STA 135 Project 2

Grant Gambetta

2/10/2022

Introduction

For this project, the goal was to perform statistical data analysis on a dataset consisting of tail and wing length measurements for male and female hook-billed kites. First, exploratory 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. Next, to support the visual inspection from the exploratory analysis, statistical analysis was performed using Hotelling's T^2 statistic, 95% confidence region, and 95% simultaneous confidence intervals to determine if there was a statistically significant difference in the mean vectors for male and female hook-billed kites, as well as to determine whether male or female hook-billed kites are larger based on their tail and wing lengths.

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. Additionally in this section, I detail the specific formulas behind the statistical calculations. 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 statistical analysis, which is where the results of the exploratory analysis and statistical analysis 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 90 rows, where two of the variables were numeric/continuous and the third was a binary categorical variable. The two continuous variables were "tail" and "wing," which denoted the tail length and wing length in millimeters for each hook-billed kite, respectively. The factor variable represented the sex of each hook-billed kite, where 0 denoted male and 1 denoted female. To make the column names a bit more descriptive, I renamed "tail" and "wing" to `tail_length` and `wing_length`, respectively. There were 45 male hook-billed kites and 45 female hook-billed kites in the dataset.

For the statistical analysis, a hypothesis test was conducted to compare the mean vectors for the male and female populations of hook-billed kites. The test statistic used was Hotelling's T^2, and the formula is below.

$$T^2 = (\overline{X}_{\text{male}} - \overline{X}_{\text{female}})^T \left[\frac{1}{n_{\text{male}}} S_{\text{male}} + \frac{1}{n_{\text{female}}} S_{\text{female}} \right]^{-1} (\overline{X}_{\text{male}} - \overline{X}_{\text{female}})$$

Also, a 95% confidence region and 95% simultaneous confidence intervals were constructed for $\mu_1 - \mu_2$. The general formula for the 95% confidence region is below, where $c_\alpha = \chi^2_p(\alpha)$.

$$R(X) = \left\{ \mu : \left(\overline{X}_1 - \overline{X}_2 - \mu\right)^T \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1} \left(\overline{X}_1 - \overline{X}_2 - \mu\right) \leq c_\alpha \right\}$$

The general formula for the simultaneous confidence intervals can be found below.

$$\overline{X}_{1i} - \overline{X}_{2i} \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)_{ii}} \text{ for } i = 1, \ldots, p$$

Results

Exploratory Analysis

To begin with, exploratory analysis was performed to gain an initial understanding of the dataset and preliminary insights about each population (male and female) of hook-billed kites. I began by computing summary statistics such as the mean vectors, variance-covariance matrices, and correlation matrices for each population, which can be found below.

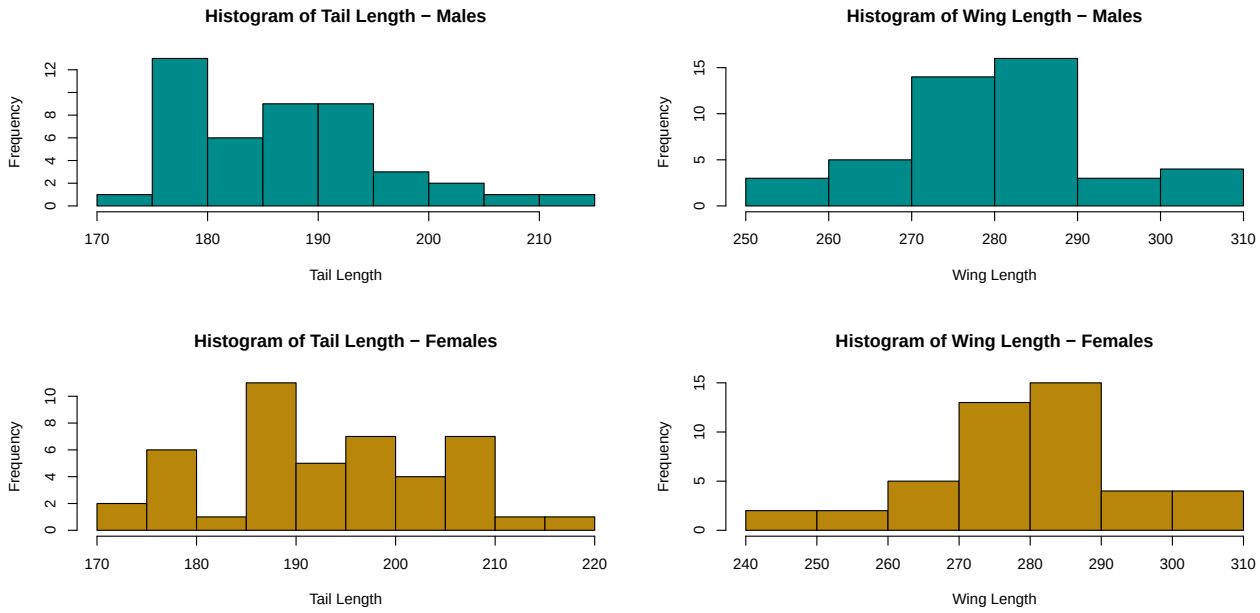
$$\overline{\mathbf{X}}_{\text{male}} = \begin{bmatrix} 187 \\ 281 \end{bmatrix} \qquad \overline{\mathbf{X}}_{\text{female}} = \begin{bmatrix} 194 \\ 280 \end{bmatrix}$$

