DiveIn CDT Penguins problem

Amparo Gimenez Rios

2025-03-25

Preparation

Loading needed packages

```
library(ggplot2)
library(dplyr)
library(GGally)
library(RColorBrewer)
library(readxl)
library(dunn.test)
```

Loading data

```
penguins <- read_excel("E:/Uni/PhD/Submission/DiveIn/Exercise-Docs/Exercise Docs/Exercise_penguins.xls"
```

Data exploration

Before proceeding to the main analyses, we will explore the data to understand its structure and content.

1. Structure of the data We can see an overview of the variables we have in this data and the type of data they contain (numerical, character, etc.)

```
str(penguins)
```

```
## tibble [344 x 9] (S3: tbl_df/tbl/data.frame)
##  $ rowid            : num [1:344] 1 2 3 4 5 6 7 8 9 10 ...
##  $ species          : chr [1:344] "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ island           : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ bill_length_mm   : chr [1:344] "39.100000000000001" "39.5" "40.299999999999997" "NA" ...
##  $ bill_depth_mm    : chr [1:344] "18.699999999999999" "17.399999999999999" "18" "NA" ...
##  $ flipper_length_mm: chr [1:344] "181" "186" "195" "NA" ...
##  $ body_mass_g      : chr [1:344] "3750" "3800" "3250" "NA" ...
##  $ sex              : chr [1:344] "male" "female" "female" "NA" ...
##  $ year             : num [1:344] 2007 2007 2007 2007 2007 ...
```

2. Summary of the data We can see number of observations and class of character variables and summary statistics of numerical variables (mean, median, min, max, etc.)

```r
summary(penguins)
```

```
##      rowid          species             island          bill_length_mm
##  Min.   :  1.00   Length:344         Length:344         Length:344
##  1st Qu.: 86.75   Class :character   Class :character   Class :character
##  Median :172.50   Mode  :character   Mode  :character   Mode  :character
##  Mean   :172.50
##  3rd Qu.:258.25
##  Max.   :344.00
##  bill_depth_mm      flipper_length_mm  body_mass_g            sex
##  Length:344         Length:344         Length:344         Length:344
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       year
##  Min.   :2007
##  1st Qu.:2007
##  Median :2008
##  Mean   :2008
##  3rd Qu.:2009
##  Max.   :2009
```

Cleaning the data

We will take a closer look at our data and transform column types, clean duplicates, missing data, rename columns, etc. if necessary.

3. Transform columns to numerical We will transform the columns that are factors to numerical.

```r
penguins <- penguins %>%
  mutate(bill_length_mm = as.numeric(bill_length_mm),
         bill_depth_mm = as.numeric(bill_depth_mm),
         flipper_length_mm = as.numeric(flipper_length_mm),
         body_mass_g = as.numeric(body_mass_g))
```

```
## Warning: There were 4 warnings in `mutate()`.
## The first warning was:
## i In argument: `bill_length_mm = as.numeric(bill_length_mm)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

4. Omit missing values We will remove rows with missing values.

```r
penguins <- na.omit(penguins)
```

There are NA string values in sex for some rows, but we can keep these as long as sex is not used in the analysis.

5. Round numerical data Round the data to 2 decimal places for bill length and bill depth to make it easier to read.

```
penguins <- penguins %>%
  mutate(length = round(bill_length_mm, 2),
         depth = round(bill_depth_mm, 2))
```

6. Rename columns We will rename flipper length and body mass columns to make them easier to read.

```
penguins <- penguins %>%
  rename(flipper_length = flipper_length_mm,
         body_mass = body_mass_g)
```

And we delete the original bill length and bill depth columns because we no longer need them.

