Palmer Penguins Analysis Report

Dimitrios Poulimenos - 200291237

Date: 17-10-2023

Introduction

Palmer penguins consist of 3 species (Adelie, Chinstrap, Gentoo) and there are 3 islands (Torgensen, Dream, Biscoe) that host these species. In this analysis, I will dive into the Palmer Penguins dataset, which was originally collected by researchers at the Palmer Station, a United States research station located on Anvers Island in Antarctica, exploring its contents, and gaining insights into the penguin population of the islands there. I am going to use a combination of statistics, visualizations, and statistical techniques to find patterns, relationships, and valuable information about the Palmer Penguins dataset. However, I am going to use only a subset of the dataset of Palmer penguins and the findings of the analysis may not reflect to the actual data.

Exploratory Data Analysis

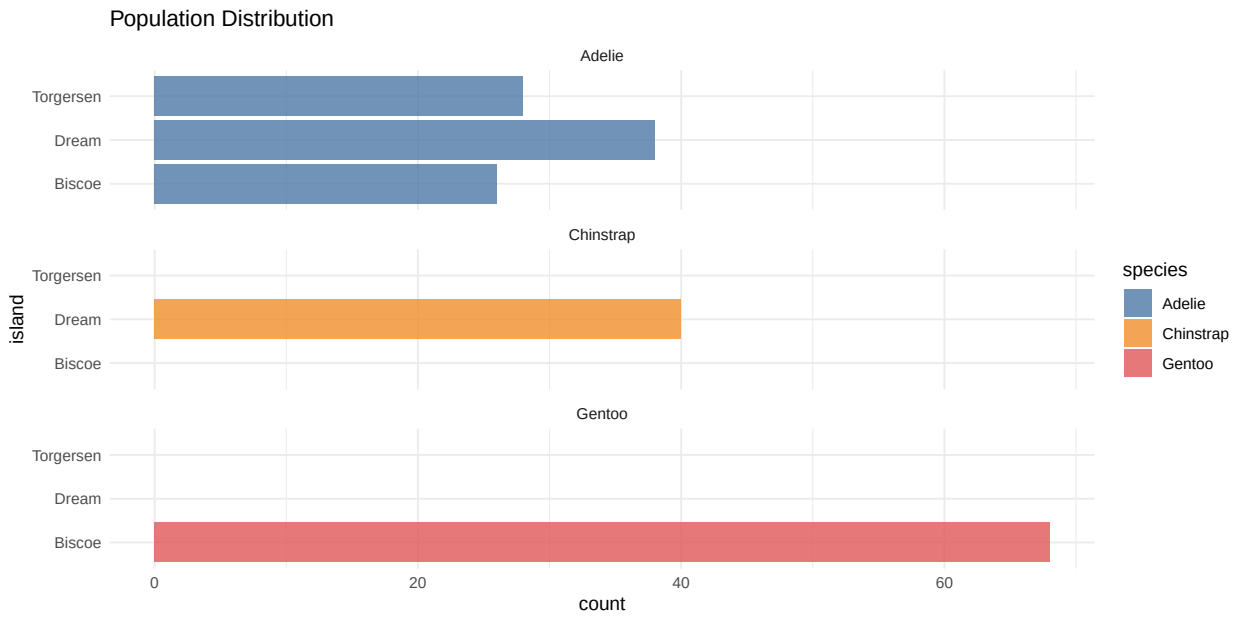
Subset Summary

Below is the summary of the subset that I have to conduct an analysis. It is important to point out that the number of males is slightly bigger than females and also that the dominant species is Adelie. Furthermore, Biscoe Island is the most populated with 94 penguins.

```
##       species         island   bill_length_mm  bill_depth_mm
##  Adelie   :92   Biscoe   :94   Min.   :32.10   Min.   :13.30
##  Chinstrap:40   Dream    :78   1st Qu.:39.20   1st Qu.:15.70
##  Gentoo   :68   Torgersen:28   Median :43.95   Median :17.50
##                                Mean   :43.91   Mean   :17.24
##                                3rd Qu.:48.70   3rd Qu.:18.73
##                                Max.   :59.60   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female: 99   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3500   male  :101   1st Qu.:2007
##  Median :196.0     Median :3975                Median :2008
##  Mean   :200.6     Mean   :4175                Mean   :2008
##  3rd Qu.:213.2     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
```