$$\mathbf{S}_{\text{male}} = \begin{bmatrix} 87 & 88 \\ 88 & 169 \end{bmatrix} \qquad \mathbf{S}_{\text{female}} = \begin{bmatrix} 121 & 122 \\ 122 & 209 \end{bmatrix}$$

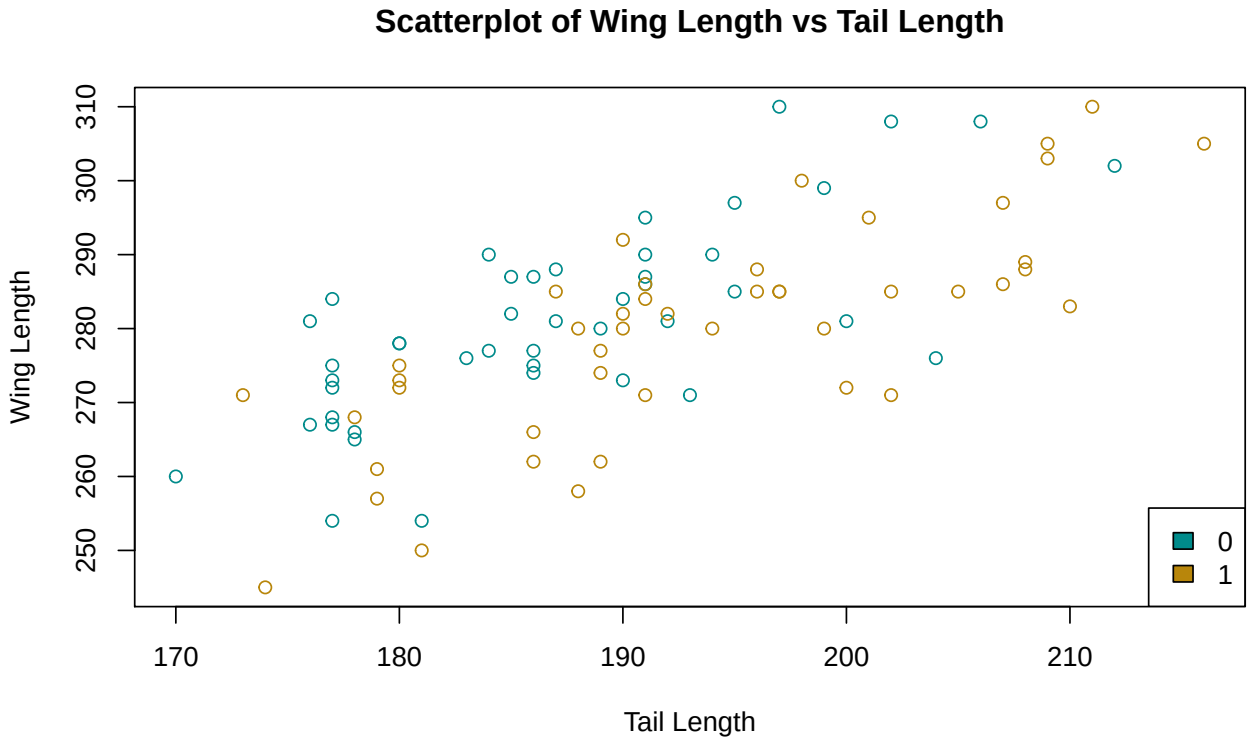
$$\mathbf{r}_{\text{male}} = \begin{bmatrix} 1.00 & 0.73 \\ 0.73 & 1.00 \end{bmatrix} \qquad \mathbf{r}_{\text{female}} = \begin{bmatrix} 1.00 & 0.77 \\ 0.77 & 1.00 \end{bmatrix}$$

