STA 135 Project 4

Grant Gambetta

3/4/2022

Introduction

For this project, the goal was to develop a classification rule for the genders of Concho Water Snakes using tail length (mm) and snout to vent length (mm) measurements. First, exploratory data analysis was conducted to gain initial insights about the dataset. Exploratory analysis was performed using several graphical and numerical techniques such as sample mean vectors, correlation and covariance matrices, scatterplots, boxplots, and histograms. To develop the classification rule, Linear Discriminant Analysis was performed and the confusion matrix was calculated to determine the Apparent Error Rate (APER).

This report consists of four main sections: introduction, materials and methods, results, and appendix. In the materials and methods section, I provide an overview of the dataset and discuss the statistical, numerical, and visualization methods I used to describe and analyze the data. In the results section, I discuss the main findings of the analysis and interpret some of the R outputs that I generated. Within the results section, there are two subsections: exploratory analysis and classification rule, which is where the results of the exploratory analysis and classification rule are discussed, respectively. Lastly, the appendix is where the entire R code can be found.

Materials and Methods

To begin with, it is important to understand the dataset behind this project. The dataset consisted of three variables and 66 observations, where two of the variables were numeric/continuous and the third was a binary categorical variable. The two continuous variables were originally named `X2` and `X3`, which denoted the tail length and snout to vent length in millimeters of Concho Water Snakes. To make the interpretation more straightforward, I decided to rename `X2` and `X3` to `X1` and `X2`, respectively. The categorical variable, now denoted as `X3`, represented the gender of each Concho Water Snake, specifically male or female. There were 37 observations for females, denoted n_1, and 29 observations for males, denoted n_2.

To develop the classification rule, Linear Discriminant Analysis (LDA) was performed and evaluation methods such as the confusion matrix, holdout procedure, and Apparent Error Rate (APER) were used to assess the performance of the classification rule.

Results

Exploratory Data Analysis

To begin with, exploratory data analysis was performed to gain preliminary insights about tail length and snout to vent length measurements for each gender of Concho Water Snake. I computed summary statistics for each population of Concho Water Snakes (male and female) using methods such as the sample mean vectors, variance-covariance matrices, and correlation matrices, which can be found below.

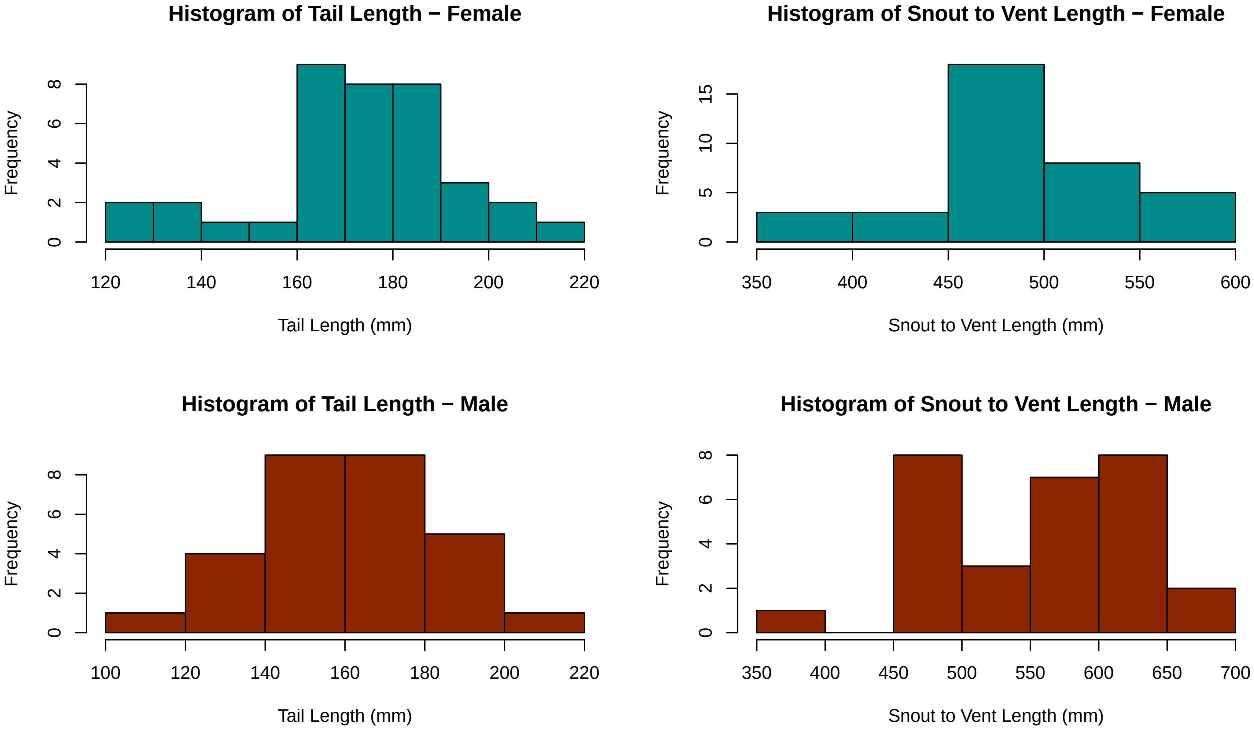
$$\bar{\mathbf{X}}_{\text{female}} = \begin{bmatrix} 173 \\ 489 \end{bmatrix} \qquad \bar{\mathbf{X}}_{\text{male}} = \begin{bmatrix} 161 \\ 554 \end{bmatrix}$$

