significance of separation.**

Compute group mean vectors

```
xbar1 <- colMeans(beetle1[, 1:4])
xbar2 <- colMeans(beetle2[, 1:4])
```

Compute group covariance matrices

```
S1 <- cov(beetle1[, 1:4])
S2 <- cov(beetle2[, 1:4])
```

```
boxM(cbind(Y1, Y2, Y3, Y4)~ Species, data = beetles)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  beetles
## Chi-Sq (approx.) = 8.7457, df = 10, p-value = 0.5564
```

The Box'sM test did not reveal a significant difference (p>0.05) in covariance matrices across the treatment groups. Therefore, the assumption of homogeneity of covariance matrices is met.

Pooled within-group covariance matrix $S_p = \frac{(n_1-1)S_1+(n_2)S_2}{n_1+n_2-2}$

```
n1 <- nrow(beetle1)
n2 <- nrow(beetle2)

Sp <- ((n1-1) * S1 + (n2-1) * S2) / (n1+n2-2)
kable(Sp)
```

	Y1	Y2	Y3	Y4
Y1	143.55910	151.8034	42.52660	71.99253
Y2	151.80341	367.7878	121.87653	106.24467
Y3	42.52660	121.8765	118.31408	42.06401
Y4	71.99253	106.2447	42.06401	208.07290

Compute $a = S_p^{-1}(\bar{x_1} - \bar{x_2})$

```
a <- solve(Sp, xbar1-xbar2)
kable(a, col.names = "a")
```

	a
Y1	0.3452490
Y2	-0.1303878
Y3	-0.1064338
Y4	-0.1433533

Test significance of separation using Hotelling's T^2

```
diff <- xbar1 - xbar2

T2 <- (n1*n2) / (n1+n2) * t(diff) %*% solve(Sp) %*% diff
T2 <- as.numeric(T2)
p=ncol(beetles)-1

df1 <- p
df2 <- n1 + n2 - p - 1

F_stat <- (df2*T2) / (df1*(n1+n2- 2))
```

```r
p_val <- pf(F_stat, df1, df2, lower.tail = FALSE)

kable(data.frame(
  T2 = T2,
  F_stat = F_stat,
  df1 = df1, df2 = df2,
  p_val = p_val
))
```

T2	F_stat	df1	df2	p_val
133.4873	30.666	4	34	0

Using hotelling.test() from Hotellings library

```r
htest <- hotelling.test(beetle1[,1:4], beetle2[,1:4])
print(htest)
```

```
## Test stat:  133.49
## Numerator df:  4
## Denominator df:  34
## P-value:  7.522e-11
```

Since $p < 0.05$, T^2 is statistically significant and it follows that the discriminant function coefficient, a, is significantly different from 0.

(b) **Find the standardized coefficients.**

```r
within_group_sds <- sqrt(diag(Sp))
kable(within_group_sds * a, col.names = "std. a")
```

	std. a
Y1	4.136640
Y2	-2.500550
Y3	-1.157705
Y4	-2.067833

Hence, we have the following discriminant function:

$$z = 4.137y_1 - 2.5y_2 - 1.158y_3 - 2.068y_4$$

By rank of absolute values, Y1 is the most significant variable. Y2 and Y4 have roughly similar contributions to the separation, while Y3 has the least.

Alternatively, linear discriminant analysis may be performed using the MASS package in R.

```r
beetle.lda <- lda(Species ~ ., data = beetles)
kable(beetle.lda$scaling)
```

	LD1
Y1	0.0932764
Y2	-0.0352271
Y3	-0.0287554
Y4	-0.0387300

The values are different from the results of manual calculations because the lda() function uses another scaling method. Nevertheless, it remains that Y1 is the most significant variable, then Y2 and Y4, and Y3.