Looking at the sample mean vectors, we see that on average, female hook-billed kites tend to have larger tail length than males and that both males and females have similar wing length (281mm and 280mm, respectively). From looking at the variance-covariance matrices, we see that females tend to have a larger variance for tail length and wing length and a larger covariance between tail length and wing length than males do. Lastly, tail length and wing length are highly correlated for both males and females, with correlation coefficients of 0.73 and 0.77, respectively. This implies that there is a clear linear relationship between tail length and wing length for both males and females.

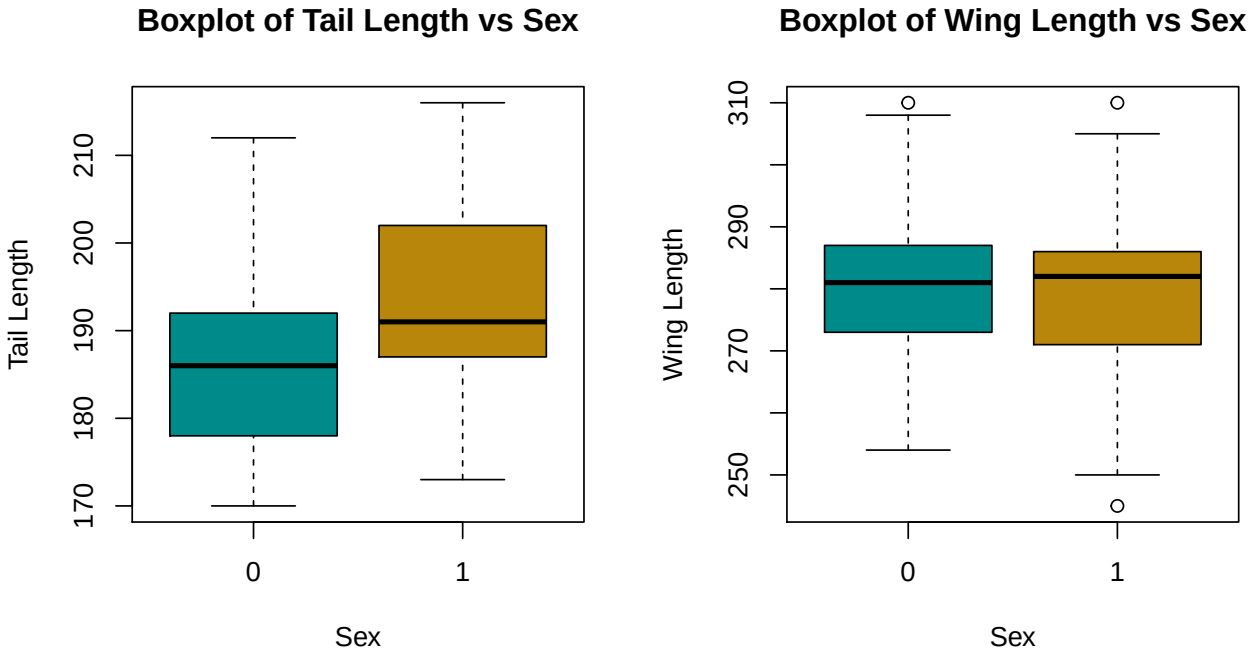
Next, several visualizations were created to analyze the data through visual inspection. First, I created side-by-side histograms to observe the distributions of tail length and wing length for each population.



Looking at the histograms, it seems that wing length for both males and females follows a normal distribution, while tail length for both males and females is a bit right skewed. Next, I created a scatterplot of wing length vs tail length, color coded by gender (0 = male, 1 = female) which is below.



After looking at the scatterplot, it appears that females tend to have larger values for tail length than males do and the values for wing length for both genders are quite similar. These insights support the conclusions made from the sample mean vectors. Additionally, the scatterplot supports the high correlation between tail length and wing length for both males and females since the points for both genders have a clear linear relationship. Lastly, I created boxplots to examine the data for outliers and to determine if any discriminating effects are present among the two populations. The boxplots are below and are color coded based on gender.