```
penguins <- penguins %>%
  select(-bill_length_mm, -bill_depth_mm)
```

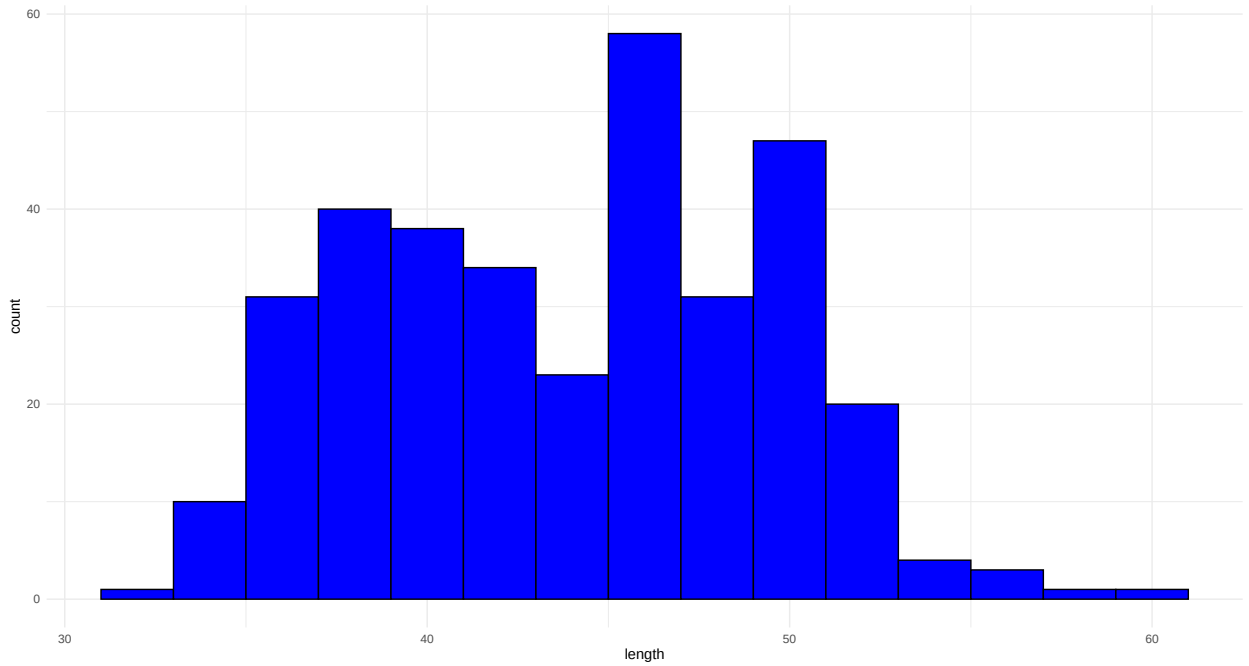
7. Convert to factors Convert species to factor so we can use this in the analysis.

```
penguins$species <- as.factor(penguins$species)
```

FIGURE 1: Culmen Length Across Species

8. Inspect data distribution Before proceeding to generate the boxplot, we will take a look at the distribution of bill length across species.

```
ggplot(penguins, aes(x = length)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  theme_minimal()
```



There seems to be a bimodal distribution (two peaks), which could be due to sex differences within species, but we should confirm this with a statistical test for normality. This is useful to decide whether to use parametric or non-parametric tests to check for differences between species.

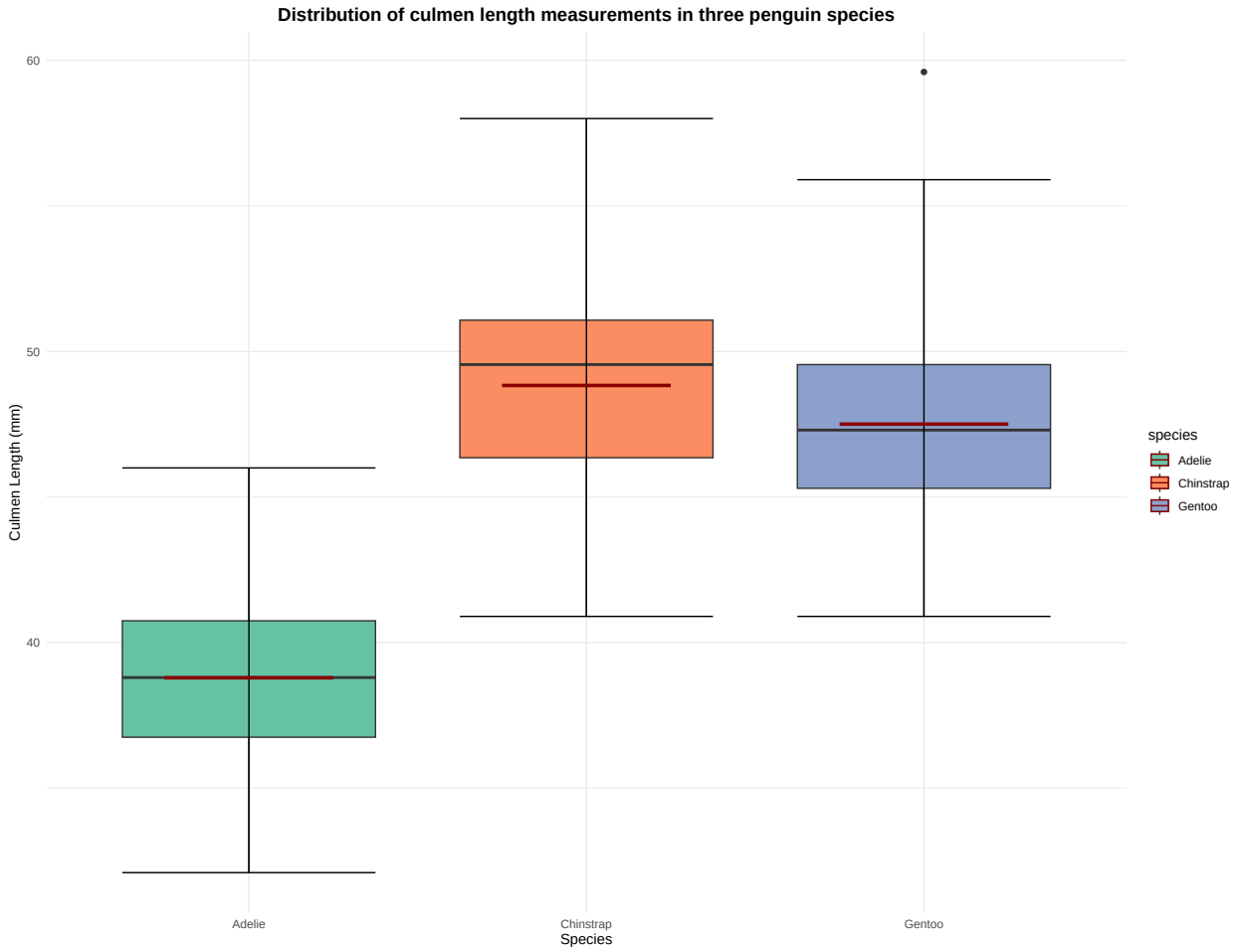
```
shapiro.test(penguins$length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  penguins$length
## W = 0.97485, p-value = 1.12e-05
```

Null hypothesis of normality is rejected. We cannot assume normality of distribution. We will use non-parametric tests for the analysis.

9. Generate boxplot We will generate a boxplot to visualize the distribution of bill length across species.

