MAS8403 Final Assignment Report - Palmer Penguins

Sandra M Nino Arbelaez

2023-10-20

Contents

0.1 Exploratory Data Analysis 1
0.2 Distribution of our data 2
0.3 Estimates for the parameters of the distribution 4
0.4 Sexing Analysis 6
0.5 Island Analysis 8
0.6 Conclusion 8
0.7 References 8

0.1 Exploratory Data Analysis

The aim of an Exploratory Data Analysis is to identify relationships between the variables, patterns or trends in our data, identify outliers in each variable, etc. To achieve this, we are going to make use of statistics and visual representations to understand better our data.

The Palmer penguins dataset contains 200 penguins and 8 variables. We have three categorical variables, which are, the specie of the penguin, the island on which the penguin lives and the sex. Also, there are four quantitative variables which describes the characteristics of each penguin. These are the bill length in millimeters, the bill depth in millimeters, the flipper length in millimeters, and the body mass in grams. Finally, the measurements were taken between 2007 and 2009.

From the box plots we can clearly explain the following variables:

- **Bill length:** Adelie species have shorter bills independent from the island the penguin is from, however, the differences between the bill length of male Adelie and female are not notorious. When we compare the penguins living in Biscoe, we can assume that the penguins with longer bills are male Gentoo specie with a minimum of around 45 millimeters. When we compare the penguins living in Dream island, we can assume that Chinstrap species have longer bills and the female from this specie tend to be shorter. However, the distribution for the Chinstrap female specie is positive skewed.
- **Bill depth:** Gentoo species always have less deeper bills and the island is independent because they live in just the Biscoe island. It is difficult to compare the bill depth of Adelie species because they tend to be similar, however, for male Adelie species the bill is deeper than female, but this difference is not notorious. Also, for this particular specie we have outliers in the Torgersen and Dream island. When comparing the bill depth of Chinstrap specie, the female penguin has shorter depth than the male and the comparison is independent from the island because they just live in Dream.
- **Flipper length:** Gentoo species have longer flippers than any other specie. For this specie, the male penguin have longer flippers than the female. The differences between Adelie and Chinstrap is that Chinstrap species have longer flippers and male penguins for both species have longer flippers than female. The differences between male and female Adelie species are more difficult to perceive and the distribution between islands is also similar. Moreover, there are some outliers for this specie in different islands.
- **Body mass:** Gentoo species have greater body mass than any other specie. For this specie, the male penguins are heavier than the female. The differences between Adelie and Chinstrap is that

female Chinstrap species have greater body mass than the Adelie specie, and the male Adelie penguins have greater body mass than the Chinstrap specie. The differences between male and female Adelie species are more difficult to perceive between islands, however, male penguins have greater body mass compare to female. There are some outliers for the males Gentoo species and Chinstrap specie.