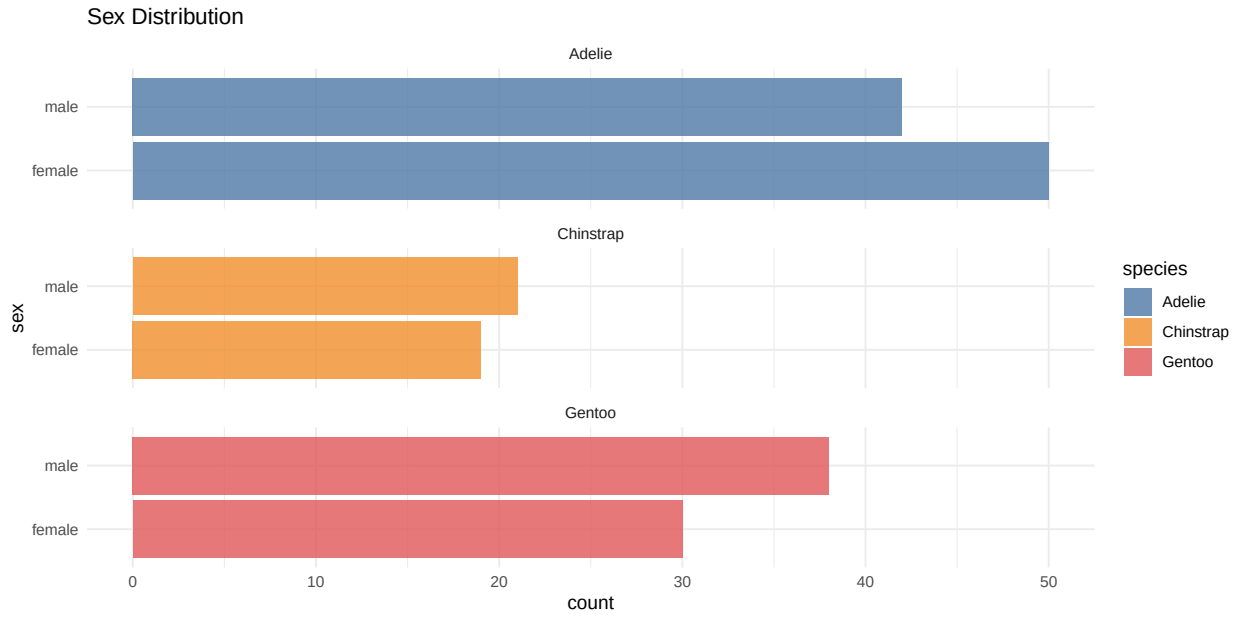
Species distribution in each island

The plot below depicts the distribution of three penguin species across different islands and their respective populations. Notably, the Adelie species is the only one to inhabit all three islands, while the other two species have more restricted habitats. The Chinstrap species exclusively resides on Dream Island, and the Gentoo species is found solely on Biscoe Island.