After observing the boxplots, it appears that tail length does not have any outliers for either gender while wing length has one outlier for males (group 0) and two outliers for females (group 1). Also, it is clear that females have a larger median tail length than males do and that both genders have a similar median wing length. There is also a discriminating effect present for tail length because the values of tail length are larger for females.

Statistical Analysis

To verify and support the preliminary results from the exploratory analysis, statistical analysis was performed. First, a hypothesis test for equality of population mean vectors was conducted at $\alpha = 0.05$ with the following null and alternative hypotheses

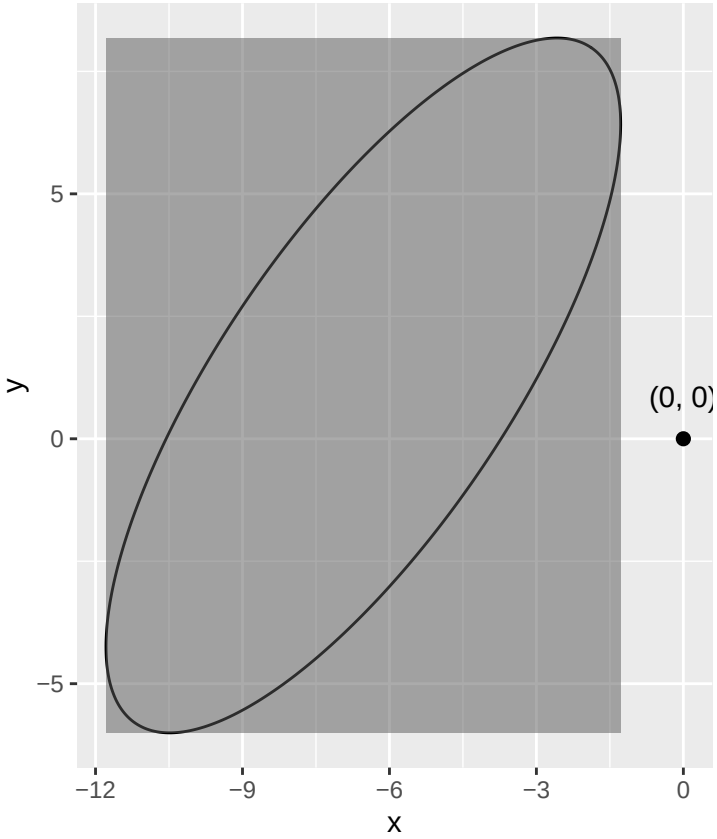
$$H_0 : \mu_{\text{male}} = \mu_{\text{female}} \text{ vs } H_a : \mu_{\text{male}} \neq \mu_{\text{female}}$$

where μ_{male} represents the mean vector for male hook-billed kites and μ_{female} represents the mean vector for female hook-billed kites.

The test statistic for this hypothesis test was Hotelling's T^2 (equation can be found in the materials and methods section) and the critical value was $\chi^2_p(\alpha) = 5.99$. After calculating $T^2 = 25.67$, we determine that $T^2 > \chi^2_p(\alpha)$ which allows us to reject H_0 and conclude that the mean vectors for male and female hook-billed kites are not equal.

Next, a 95% confidence region and 95% simultaneous confidence intervals were computed for $\mu_1 - \mu_2$, where $\mu_1 = \mu_{\text{male}}$ and $\mu_2 = \mu_{\text{female}}$. The equations for both the confidence region and simultaneous confidence intervals can be found in the materials and methods section. The resulting difference vector from $\mu_1 - \mu_2$ and the plot for the 95% confidence region is below. We see that the difference in tail length between male and female hook-billed kites is -6.53 and the difference in wing length between male and female hook-billed kites is 1.09.

$$\mu_1 - \mu_2 = \begin{bmatrix} -6.53 \\ 1.09 \end{bmatrix}$$



From using the formula for simultaneous confidence intervals as listed in the materials and methods section, the following 95% confidence intervals are obtained for $\mu_1 - \mu_2$

$$\text{Tail Length} : [-1.28, -11.79] \qquad \text{Wing Length} : [-6.003, 8.18]$$

After observing the confidence intervals, we notice that the confidence interval for tail length does not include 0, which means we can conclude that the difference in means for tail length among male and female hook-billed kites is statistically significant. In other words, female hook-billed kites have larger tail length on average than males do. The confidence interval for wing length contains 0, which means we fail to reject the null hypothesis and therefore conclude that the difference in means for wing length among male and female hook-billed kites is not statistically significant.