(c) **Calculate t-tests for individual variables.**

```
vars <- c("Y1", "Y2", "Y3", "Y4")
ttest_results <- p_vals <- numeric(length(vars))

for (i in 1:length(vars)) {
  var <- vars[i]
  test <- t.test(beetle1[[var]], beetle2[[var]])
  ttest_results[i] <- test$statistic
  p_vals[i] <- test$p.value
}

kable(data.frame(variable = vars, t = ttest_results, p_val = p_vals))
```

variable	t	p_val
Y1	3.857721	0.0005024
Y2	-3.871290	0.0004251
Y3	-5.756317	0.0000021
Y4	-5.022938	0.0000144

The t-tests for each variable comparing the means of two groups, all returned to be significant ($p<0.05$). This indicates a statistically significant difference between the groups for each variable.

(d) **Compare the results of (b) and (c) as to the contribution of each variable to separation of the groups.**

All between-group differences of each variable were significantly different, suggesting that all variables likely have strong discriminatory power. Furthermore, the absolute value of the t-values reflect the significance or strength of evidence for differences between groups on each variable. For larger absolute t-values, there is stronger evidence that the variable differs between groups.

Based on the results of (c), Y3 shows the largest between-group difference, while Y1 shows the smallest. From the standardized discriminant function coefficients and the partial F values, we found that Y1 has the most contribution to the separation of groups, while Y3 contributes the least. So, even though Y3 differs strongly between groups, Y1 is more important in discriminating between them This pattern can occur because Y3's information may overlap with other variables, so it provides little unique contribution to the separation once other variables are accounted for.

(e) **Find the partial F for each variable. Do the partial F's rank the variables in the same order of importance as the standardized coefficients?**

$$F = (v - p + 1)\frac{T_p^2 - T_{p-1}^2}{v + T_{p-1}^2}$$

Partial F for each variable

```
v = n1+n2-2

T2_y1 <- hotelling.test(beetle1[,2:4], beetle2[,2:4])$stats[[1]]
F_y1 <- ((v-p+1)*(T2-T2_y1))/(v+T2_y1)

T2_y2 <- hotelling.test(beetle1[,c(1,3,4)], beetle2[, c(1,3,4)])$stats[[1]]
F_y2 <- ((v-p+1)*(T2-T2_y2))/(v+T2_y2)

T2_y3 <- hotelling.test(beetle1[, c(1,2,4)], beetle2[, c(1,2,4)])$stats[[1]]
F_y3 <- ((v-p+1)*(T2-T2_y3))/(v+T2_y3)

T2_y4 <- hotelling.test(beetle1[,1:3], beetle2[,1:3])$stats[[1]]
F_y4 <- ((v-p+1)*(T2-T2_y4))/(v+T2_y4)

kable(data.frame(Variable = vars, F_partial = c(F_y1, F_y2, F_y3, F_y4)))
```

Variable	F_partial
Y1	35.933601
Y2	5.799435
Y3	1.774944
Y4	8.259241

The partial F-values are calculated for each variable. This test ranked Y1 as the most significant variable, then Y4, Y2, and Y3. This rank is almost similar with the standardized coefficients, with the exception of Y2 and Y4 whose ranks were reversed.

Scores on Fish Dataset

```
df2 <- suppressMessages(read_excel("Scores on Fish.xlsx",skip=2))
method1 <- df2[, 1:4]
method2 <- df2[, 5:8]
method3 <- df2[, 9:12]

names(method1) <- c("Y1","Y2","Y3","Y4")
names(method2) <- c("Y1","Y2","Y3","Y4")
names(method3) <- c("Y1","Y2","Y3","Y4")

method1$Method <- 1
method2$Method <- 2
method3$Method <- 3

fish <- bind_rows(method1, method2,method3)
fish$Method <- factor(fish$Method)
```

(a) **Find the eigenvectors.**

(b) **Carry out tests of significance for the discriminant functions and find the relative importance of each. Do these two procedures agree as to the number of important discriminant functions?**

(c) **Find the standardized coefficients and comment on the contribution of the variables to separation of groups.**

(d) **Find the partial F for each variable. Do they rank the variables in the same order as the standardized coefficients for the first discriminant function?**

(e) **Plot the first two discriminant functions for each observation and for the mean vectors.**

(f) **Carry out a stepwise selection of variables.**