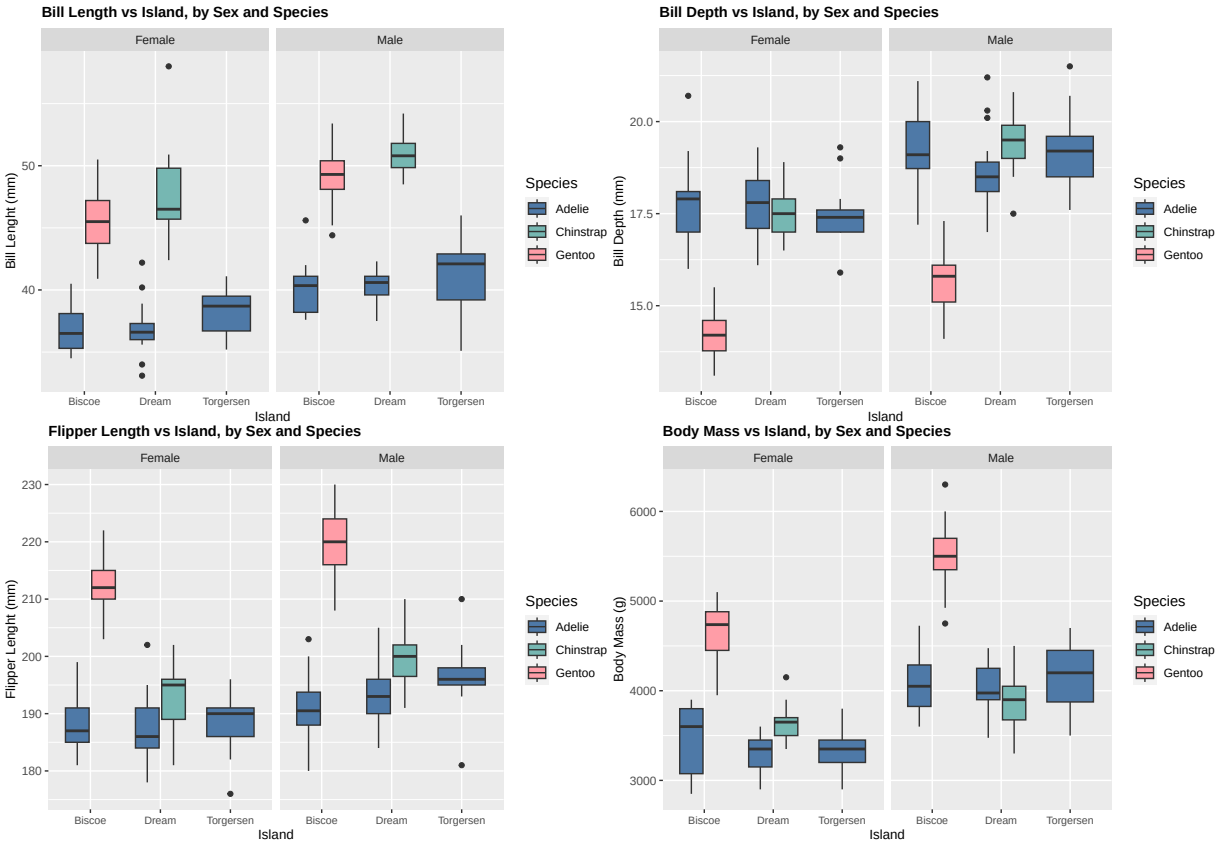


Figure 1: Boxplots between each quantitative variable vs Sex, Specie and Island

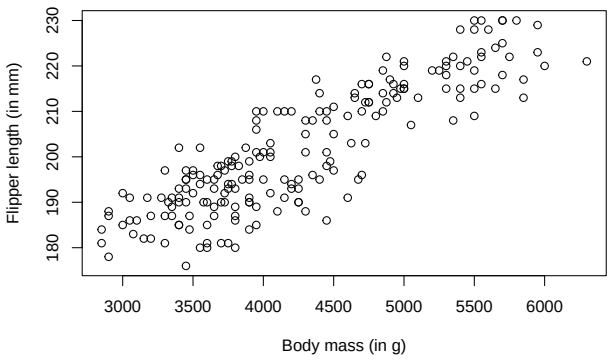


Figure 2: Scatterplot between Body mass (in grams) and Flipper length (in millimetters)

From this scatterplot we can see that there is a linear relationship between two quantitative variables: the body mass and the flipper length. From the analysis we have done in the previous boxplots we see that these variables behave similar for each penguin, therefore, we can assume this relationship is strong.

0.2 Distribution of our data

In order to determine the underlying distribution of the variable bill length, we are going to use some visual representations. We are going to plot a **histogram**, which is a common plot that shows the distribution of a variable. Then, having in mind that bill length is a continuous variable, we will overlay

the **Probability Density Function (PDF)**, which describes the shape of the distribution. If the PDF fits the histogram, we can say that the distribution fits our data. For this case, we are going to try to fit a normal distribution.