$$\mathbf{S}_{\text{female}} = \begin{bmatrix} 413 & 873 \\ 873 & 2418 \end{bmatrix} \qquad \mathbf{S}_{\text{male}} = \begin{bmatrix} 420 & 1402 \\ 1402 & 5356 \end{bmatrix}$$

$$\mathbf{r}_{\text{female}} = \begin{bmatrix} 1.00 & 0.87 \\ 0.87 & 1.00 \end{bmatrix} \qquad \mathbf{r}_{\text{male}} = \begin{bmatrix} 1.00 & 0.93 \\ 0.93 & 1.00 \end{bmatrix}$$

Looking at the sample mean vectors, we observe that females have a larger mean tail length and males have a larger mean snout to vent length. From the variance-covariance matrices, we see that the covariance between tail length and snout to vent length for both males and females is positive, which means these two variables have a positive relationship. We can use the correlation matrices to determine the strength of linear relationship between these two variables. Observing the correlation matrix for females, we see that the correlation coefficient between tail length and snout to vent length is 0.87, which implies these two variables are positively correlated for female Concho Water Snakes. Observing the correlation matrix for males, we see that the correlation coefficient between tail length and snout to vent length is 0.93, which also implies these two variables are positively correlated for male Concho Water Snakes.

Next, side-by-side histograms were created to observe the distribution of tail length and snout to vent length for each population.