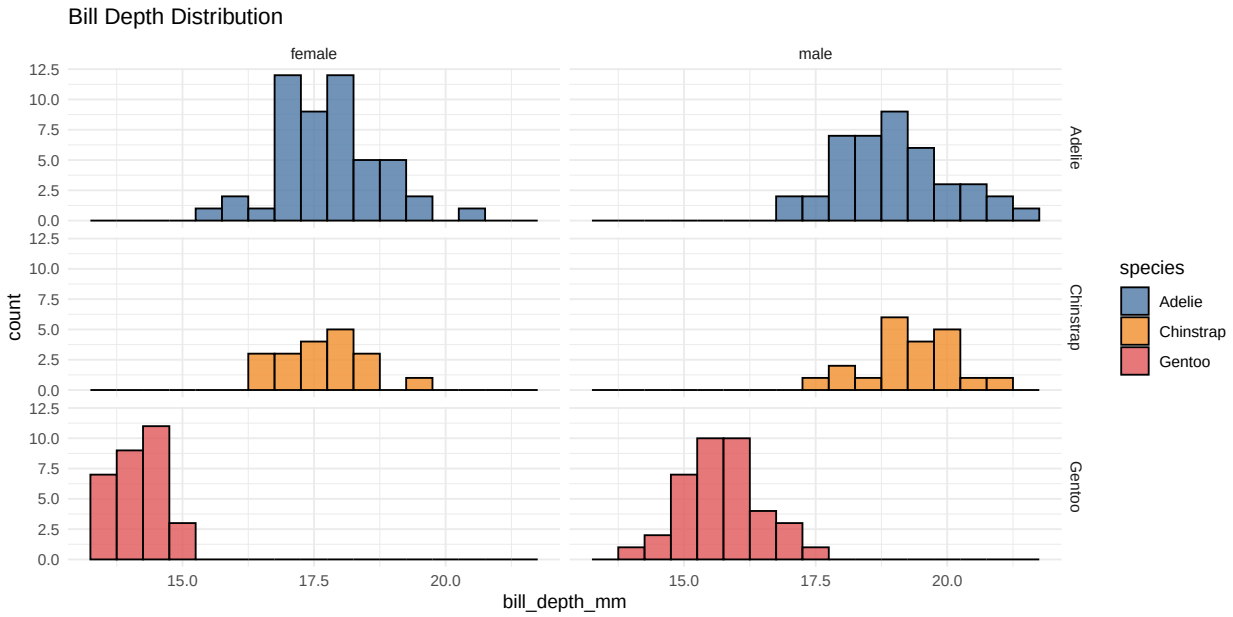
Sex distribution in each species

In the below plot we observe distinct sex distributions across different penguin species. Specifically, for Gentoo and Chinstrap species, the female population predominates, whereas in the Adelie species, the male population is more prevalent.



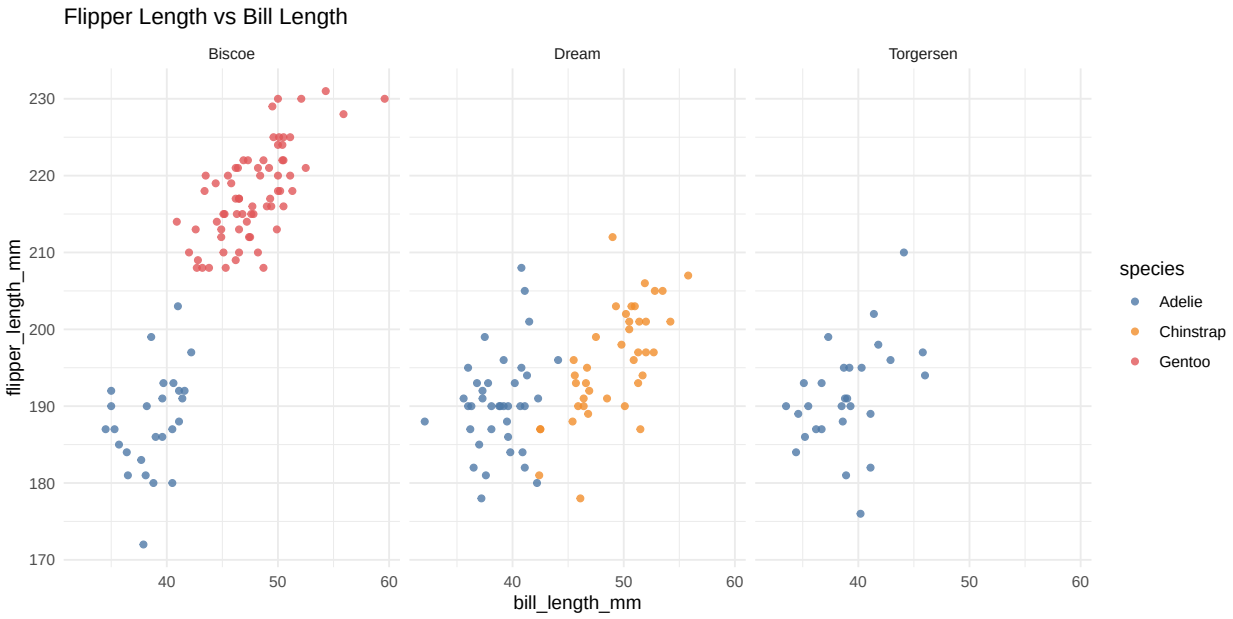
Bill depth distribution

The histograms below illustrate the distribution of bill depths among the three penguin species, differentiated by gender. It's quite evident that, across all three species, male penguins consistently exhibit larger bill depths compared to their female counterparts. Notably, within these species, Gentoo penguins stand out with the smallest bill depth, both among males and females, in comparison to the other two species.