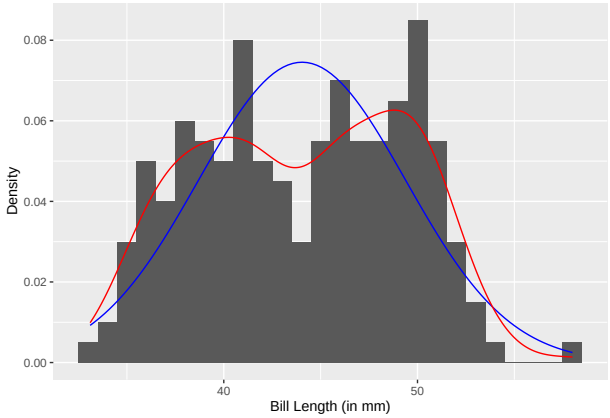


Figure 3: Histogram for bill length with the theoretical Probability Density Function overlaid for a normal distribution

It is very clear from the figure that the PDF (blue line) does not fit the histogram. Therefore, we can say it does not follow a normal distribution completely. Also, by drawing the empirical density (red line), it suggest the distribution is multimodal. We do not bother to use a Q-Q plot because, again, it does not follow a normal distribution.

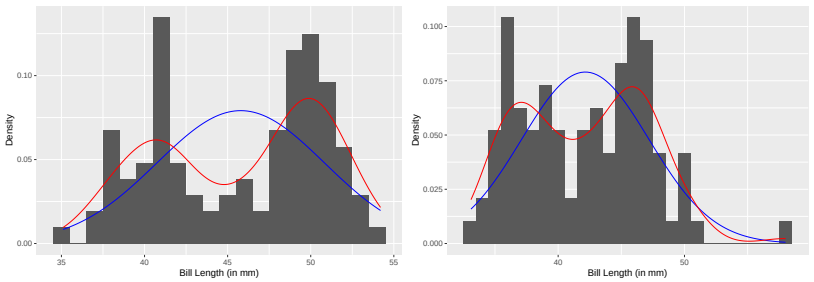


Figure 4: a) Histogram for bill length for the male penguins overlaid by the PDF. b) Histogram for bill length for the female penguins overlaid by the PDF.

We are splitting our penguins dataset by sex to try to fit a normal distribution. However, it is clear from the figures that they do not follow a normal distribution by looking at the PDF (blue line) and they suggest the distribution is also multimodal by the empirical density (red line).

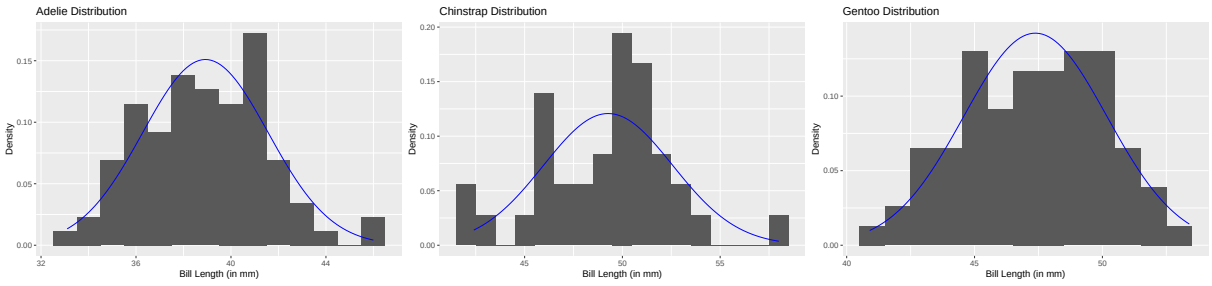


Figure 5: a) Histogram for bill length for Adelie specie overlaid by the PDF. b) Histogram for bill length for Chinstrap specie overlaid by the PDF. c) Histogram for bill length for Gentoo specie overlaid by the PDF

Now, when we split our data by species, we can see better results when fitting a normal distribution for bill length. For each figure representing a specie, it does not follow completely a normal distribution

because we have less data by doing this split. However, we can see that the PDFs for each specie seems to fit the histograms, which is an indication that the distribution are closer to a normal distribution.