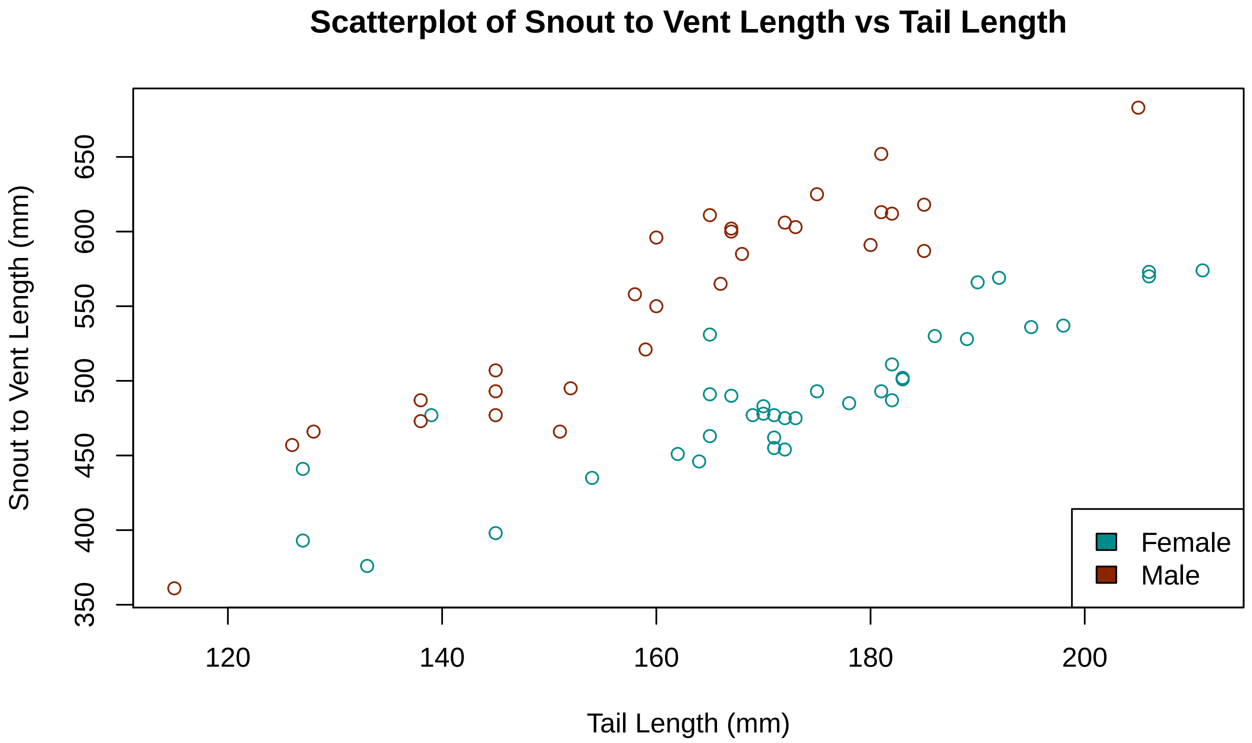

Histogram of Tail Length − Female

Histogram of Snout to Vent Length − Female

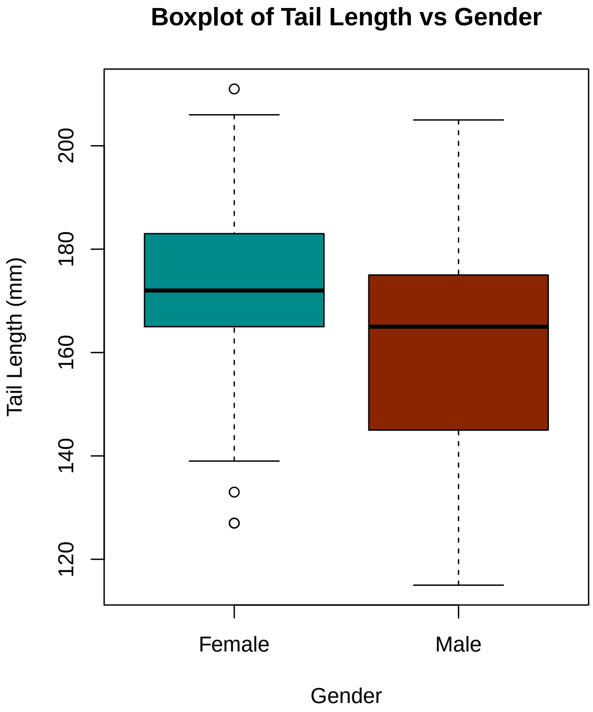
Histogram of Tail Length − Male

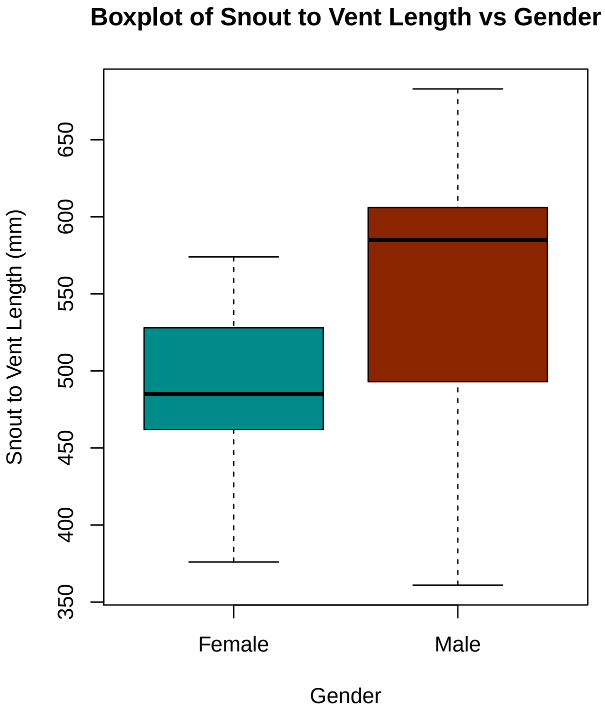
Histogram of Snout to Vent Length − Male

Observing the histograms, it appears that tail length for males is the only one that follows a normal distribution. For females, tail length and snout to vent length look a bit right skewed. Additionally, a scatterplot of snout to vent length vs tail length, color coded by population (male or female), was created to observe the overall trend of the data.



Observing the scatterplot, we see that females tend to have smaller snout to vent length than males do. Additionally, males and females seem to have similar tail length overall, however, there are a few observations for females that are significantly larger (greater than 200mm) than the rest which could be the reason why females had a larger mean tail length than males. Also, the data for both populations has a clear linear relationship, which supports the variance-covariance and correlation matrices from earlier. Lastly, side-by-side boxplots were created to observe the data for discriminating effects and outliers.