Flipper length vs Bill lenght scatterplot for Islands

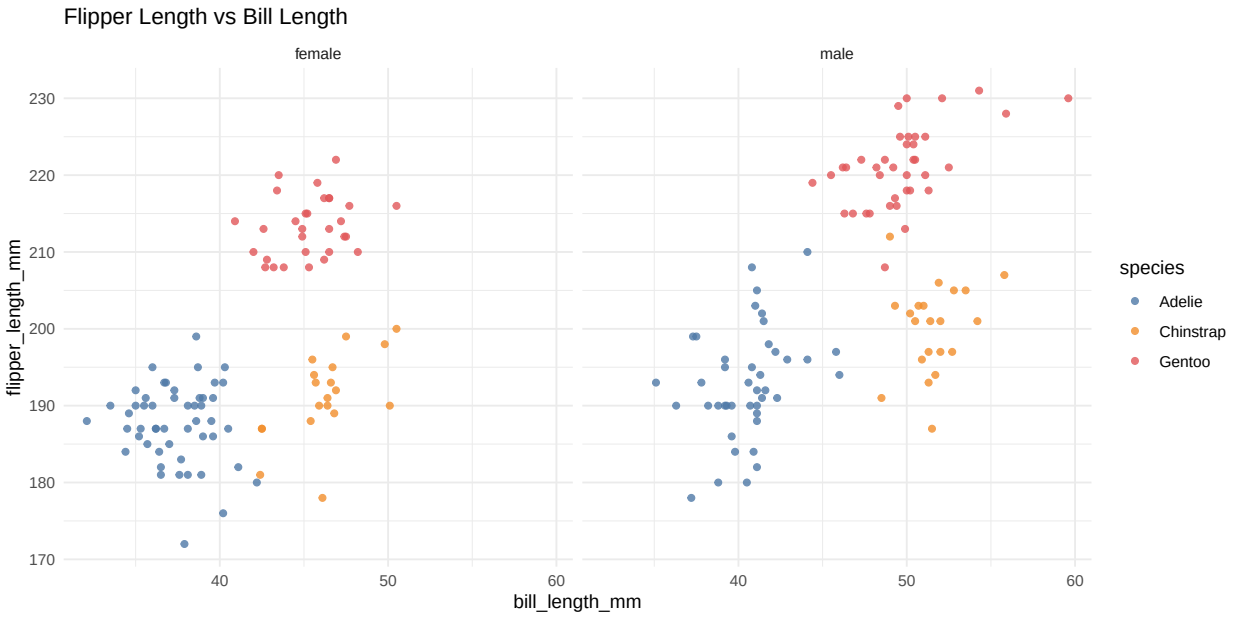
The scatter plot clearly illustrates that Gentoo penguins exhibit the largest flipper lengths in comparison to the other two species. Following closely behind are the Chinstrap penguins, which have the second-longest flipper lengths, with the Adelie species trailing behind. Additionally, it's noteworthy that Adelie penguins consistently display smaller bill and flipper lengths when compared to the other two species across all the islands they inhabit.



Flipper length vs Bill lenght scatterplot for Sex

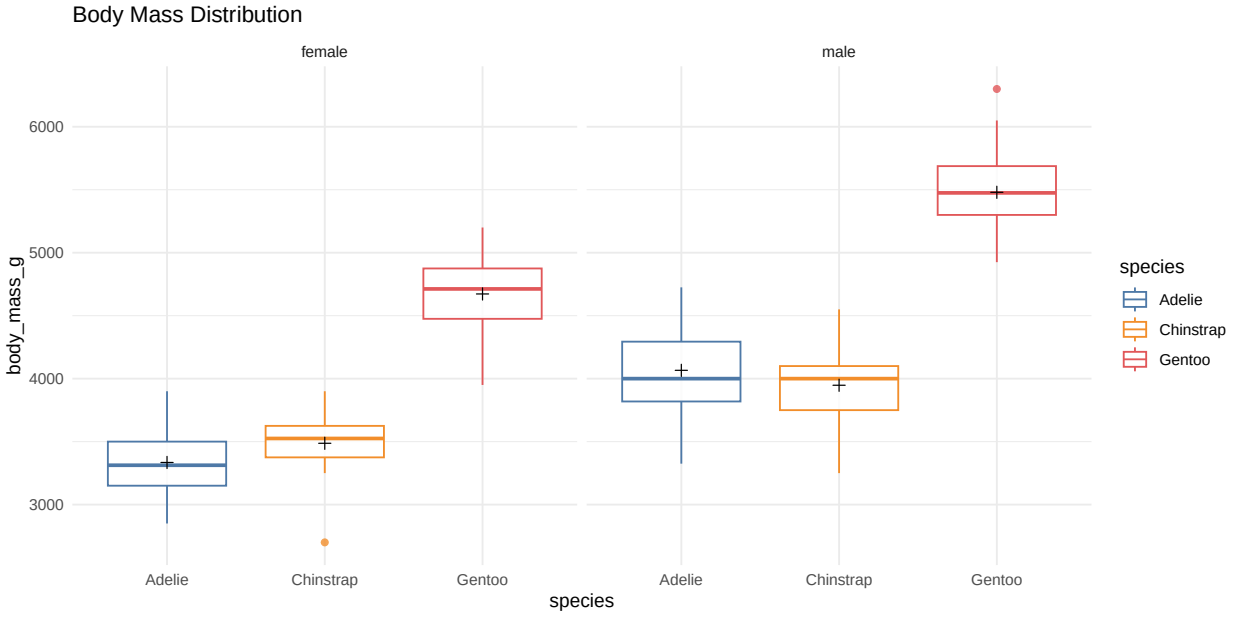
The scatter plot presented below reveals a notable trend: male penguins consistently exhibit larger flipper length and bill length. However, it's worth noting that within the Adelie species, there is a remarkable