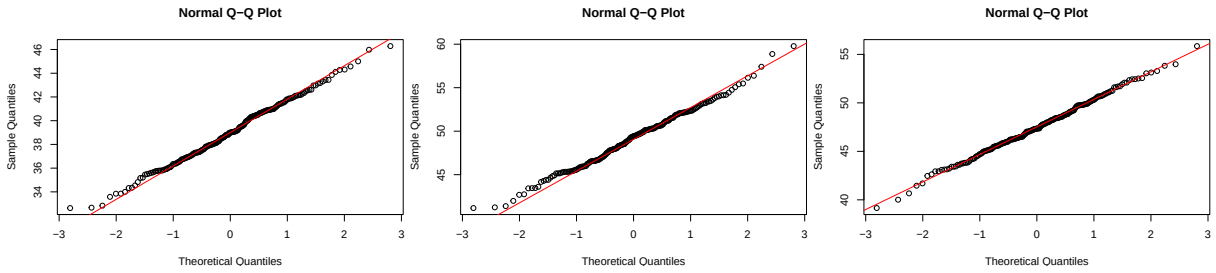


Figure 6: a) Q-Q plot for Adelie specie. b) Q-Q plot for Chinstrap specie c) Q-Q plot for Gentoo Specie

Q-Q plots are also a visual method to compare our data to a normal distribution. We can see from the Q-Q plots that the three distributions are close to follow a normal distribution because the points fall approximately around a diagonal line.

We have done visual analysis of the distribution for bill length by plotting the PDF over the histogram and comparing if it fits the bell shape to conclude it follows a normal distribution. However, we have proved that this method is not enough to estimate population proportions. Thus, there are some significance tests called "Goodness-of-fit" which measure "how well the distribution of the data fits a probability model" (Verzani, 2014). They are hypothesis tests where the null and alternative hypothesis are:

H_0: The sample data come from the stated distribution H_1: The sample data do not come from the stated distribution

For continuous distributions, some tests that can be applied are:

- **Kolmogorov-Smirnov:** We have to specify the distribution and we need to know the parameters of it.
- **Shapiro-Wilk:** Allows us to perform normality tests. It is based on the ideas behind the Q-Q plot.
- **Jarque-Bera:** Allows us to perform normality tests. It is based on skewness and kurtosis measures.
- **Anderson-Darling:** It is restricted for specific distributions.

0.3 Estimates for the parameters of the distribution

In order to estimate the parameters of the distribution for bill length, we are going to use the Maximum Likelihood Estimation. Previously, we have conclude that the variable bill length is close to follow a **normal distribution** when we split our data by species. In this way, we are going to maximize the likelihood function to find the estimators for μ and σ.

1. We start by taking the logarithm of the Probability Density Function for a normal distribution:

$$\begin{aligned} f(x|\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ \log f(x|\mu,\sigma) &= \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) \\ &= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \end{aligned}$$

2. We calculate the log-likelihood function:

$$\begin{aligned} \ell(\mu,\sigma|x_1,\ldots,x_n) &= \log f(x_1|\mu,\sigma) + \cdots + \log f(x_n|\mu,\sigma) \\ &= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x_n-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{(x_n-\mu)^2}{2\sigma^2} \end{aligned}$$

3. First, we are going to find the estimator for μ. Therefore, we differentiate the log-likelihood function with respect to μ:

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell(\mu,\sigma|x_1,\ldots,x_n) &= \frac{\partial}{\partial\mu}(-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{(x_n-\mu)^2}{2\sigma^2}) \\ &= \frac{x_1-\mu}{\sigma^2} + \cdots + \frac{x_n-\mu}{\sigma^2} \\ &= \frac{1}{\sigma^2}[(x_1+\cdots+x_n) - n\mu]\end{aligned}$$

4. We maximize μ:

$$\begin{aligned}\frac{1}{\sigma^2}[(x_1+\cdots+x_n) - n\hat{\mu}] &= 0 \\ \frac{x_1+\cdots+x_n}{n} &= \hat{\mu} \\ \frac{1}{n}\sum_{i=1}^{n} x_i &= \hat{\mu} \\ \therefore \qquad \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

Therefore, we can conclude that $\hat{\mu}$ is the mean.