Looking at the boxplots, it appears that there is a discriminating effect for both tail length and snout to vent length. This is because the values for tail length among females are larger than the values for tail length among males and the snout to vent length values are larger for males than they are for females. Additionally, there appear to be three potential outliers for tail length among females.

Classification Rule

Lastly, a classification rule was developed to classify the gender of Concho Water Snakes as female or male based on their tail length and snout to vent length. The two populations for this classification rule are defined as π_1 (Female Concho Water Snakes) and π_2 (Male Concho Water Snakes). The formula for calculating the classification rule is

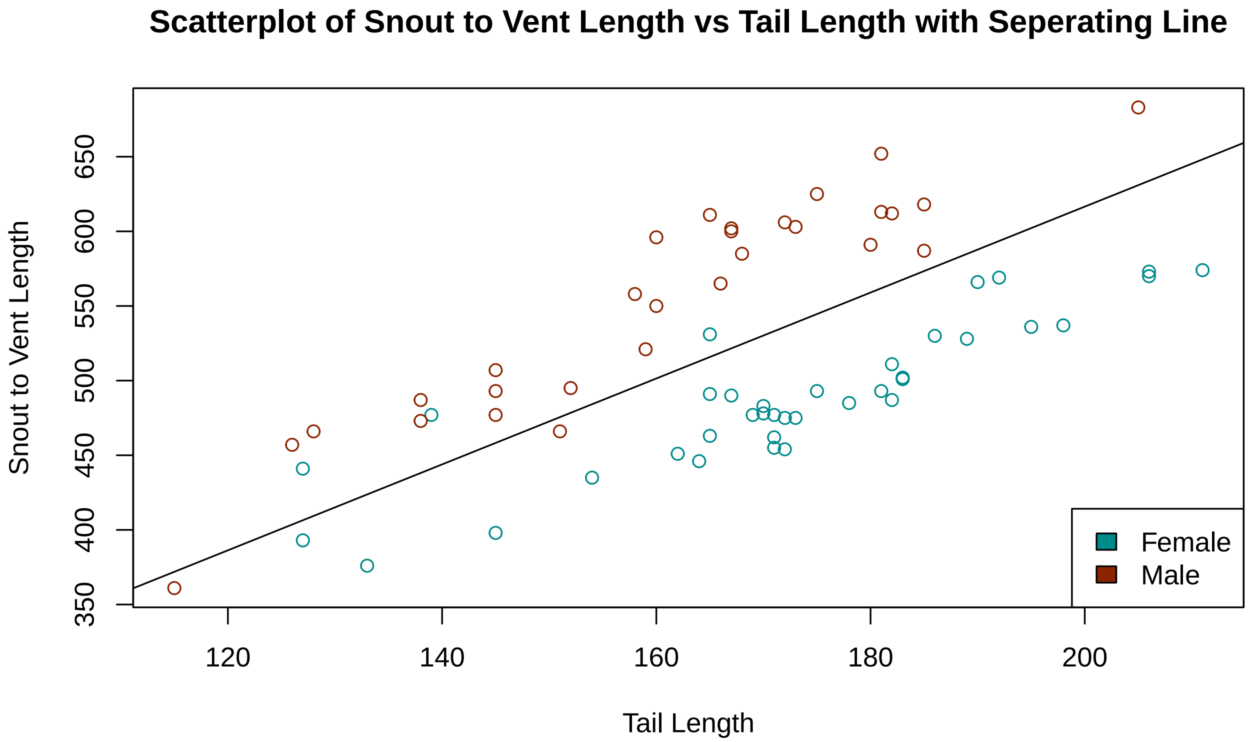
$$(\bar{\mathbf{X}}_{\text{female}} - \bar{\mathbf{X}}_{\text{male}})' S_{\text{pooled}}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \geq c$$