Appendix

```
library(dplyr)
library(ggplot2)
library(stargazer)
library(ggpubr)
library(ggforce)

# read in the data, change column names, convert sex to factor variable
data <- read.table('Project_2_Data.txt', header = T)
colnames(data)[1] <- 'tail_length'
colnames(data)[2] <- 'wing_length'
data$sex <- as.factor(data$sex)
head(data)
```

```r
# split dataset into male and female groups
male <- data %>% filter(sex == 0) %>% select(tail_length, wing_length)
female <- data %>% filter(sex == 1) %>% select(tail_length, wing_length)

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(data[, c('tail_length', 'wing_length')], data$sex), function(x) colMeans(x))
# var-cov matrices
var_cov_mat <- lapply(split(data[, c('tail_length', 'wing_length')], data$sex), cov)
# correlation matrices
cor_mat <- lapply(split(data[, c('tail_length', 'wing_length')], data$sex), cor)

# side by side histograms
par(mfrow=c(2,2))
hist(male$tail_length, xlab = 'Tail Length', main = 'Histogram of Tail Length - Males', col = 'darkcyan
hist(male$wing_length, xlab = 'Wing Length', main = 'Histogram of Wing Length - Males', col = 'darkcyan
hist(female$tail_length, xlab = 'Tail Length', main = 'Histogram of Tail Length - Females', col = 'dark
hist(female$wing_length, xlab = 'Wing Length', main = 'Histogram of Wing Length - Females', col = 'dark
```

```r
cols <- c('darkcyan', 'darkgoldenrod')
plot(data$tail_length, data$wing_length, col = cols[data$sex], main = 'Scatterplot of Wing Length vs Ta
legend('bottomright', fill = cols, legend = c(levels(data$sex)))

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

```r
boxplot(tail_length ~ sex, data = data, col = cols[unique(data$sex)], xlab = 'Sex', ylab = 'Tail Length
boxplot(wing_length ~ sex, data = data, col = cols[unique(data$sex)], xlab = 'Sex', ylab = 'Wing Length
```

```r
mu_male <- colMeans(male)
mu_female <- colMeans(female)

sigma_male <- cov(male)
sigma_female <- cov(female)

n1 <- nrow(male)
n2 <- nrow(female)

diff <- mu_male - mu_female
S <- sigma_male/n1 + sigma_female/n2

alpha <- 0.05
p <- ncol(female)
```

```r
T2 <- t(diff) %*% solve(S, diff)
c2 <- qchisq(alpha, p, lower.tail = FALSE)
T2 > c2

e.decom <- eigen(S)

a1 <- sqrt(e.decom$values[1]) * sqrt(c2)
b1 <- sqrt(e.decom$values[2]) * sqrt(c2)
theta <- atan2(e.decom$vectors[, 1][2], e.decom$vectors[, 1][1])

# confidence region
x_range <- diff[1] + c(-1, 1) * sqrt(c2) * sqrt(diag(S))[1]
y_range <- diff[2] + c(-1, 1) * sqrt(c2) * sqrt(diag(S))[2]
rect_range <- data.frame(xmin = x_range[1], xmax = x_range[2], ymin = y_range[1], ymax = y_range[2])

ggplot() +
  geom_ellipse(aes(x0 = diff[1], y0 = diff[2], a = a1, b = b1, angle = theta)) +
  coord_fixed() +
  geom_rect(data = rect_range, mapping = aes(xmin = xmin, xmax = xmax,
  ymin = ymin, ymax = ymax), alpha = 0.5) +
  geom_point(aes(x = 0, y = 0), size = 2) +
  geom_text(aes(x = 0, y = 0), label = "(0, 0)", vjust = -1.5, show.legend = TRUE)
```

```r
# simultaneous CIs
diff
diff[1] + sqrt(c2)*sqrt(S)[1,1]
diff[1] - sqrt(c2)*sqrt(S)[1,1]
diff[2] + sqrt(c2)*sqrt(S)[2,2]
diff[2] - sqrt(c2)*sqrt(S)[2,2]
```