5. Now, we are going to find the estimator for σ. Therefore, we differentiate the log-likelihood function with respect to σ:

$$\begin{aligned}\frac{\partial}{\partial\sigma}\ell(\mu,\sigma|x_1,\ldots,x_n) &= \frac{\partial}{\partial\sigma}[-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{(x_n-\mu)^2}{2\sigma^2}] \\ &= -\frac{n}{\sigma} + \frac{(x_1-\mu)^2}{\sigma^3} + \cdots + \frac{(x_n-\mu)^2}{\sigma^3} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\mu)^2 + \cdots + (x_n-\mu)^2]\end{aligned}$$

6. We maximize σ:

$$\begin{aligned}-\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\hat{\mu})^2 + \cdots + (x_n-\hat{\mu})^2] &= 0 \\ \frac{(x_1-\hat{\mu})^2 + \cdots + (x_n-\hat{\mu})^2}{n} &= \hat{\sigma}^2 \\ \sqrt{\frac{(x_1-\hat{\mu})^2 + \cdots + (x_n-\hat{\mu})^2}{n}} &= \hat{\sigma} \\ \sqrt{\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}} &= \hat{\sigma} \\ \therefore \qquad \hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2}{n}}\end{aligned}$$

Therefore, we can conclude that $\hat{\sigma}$ is the standard deviation of our data.

Now we can find the values for $\hat{\mu}$ and $\hat{\sigma}$ for the three different distributions.

Adelie mean and standard deviation (using formula from step 4 and 6, respectively, and `mean` and `sd` from R):

```
## [1] 38.92414
```

```
## [1] 2.626959
```

```
## [1] 38.92414
```

```
## [1] 2.642187
```

Chinstrap mean and standard deviation (using `mean` and `sd` from R):

```
## [1] 49.26944
```

```
## [1] 3.303662
```

Gentoo mean and standard deviation (using `mean` and `sd` from R):

```
## [1] 47.37792
```

```
## [1] 2.805053
```

0.4 Sexing Analysis

In order to determine the best variables to distinguish between male and female penguins, first we are going to do the analysis based on the visual representations from the Exploratory Data Analysis.

For the three species we see that the male bill is longer than the female penguins. However, there is a more noticeable difference in Gentoo and Chinstrap and might be difficult to identify for Adelie specie because they seem more similar. However, if we focused on the island, we can assume that in Biscoe, male Gentoo have longer bills than Adelie, and in Dream, male Chinstrap have longer bills than Adelie. Also, the bill is deeper for all three male species than female. If we focus on the island, in Biscoe, female Gentoo have less deeper than male and any sex of Adelie, and in Dream, male Chinstrap have deeper than Adelie. Regarding the flippers, for the three species, female penguins have shorter flippers than male for three species. This is more notorious for Gentoo and might be difficult to see the differences in Adelie and Chinstrap. Finally, the behaviour for the body mass is similar to the flippers. In conclusion, the characteristics to identify the sex is highly affected by the specie and the island on which the penguin lives. However, from the initial box plots we cannot determine which variable is more relevant than other to distinguish the sex.

Therefore, we can perform t-tests to compare the means between the male and female penguins by each quantitative variable. By doing this we can see the effect of sex in the characteristics of the penguins. Due to page limit, I am going to show just the results of two t-tests which gives us valuable information for the sexing analysis. These are the ones that show a greater difference on the mean, thus, we can assume that those characteristics are more relevant to identify the sex of the penguins. It is important to mention that all t-tests performed for each quantitative variable reject the null hypothesis.

To determine whether the mean of the body mass of two different sexes of penguins is equal, we need to check if the variances are equal or not by doing a Bartlett test.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  body_mass_g by sex
## Bartlett's K-squared = 3.1326, df = 1, p-value = 0.07674
```

We have a p-value of 0.07674, which is greater than 0.05, so the assumption of equal variances is valid for our t-test.

Now, we are going to perform a t-test for the body mass of the penguins affected by sex.