similarity in both flipper length and bill length between male and female individuals. This pattern suggests that flipper length and bill length can serve as reliable indicators of sex differentiation across various penguin species, except in the case of Adelie penguins where these characteristics show less variability between genders.



Body mass distribution

Regarding body mass, the box plot data unmistakably reveals that males of all three species consistently exhibit larger body mass than their female counterparts. It's worth noting that within the Gentoo species, females tend to have a notably greater body mass compared to the males of both the Adelie and Chinstrap species.



Sex Indentification with Hypothesis Testing

In hypothesis testing, we usually compare two hypotheses. The first, called the null hypothesis H_0, generally reflects the status-quo. The second, called the alternative hypothesis H_1, is the conclusion to be reached if we find evidence to reject H_0. The objective is then to determine whether the null hypothesis is plausible in light of a sample of data.

To determine which variables effectively distinguish between male and female penguins, I plan to conduct a hypothesis test, specifically a two-sample t-test. This test aims to assess whether certain variables within a subset of penguins can reliably differentiate between males and females of the same species. In this particular analysis, I will focus on the "body mass" variable and compare it between male and female penguins. By doing so, I intend to evaluate whether body mass provides valuable insights for distinguishing the gender of the penguins.

Equal Variance Assumption

We have to test the hypothesis for variances of the body mass between male and female penguins using this:

$$H_0 : \sigma_A^2 = \sigma_B^2$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

Let's assume that A stands for male and B for female

A commonly used test which comes in built to base R is Bartlett's test, which we carry out using the bartlett.test command. If our test gives a p-value less than 0.05, we reject the null hypothesis, and therefore cannot assume the population variances are equal.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  body_mass_g by sex
## Bartlett's K-squared = 2.7821, df = 1, p-value = 0.09533
```

Two sample t-test

From Bartlett test results we can see that the p-value is bigger than 0.05 which means that we have to follow the equal variances path. Now we are conducting this test known as the Welch test. In this case the test statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Again we can carry out this test in R using the t.test command and specifying the var.equal parameter to **"TRUE"**.

```
##
##  Two Sample t-test
##
## data:  male_penguins and female_penguins
## t = 7.8661, df = 198, p-value = 2.323e-13
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##   602.4947 1005.6561
## sample estimates:
## mean of x mean of y
##  4573.267  3769.192
```

Since the p-value is extremely small, much less than a typical significance level like 0.05, we reject the null hypothesis:

$$H_0 : \mu_A = \mu_B$$

This suggests strong evidence that there is a significant difference in body masses between male and female penguins.

Based on the comprehensive exploratory data analysis conducted on the selected subset, it can be concluded that the most reliable variables for identifying the sex of a penguin are body mass, flipper length, and bill length. These particular attributes consistently exhibit distinct patterns between male and female penguins across various species, making them robust indicators of sex differentiation. While the reliability of these indicators may vary across species, these insights provide valuable and suggestive information for inferring the sex of a penguin, enhancing our ability to make accurate sex-based distinctions.

Characteristics Impact Based On Island

In order to be able to tell if the island the penguin is from appear to have a significant impact on any of its physical, characteristics we can use the analysis of variance or (ANOVA) in R. Using the aov() function we can answer to our question. Because only one species inhabits all three islands I will adjust my function to use only the Adelie species.

Using ANOVA for body mass

