STA 135 - Multivariate Data Analysis - Winter 2023

Final Project - Multivariate Analysis on Leptoconops Biting Flies

Yizhang Huang - 917099557

3/20/2023

Introduction

There are two types of biting flies called Leptoconops that appear very similar to each other. For a long time, people believed that they were the same species due to their physical similarities. However, further research has revealed that there are biological differences between the two, including varying sex ratios of newly hatched flies and different biting patterns. In this project, I will implement various statistical tools to examine the difference between the two species and use classification methods to classify a particular fly with given information.

Analysis

Visualization

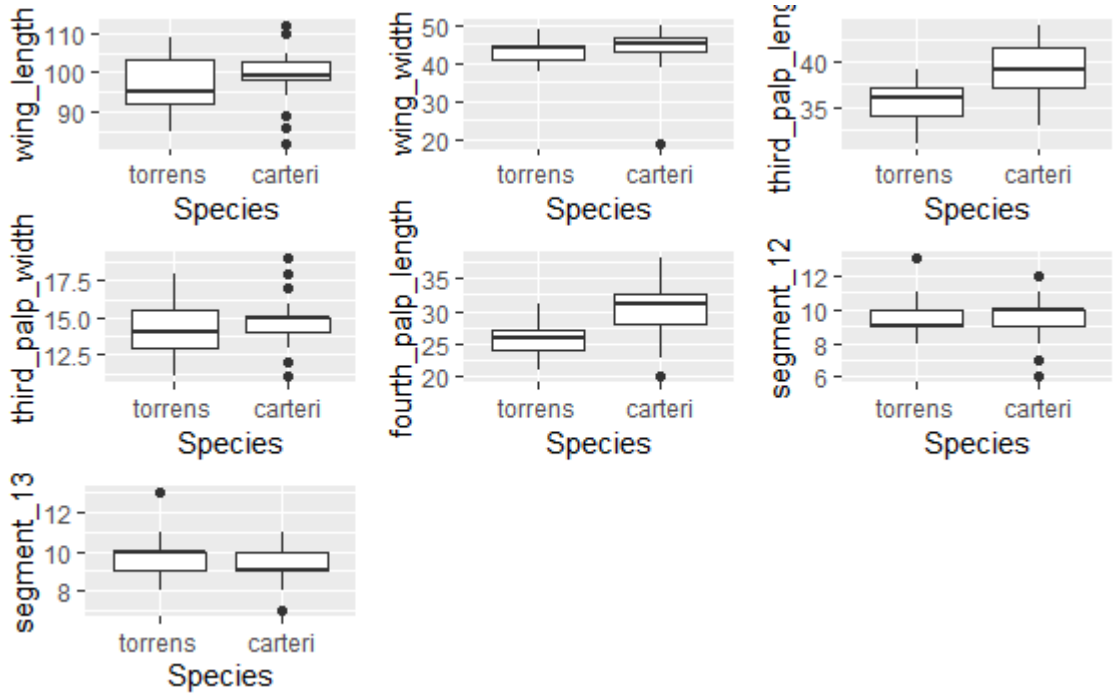

Figure 1. Boxplots of each variable of the two different species.

To analyze the differences between two species of flies, torrens and carteri, we first used data visualization. Boxplots were created to compare the features of the two species. From the boxplots, it was observed that while many of the features of the two species were quite similar, there were certain features that could be used to easily distinguish between them.

Specifically, the third and fourth palp lengths appeared to be key distinguishing factors between the two species. These lengths were notably different in the two species, and could therefore be used to quickly identify which group a particular fly belonged to.

Hotellings' T^2 Test

Next we used Hotellings' T^2 to test $H_0 : \vec{\mu}_1 - \vec{\mu}_2 = 0$ vs $H_A : \vec{\mu}_1 - \vec{\mu}_2 \neq 0$. By comparing T^2 with

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

where n_1 and n_2 are numbers of observations from both species, p is the numbers of variates, and $\alpha = 0.05$. By using R, we obtain $T^2 = 106.1348$ and $critical value = 16.59331$. Since $106.12 > 16.59$, we reject H_0 and conclude that there is significant difference between the two species

Confidence Interval of mean difference

Next, we conducted a 95% Confidence Interval of mean difference.

```
95% simultaneous confidence interval
                        [,1]      [,2]
wing_length        -2.9574687 8.7288972
wing_width         -3.1434050 4.8005479
third_palp_length   1.4723367 6.4133776
third_palp_width   -1.5545588 1.8402731
fourth_palp_length  0.7584051 7.9844521
segment_12         -0.9865633 1.1579919
segment_13         -1.3116852 0.6259709
```

As we can see from the output above, there can be a significant difference in wing length, third palp length, and fourth palp length between the two species, which helps reinforce our finding.