```
##
##  Two Sample t-test
##
## data:  body_mass_g by sex
## t = -6.6883, df = 198, p-value = 2.255e-10
## alternative hypothesis: true difference in means between group female and group male is not equal
## 95 percent confidence interval:
##  -908.7787 -494.9072
## sample estimates:
```

```
## mean in group female   mean in group male
##             3904.167             4606.010
```

We have a very small p-value, which means that we have a very strong evidence against the null hypothesis (the mean for the body masses are equal), so we reject it and go with the alternative hypothesis (the mean for the body masses are not equal). There is a difference in the means of the two groups, and the male penguins are bigger than female penguins, so we can assume that the body mass is affected by sex. However, we have to keep in mind if the body mass is affected by another variable such as the island or the species.

To determine whether the mean of the flipper length of two different sexes of penguins are equal, we need to check if the variances are equal or not by doing a Bartlett test.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  flipper_length_mm by sex
## Bartlett's K-squared = 1.7047, df = 1, p-value = 0.1917
```

We have a p-value of 0.1917, which is greater than 0.05, so the assumption of equal variances is valid for our t-test.

We can continue by performing a t-test for the flipper length of the penguins to see if is affected by sex.

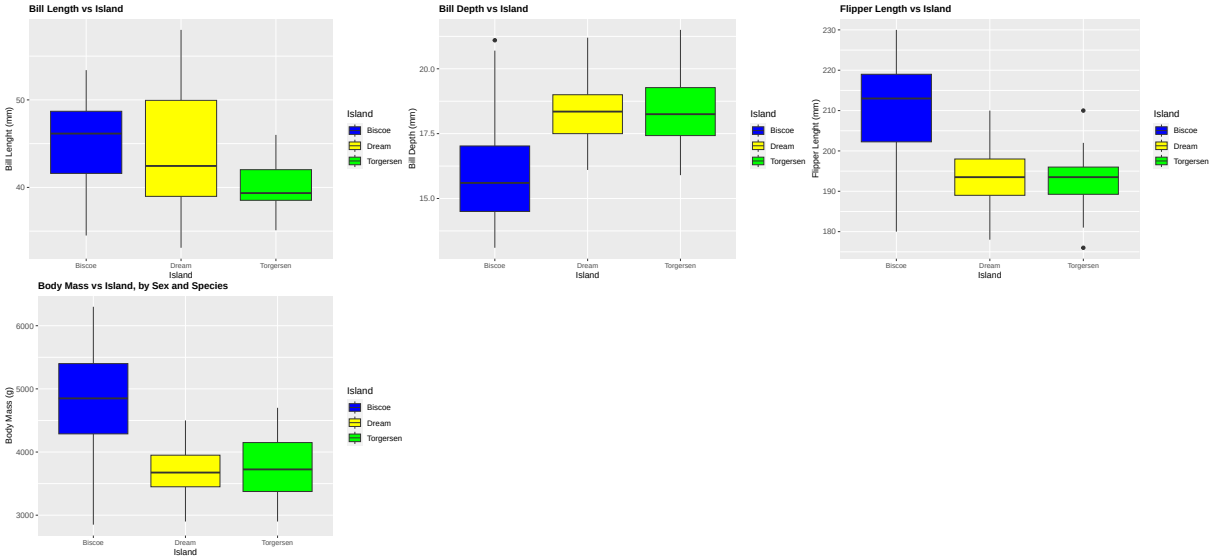
```
##
##  Two Sample t-test
##
## data:  flipper_length_mm by sex
## t = -3.8954, df = 198, p-value = 0.000134
## alternative hypothesis: true difference in means between group female and group male is not equal
## 95 percent confidence interval:
##  -10.995077  -3.604282
## sample estimates:
## mean in group female   mean in group male
##             197.9792             205.2788
```

We have a small p-value, which means that we have a very strong evidence against the null hypothesis (the means of the flippers length are equal for male and female), so we reject it and go with the alternative hypothesis (the means of the flippers length are not equal for male and female). This test suggest that the flipper length is affected by sex because there is a difference in the means of the two groups and the male penguins have longer flippers than female penguins. Again, this characteristic can be affected by specie and/or the island.