```
##             Df   Sum Sq Mean Sq F value Pr(>F)
## island       2   379392  189696   0.835  0.437
## Residuals   89 20225445  227252
```

Using ANOVA for flipper length

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2    154   76.77   1.773  0.176
## Residuals   89   3853   43.30
```

Using ANOVA for bill length

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2    1.1   0.548   0.076  0.927
## Residuals   89  643.0   7.225
```

According to the results from the ANOVA function above suggests that the island of origin does not appear to have a significant impact on body mass, flipper length and bill length as well. And that's because only one species inhabits all the three islands so if I take into consideration all species I will get false results.

Probability Distribution

Maximum likelihood estimation

In the forthcoming analysis, I aim to determine the optimal parameters, namely the mean and standard deviation, that effectively characterize the distribution of bill lengths in the penguin subset. To achieve this, I will employ the Normal Distribution, a widely-used statistical model. The following formula will be instrumental in estimating these parameters:

$$\log(\mathcal{L}(\mu, \sigma)) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right)$$

This formula expresses the logarithm of the likelihood function, where μ represents the mean and σ signifies the standard deviation. By maximizing this likelihood, I can obtain the most accurate parameter values that describe the bill lengths of the penguins in the analysis.

Below are the estimations of **mean** and **standard deviation**.

```
## Estimated Mean: 43.909
```

```
## Estimated Standard Deviation: 5.586673
```

Confidence Intervals

Given a confidence level $1 - \alpha$ (where α is the significance level), the formula for the confidence interval is:

Confidence Interval = [estimated mean − margin of error, estimated mean + margin of error]

Where:

"Estimated mean" is the point estimate of the population mean obtained through Maximum Likelihood Estimation (MLE).

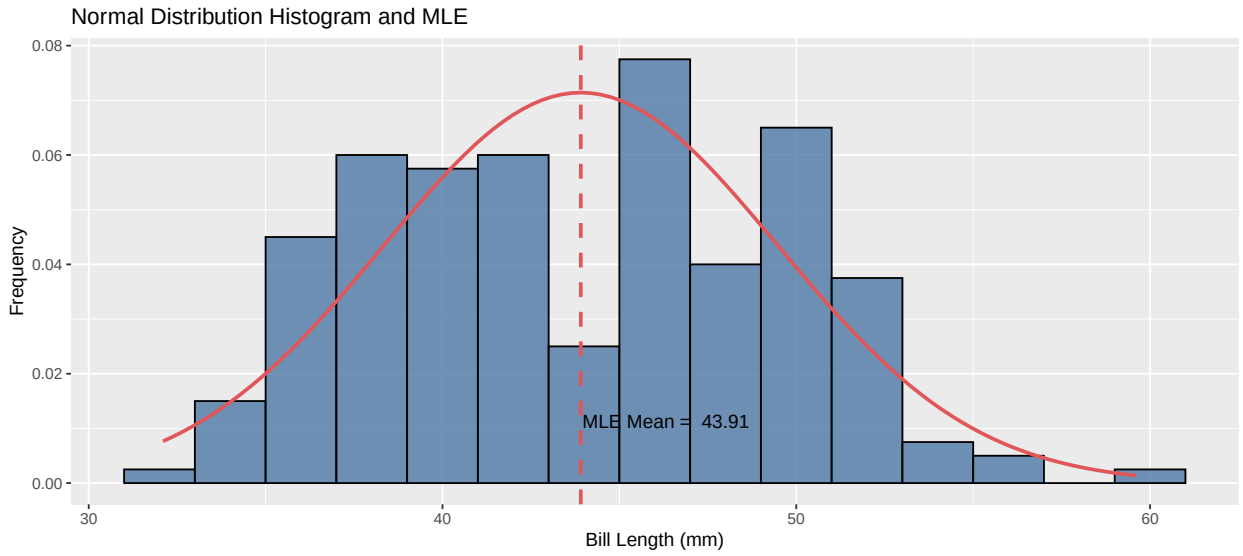
"Margin of error" is calculated as: Margin of Error = critical value × standard error

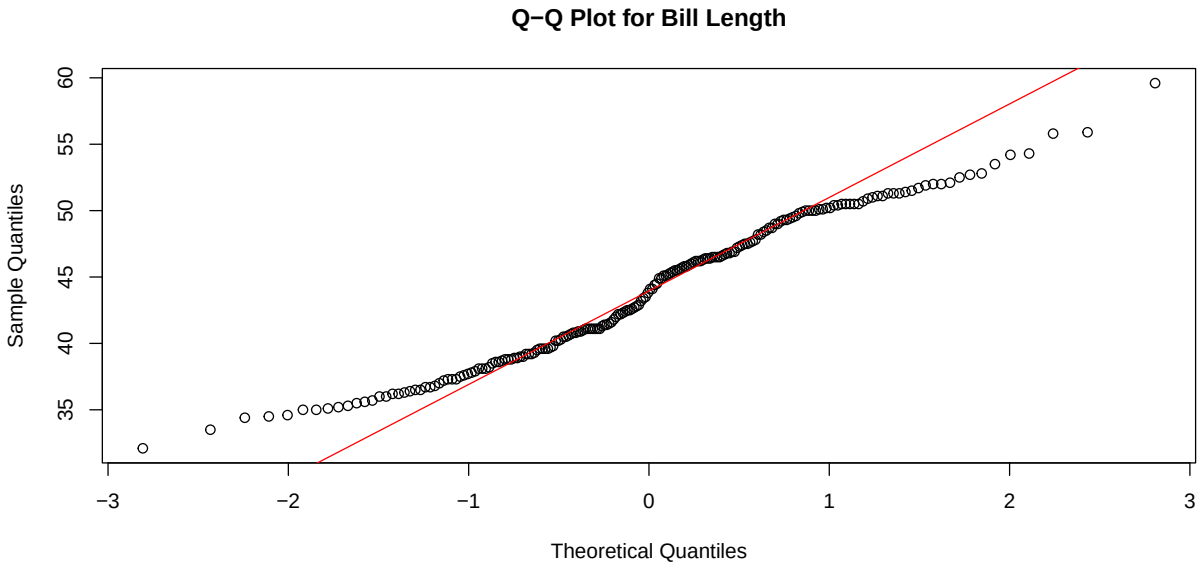
The "critical value" is obtained from the standard normal distribution (e.g., using the quantile function qnorm) and is determined by the desired confidence level. For example, for a 95% confidence level, $\alpha = 0.05$, so the critical value corresponds to the 97.5th percentile of the standard normal distribution.

The "standard error" is calculated as: Standard Error = $\frac{\text{estimated standard deviation}}{\text{sample size}}$