```
boxplot <- ggplot(penguins, aes(x = species, y = length, fill = species)) +
  geom_boxplot(outlier.shape = 16, outlier.size = 2, coef = 1.5) +  # `coef = 1.5` ensures proper whisk
  scale_fill_brewer(palette = "Set2") +  # Colorblind-friendly palette
  labs(title = "Distribution of culmen length measurements in three penguin species",
       x = "Species",
       y = "Culmen Length (mm)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14)) + # Center and bold title
  stat_boxplot(geom ='errorbar') +
  stat_summary(fun = mean, geom = "crossbar", color = "darkred", width = 0.5)
boxplot
```



This code generates a boxplot showing the distribution of culmen length measurements in three penguin species. The boxplot includes the median (middle black line), interquartile range (box), and whiskers extending to 1.5 times the interquartile range. Outliers are shown as individual points. The mean is represented by a red crossbar. RColorBrewer Set2 palette is used for colourblind-friendly colours.

10. Further analysis We can test for significant differences in culmen length between species. Taking into account the data is not normally distributted, the statistical test needs to be suitable to three categories with variable numerical. We will use the Kruskal-Wallis test.

```
kruskal.test(length ~ species, data = penguins)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  length by species
## Kruskal-Wallis chi-squared = 244.14, df = 2, p-value < 2.2e-16
```

The resulting p-value is statistically significan (over 0.05), which means there are significant differences in culmen length between species. However, we should perform post-hoc tests to determine which species differ from each other. We will perform Dunn text, which is a non-parametric post-hoc test to compare all pairs of species.

```
dunn.test::dunn.test(penguins$length, penguins$species, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 244.1367, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                   (Bonferroni)
## Col Mean-|
## Row Mean |     Adelie   Chinstra
## ---------+----------------------
## Chinstra |  -12.75351
##          |    0.0000*
##          |
##   Gentoo |  -13.13563   1.767498
##          |    0.0000*     0.1157
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

There is a statistically significant difference in culmen length between Adelie and Chinstra species, and between Adelie and Gentoo species. There is no significant difference between Chinstra and Gentoo species.