We have used these t-test for each variable and the boxplots from the beginning of the report to conclude information about the variables that can help us to distinguish between male and female penguins. However, there are other methods to identify better the relationship between variables including the strength of it.

In this case, we want to predict a categorical variable which has two possible values: male and female. Then, a better approach would be making use of **Logistic Regression**. This is an extension to the linear regression model and it "covers the situation where the response variable is a binary variable"(Varzani, 2014). That means it just distinguishes between two classes. It uses a sigmoid function (the cumulative distribution function of the logistic distribution) which it is S-shaped, so it transforms the input values between 0's and 1's, then there is a cut-off threshold to classify the output into one class or the other. "The model assigns weights to the predictors based on how they impact the target variable and combines them to calculate the normalized score" (Chugh, 2023)

0.5 Island Analysis



In order to determine if there is a significant difference in the physical characteristics of penguins living on different island, first we are going to do the analysis based on the visual representations.

For the bill length we see that the boxplots from Biscoe and Dream island are overlaid so it might be difficult to identify the differences. In Torgersen we just have one specie, therefore, it might be easy to identify the characteristics and compare them with other islands. For the bill depth, it is clear from the box plots that the penguins from Biscoe have less deeper bills but it is not easy to identify this characteristic between the penguins from Dream and Torgersen island. Regarding the flipper length and the body mass, in Biscoe island the penguins have longer flippers and greater body mass, but it is not straightforward to this comparison between penguins living in Dream and Torgersen islands.

We cannot apply Logistic Regression for this task because, by default, it is limited for two classes. However, there are some extensions like One vs All classification model that allows to apply the same idea behind logistic regression but for multi-class classification, although this requires to transform the problem into multiple binary classification problems. This approach might be reasonable for this specific case because we just have three classes, but it becomes inefficient when the number of classes increases. This problem uses a softmax function which can be used in multi-class classification to predict a single label from many classes.

Therefore, we can make use of one-way ANOVA. This approach is a "generalisation of the t-test for for two independent samples, allowing us to compare population means for several independent samples" (Verzani, 2014). In this way, we can compare the means between the three different islands for each quantitative variable. ANOVA will have a known distribution: F-distribution with $k-1$ and $n-k$ degrees of freedom. So it produces an F test statistic to see if there is a significant differences between the groups.

0.6 Conclusion

We have used statistics and visual representations to analyze the differences between the characteristics of the Palmer penguins based on the sex, the island on which the penguins lives and the specie. Also, we were able to identify the distributions of our variables, however, there was a limitation because our dataset is so small to fit a distribution properly. We found limitations to fit a distribution and suggested new Goodness of fit methods to achieve this. Moreover, we demonstrate that t-tests have limitations to provide strong conclusions regarding the sexing and the relevance of the island, therefore, we introduce new methods to achieve this such as the logistic regression and one-way ANOVA.

0.7 References

Verzani, John (2014). *Using R for introductory statistics* (Second edition..). Boca Raton : CRC Press, Taylor & Francis Group;

Ricci, Vito. (February 2005). *Fitting distributions with R.* https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Muller, Marie L, Dutang, Christophe (2023, March 25). *Overview of the fitdistrplus package* https://cran.r-project.org/web/packages/fitdistrplus/vignettes/fitdistrplus_vignette.html

Chugh, Vidhi. (2023, March). *Logistic Regression in R Tutorial* https://www.datacamp.com/tutorial/logistic-regression-R

StatQuest with Josh Starmer. (September 10, 2018). *Maximum Likelihood For the Normal Distribution, step-by-step!!!.* Youtube. https://www.youtube.com/watch?v=Dn6b9fCIUpM