where $\bar{\mathbf{X}}_{\text{female}} = \begin{bmatrix} 173 \\ 489 \end{bmatrix}$, $\bar{\mathbf{X}}_{\text{male}} = \begin{bmatrix} 161 \\ 554 \end{bmatrix}$, $S_{\text{pooled}} = \begin{bmatrix} 416 & 1104 \\ 1104 & 3703 \end{bmatrix}$, and $c = (\bar{\mathbf{X}}_{\text{female}} - \bar{\mathbf{X}}_{\text{male}})' S_{\text{pooled}}^{-1} \left(\frac{\bar{\mathbf{X}}_{\text{female}} + \bar{\mathbf{X}}_{\text{male}}}{2} \right) = -5.07174$.

Assuming multivariate normal distributions with common variance-covariance matrices, equal prior probabilities ($p_1 = p_2 = 0.5$), and equal costs of misclassification, the minimum ECM rule for classifying the gender of Concho Water Snakes was calculated as follows:

$$\text{Allocate (X1, X2) to } \pi_1 \text{ if } 0.3564042 X_1 - 0.1238378 X_2 \geq -5.07174$$
$$\text{Allocate (X1, X2) to } \pi_2 \text{ if } 0.3564042 X_1 - 0.1238378 X_2 < -5.07174$$

where `X1` represents tail length (in mm) and `X2` represents snout to vent length (in mm). Below is the scatterplot of the two populations with the separating line from the classification rule.



Next, the confusion matrix and the apparent error rate (APER) were calculated to evaluate the classification rule that was specified above. The confusion matrix calculated without the holdout procedure is below.

Table 1: Confusion Matrix calculated without the holdout procedure.

	Female (Predicted)	Male (Predicted)
Female	34	3
Male	2	27

Observing the confusion matrix, we see that 34 out of 37 females were classified correctly and 27 out of 29 males were classified correctly. More specifically, there were 3 females that were mistakenly classified as male and 2 males that were mistakenly classified as female.

Additionally, the apparent error rate (APER) was calculated to further evaluate the accuracy of the classification rule and is defined as

$$\frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

where n_{1M} and n_{2M} represent the number of items in π_1 and π_2 that were misclassified, respectively. The APER was roughly 0.076, which implies this classification rule is accurate when it comes to classifying Concho Water Snakes as male or female. The final step of this project was to calculate the confusion matrix using the holdout procedure, which can be found below.

Table 2: Confusion Matrix calculated using the holdout procedure.

	Female (Predicted)	Male (Predicted)
Female	34	3
Male	2	27

Observing the confusion matrix that was calculated using the holdout procedure, we see that it is the same as the confusion matrix from without the holdout procedure and therefore the APER for the holdout procedure was 0.076. This also shows that the classification rule is a sufficient estimate for this data.