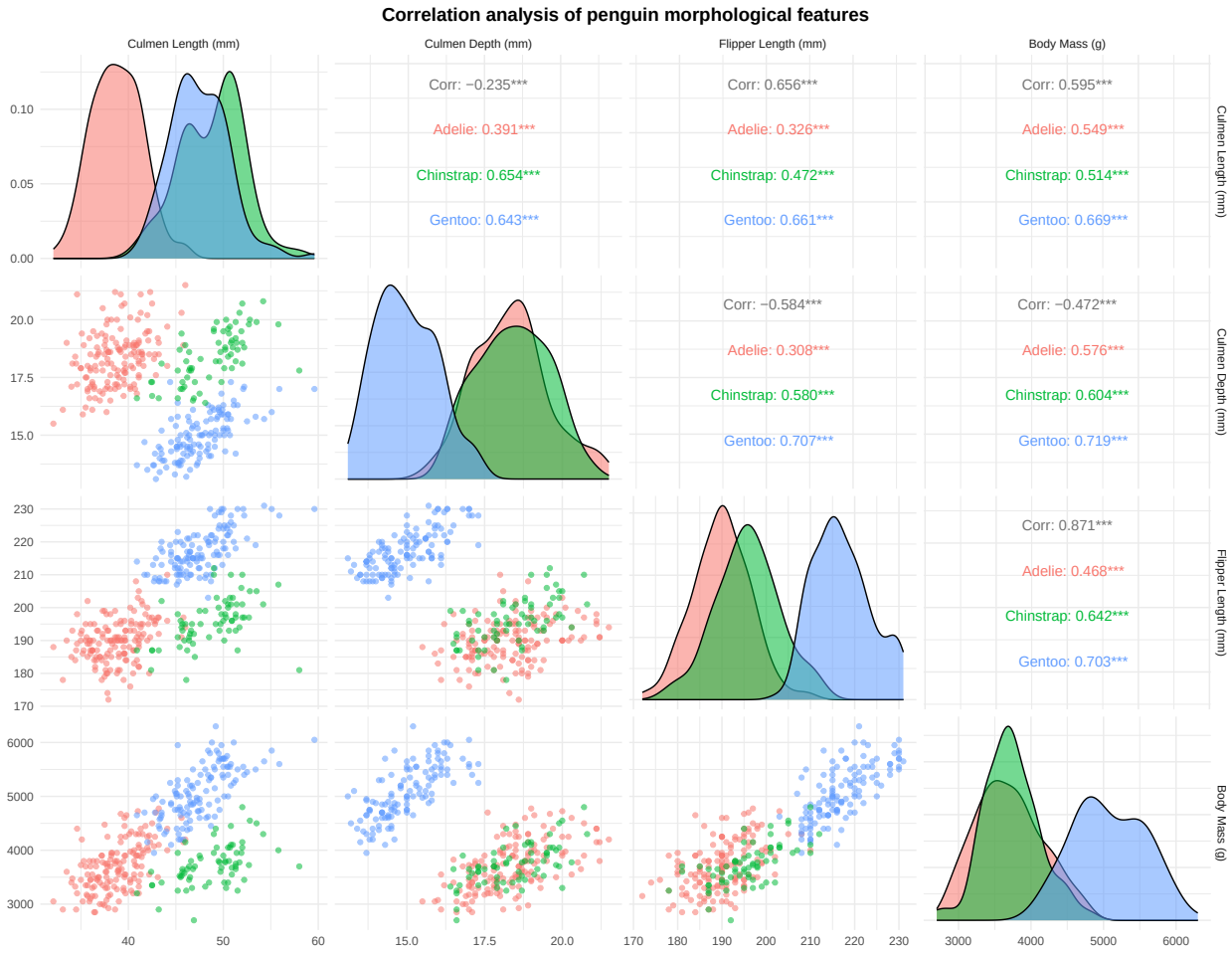
FIGURE 2: Correlation analysis

11. Edit column names for cosmetic purposes We will edit the column names so they look better in the resulting plot.

```
colnames(penguins)[which(colnames(penguins) == "length")] <- "Culmen Length (mm)"
colnames(penguins)[which(colnames(penguins) == "depth")] <- "Culmen Depth (mm)"
colnames(penguins)[which(colnames(penguins) == "flipper_length")] <- "Flipper Length (mm)"
colnames(penguins)[which(colnames(penguins) == "body_mass")] <- "Body Mass (g)"
```

12. Generate scatterplot matrix We will generate a scatterplot matrix to visualize the relationships between the numerical variables in the dataset. We will use the package GGally to generate the scatterplot matrix, and more information can be found in the package documentation. https://www.rdocumentation.org/packages/GGally/versions/2.2.1

```
ggpairs(penguins,
  columns = c("Culmen Length (mm)", "Culmen Depth (mm)", "Flipper Length (mm)", "Body Mass (g)"),
  aes(color = species, alpha = 0.7)
) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    legend.title = element_text(face = "bold"),
    legend.position = "right"
  ) +
  labs(title = "Correlation analysis of penguin morphological features",
    color = "Penguin Species"
  )
```



Scatterplots of each pair of numeric variable are drawn on the left part of the figure. Pearson correlation is displayed on the right. Variable distribution is available on the diagonal.