```
## Confidence Interval for the estimated mean : [ 43.13474 ,  44.68326 ]
```

As we can see my estimation falls within the confidence interval which supports the validity of my estimation. However, this does not guarantee that my estimation is exactly equal to the true population parameter. Now with the estimated mean and standard deviation in hand, it's time to visualize the bill length distribution of the penguins and assess whether a normal distribution is a suitable descriptor.





In conclusion, based on the analysis of the histogram above but also the QQplot, it appears that a normal distribution is not an appropriate model for describing the distribution of the bill length variable among the penguins although it looks like it follows the normal distribution in some points. Additionally, it is important to note that estimating population proportions using this subset of penguin data may not be a reliable method. To make accurate population estimates, a comprehensive data set encompassing the entire penguin population would be necessary.

Critical evaluation

In the course of my analysis, I generated several plots and gained valuable insights into the subset of the Palmer penguins dataset. However, I encountered limitations, primarily related to the constraints on the number of pages available (8 pages) for my analysis. Additionally, as I continue my data analysis journey, I recognize the need to enhance my proficiency in R and develop a deeper understanding of data visualization techniques. This will enable me to more effectively manage and interpret diagrams and charts, ultimately enhancing the quality and depth of my future analyses.

References

ggplot: Hadley Wickham. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

dplyr: Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. https://CRAN.R-project.org/package=dplyr

knitr: Yihui Xie. (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33. https://yihui.org/knitr/

Palmer Penguin Dataset: Gorman, K. B., Williams, T. D., Fraser, W. R., & G. C. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE, 9(3), e90081. https://doi.org/10.1371/journal.pone.0090081

R For Data Science 2nd Edition: Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. R for Data Science 2nd Edition. Sebastopol, CA: O'Reilly Media.

Course Material: [Dimitrios Poulimenos]. (2023). [Statistical Foundations of Data Science]. [Newcastle University].