Appendix

```r
library(dplyr)
library(ggplot2)
library(MASS)
library(stargazer)
library(caret)

df <- read.table('Project_4_Data.txt', header = T)
df$X1 <- as.factor(df$X1)
colnames(df)[1] <- 'X3'
colnames(df)[2] <- 'X1'
colnames(df)[3] <- 'X2'
head(df)

# split dataset based on population
female <- df %>% filter(X3 == 'Female') %>% dplyr::select(-X3)
male <- df %>% filter(X3 == 'Male') %>% dplyr::select(-X3)

write_matex2 <- function(x) {
  begin <- "\\begin{bmatrix}"
  end <- "\\end{bmatrix}"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  paste(c(begin, X, end), collapse = "")
}
# mean vectors
mean_vecs <- lapply(split(df[, c('X1', 'X2')], df$X3), function(x) colMeans(x))
# var-cov matrices
var_cov_mat <- lapply(split(df[, c('X1', 'X2')], df$X3), cov)
# correlation matrices
cor_mat <- lapply(split(df[, c('X1', 'X2')], df$X3), cor)

# side by side histograms
par(mfrow=c(2,2))
hist(female$X1, xlab = 'Tail Length (mm)', main = 'Histogram of Tail Length - Female', col = 'darkcyan'
hist(female$X2, xlab = 'Snout to Vent Length (mm)', main = 'Histogram of Snout to Vent Length - Female'
hist(male$X1, xlab = 'Tail Length (mm)', main = 'Histogram of Tail Length - Male', col = 'orangered4')
hist(male$X2, xlab = 'Snout to Vent Length (mm)', main = 'Histogram of Snout to Vent Length - Male', co
```

```r
cols <- c('darkcyan', 'orangered4')
plot(df$X1, df$X2, col = cols[df$X3], main = 'Scatterplot of Snout to Vent Length vs Tail Length', xlab
legend('bottomright', fill = cols, legend = c(levels(df$X3)))

# boxplots
par(mfrow=c(1,2))
```

```r
boxplot(X1 ~ X3, data = df, col = cols[unique(df$X3)], xlab = 'Gender', ylab = 'Tail Length (mm)', main
boxplot(X2 ~ X3, data = df, col = cols[unique(df$X3)], xlab = 'Gender', ylab = 'Snout to Vent Length (m
```

```r
# (b)
n <- by(df[, 2:3], df$X3, nrow)
n1 <- n[1][[1]]
n2 <- n[2][[1]]

M <- by(df[, 2:3], df$X3, colMeans)
m1 <- M[1][[1]] # female
m2 <- M[2][[1]] # male

S <- by(df[, 2:3], df$X3, var)
S1 <- S[1][[1]]
S2 <- S[2][[1]]
Sp <- ((n1 - 1)*S1 + (n2-1)*S2)/(n1 + n2 - 2)
(a <- solve(Sp) %*% (m1 - m2))
(m <- (t(a)%*%m1 + t(a)%*%m2)/2)

plot(df$X1, df$X2, col = cols[df$X3], main = 'Scatterplot of Snout to Vent Length vs Tail Length with S
abline(a = m/a[2], b = -a[1]/a[2])
legend('bottomright', fill = cols, legend = c(levels(df$X3)))
```

```
# (c)
predictions <- t(a) %*% t(df[, 2:3]) < m[1]
Predicted <- factor(predictions, labels = c('Female (Predicted)', 'Male (Predicted)'))
CM <- table(df$X3, Predicted)
pander(CM, caption = 'Confusion Matrix calculated without the holdout procedure.', split.cells = c(1,1,2

# APER
(CM[1,2] + CM[2,1])/(nrow(df))

# (d)
lda_holdout <- lda(X3 ~ X1 + X2, data = df, CV = TRUE, prior = c(.5, .5))
cm_holdout <- table(df$X3, lda_holdout$class)
colnames(cm_holdout)[1] <- 'Female (Predicted)'
colnames(cm_holdout)[2] <- 'Male (Predicted)'
pander(cm_holdout, caption = 'Confusion Matrix calculated using the holdout procedure.', split.cells = c
```