Classification

Lastly we will utilize Naive Bayes Classifier to perform classification. We first split data into 60% training and 40% testing. Then, based on the predicted data, we obtain the following confusion matrix:

```
Confusion Matrix and Statistics

   y_pred
     0  1
  0 16  1
  1  4 14

               Accuracy : 0.8571
                 95% CI : (0.6974, 0.9519)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : 0.0003014

                  Kappa : 0.7154

 Mcnemar's Test P-Value : 0.3710934

            Sensitivity : 0.8000
            Specificity : 0.9333
         Pos Pred Value : 0.9412
         Neg Pred Value : 0.7778
             Prevalence : 0.5714
         Detection Rate : 0.4571
   Detection Prevalence : 0.4857
      Balanced Accuracy : 0.8667

       'Positive' Class : 0
```

In the confusion matrix above, 0 and 1 represent the two types of biting flies Torrens and Carteri respectively. As we can see, we successfully classified 30 observations and misclassified 5 observations, which leads us to an accuracy of 85.7%. The misclassification was likely due to the outliers that we see in figure 1. And probably would give us a better result if we remove the outliers or if we have a bigger dataset.

Conclusion and Discussion

This project aimed to differentiate between two similar-looking species of biting flies, Leptoconops. Statistical tools were used to identify distinguishing characteristics, such as the third and fourth palp length. The Hotellings' T^2 test and confidence interval analyses revealed significant differences between the two species in wing length, third palp length, and fourth palp length. Finally, a Naive Bayes Classifier was utilized to classify the species based on given information, achieving 85.7% accuracy. However, the accuracy of this method may have been influenced by outliers or the small size of the dataset. Further research with a larger dataset could provide more accurate results.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
rm(list = ls())
setwd("C:/Users/s1a1m/OneDrive/文档/study/2023winter/sta135/proj")
dat<- read.table("T6-15.dat", header=FALSE)
colnames(dat) <- c("wing_length", "wing_width", "third_palp_length","third_palp_width",
                   "fourth_palp_length","segment_12","segment_13","species")

split_data <- split(dat, dat[, "species"])
torrens=split_data$`0`
carteri=split_data$`1`
torrens=torrens[, -which(names(torrens) == "species")]
carteri=carteri[, -which(names(carteri) == "species")]

dat$species=as.factor(dat$species)
library(gridExtra)
library(ggplot2)
p0=ggplot(dat, aes(x = species, y = wing_length)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "wing_length")

p1=ggplot(dat, aes(x = species, y = wing_width)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "wing_width")

p2=ggplot(dat, aes(x = species, y = third_palp_length)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "third_palp_length")

p3=ggplot(dat, aes(x = species, y = third_palp_width)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "third_palp_width")

p4=ggplot(dat, aes(x = species, y = fourth_palp_length)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "fourth_palp_length")

p5=ggplot(dat, aes(x = species, y = segment_12)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "segment_12")

p6=ggplot(dat, aes(x = species, y = segment_13)) +
  geom_boxplot() +
  scale_x_discrete(name = "Species", labels = c("torrens", "carteri")) +
  labs(x = "Species", y = "segment_13")

grid.arrange(p0,p1, p2, p3,p4,p5,p6, ncol = 3,nrow=3)
n<-c(35,35)
p<-7

tmean<-colMeans(torrens)
cmean<-colMeans(carteri)

d<-cmean-tmean

St<-var(torrens)
Sc<-var(carteri)
Sp<-((n[1]-1)*St+(n[2]-1)*Sc)/(sum(n)-2)

t2 <- t(d)%*%solve(sum(1/n)*Sp)%*%d
cat("T-square is ", t2)

alpha<-0.05
```

```r
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)
cat('critical value is ', cval)

wd<-sqrt(cval*diag(Sp)*sum(1/n))
Cis<-cbind(d-wd,d+wd)

cat("95% simultaneous confidence interval","\n")
Cis

library(e1071)
library(caTools)
library(caret)
split <- sample.split(dat, SplitRatio = 0.6)
train_cl <- subset(dat, split == "TRUE")
test_cl <- subset(dat, split == "FALSE")

# Feature Scaling
train_scale <- scale(train_cl[, 1:7])
test_scale <- scale(test_cl[, 1:7])

# Fitting Naive Bayes Model
# to training dataset
set.seed(120)  # Setting Seed
classifier_cl <- naiveBayes(species ~ ., data = train_cl)


# Predicting on test data'
y_pred <- predict(classifier_cl, newdata = test_cl)

# Confusion Matrix
cm <- table(test_cl$species, y_pred)
cm

# Model Evaluation
confusionMatrix(cm)
```