Data Analysis of 200 Penguins from Antarctica

Kuanrou Fan

Semester 1, 2024/2025

In this report, data from three species of penguins—Adelie, Chinstrap, and Gentoo—from Biscoe, Dream, and Torgersen islands in Antarctica were analyzed. The entire dataset contains 200 rows and 8 columns. Each row provides information on a penguin from Antarctica, including its species, island, bill length, bill depth, flipper length, body mass, sex, and the year the data was recorded.

1 Exploratory Data Analysis of Penguin Sample

1.1 Overview of the Penguin Data

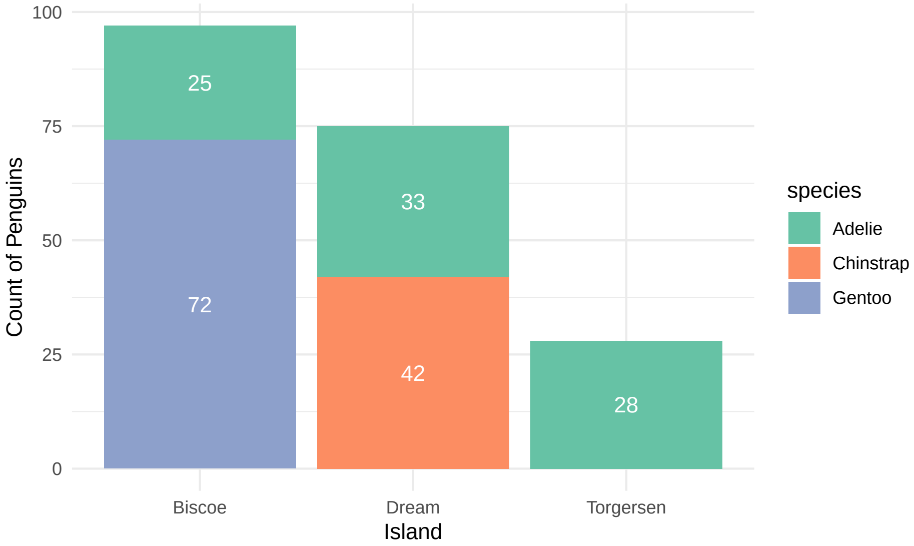
First, gain an overall understanding of the entire dataset.

```
##
## Attaching package: 'palmerpenguins'

## The following objects are masked from 'package:datasets':
##
##     penguins, penguins_raw

##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :86   Biscoe   :97   Min.   :32.10   Min.   :13.30
##  Chinstrap:42   Dream    :75   1st Qu.:39.20   1st Qu.:15.70
##  Gentoo   :72   Torgersen:28   Median :45.10   Median :17.30
##                                Mean   :44.09   Mean   :17.20
##                                3rd Qu.:49.00   3rd Qu.:18.62
##                                Max.   :58.00   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172       Min.   :2700   female: 96   Min.   :2007
##  1st Qu.:190       1st Qu.:3519   male  :104   1st Qu.:2007
##  Median :197       Median :4000                Median :2008
##  Mean   :201       Mean   :4189                Mean   :2008
##  3rd Qu.:214       3rd Qu.:4800                3rd Qu.:2009
##  Max.   :231       Max.   :5950                Max.   :2009
```

Based on the above statistical data, among the 200 penguins, the species with the highest number is Adelie, totaling 86, while the Chinstrap penguins are the least, with only 42. These penguins are distributed across three islands, with the highest number of 97 penguins on Biscoe Island, and the fewest, only 28, on Torgersen Island. The penguin bill length ranges from 32.10 mm to 58.00 mm, with an average of 44.09 mm. The bill depth ranges from 13.40 mm to 21.50 mm, with an average of 17.20 mm. The flipper length ranges from 172 mm to 231 mm, with an average of 201 mm. The penguins' body mass ranges from 2700 g to 5950 g, with an average of 4189 g. The gender distribution is relatively balanced, with 96 females and 104 males.

1.2 Distribution of the Three Penguin Species

Here is a statistical plot created using R to show the distribution of three different penguin species across three islands:



According to this chart, we can visually observe that Gentoo penguins are only found on Biscoe Island, Chinstrap penguins are only found on Dream Island, while Adelie penguins are distributed across all three islands. Regarding the possible reasons for the distribution of Adelie penguins across these three islands, I propose two hypotheses: 1.Compared to other penguin species, Adelie penguins may have stronger swimming abilities; 2.External geographical factors, such as geological or climatic changes, have led to Adelie penguins inhabiting these three different islands.

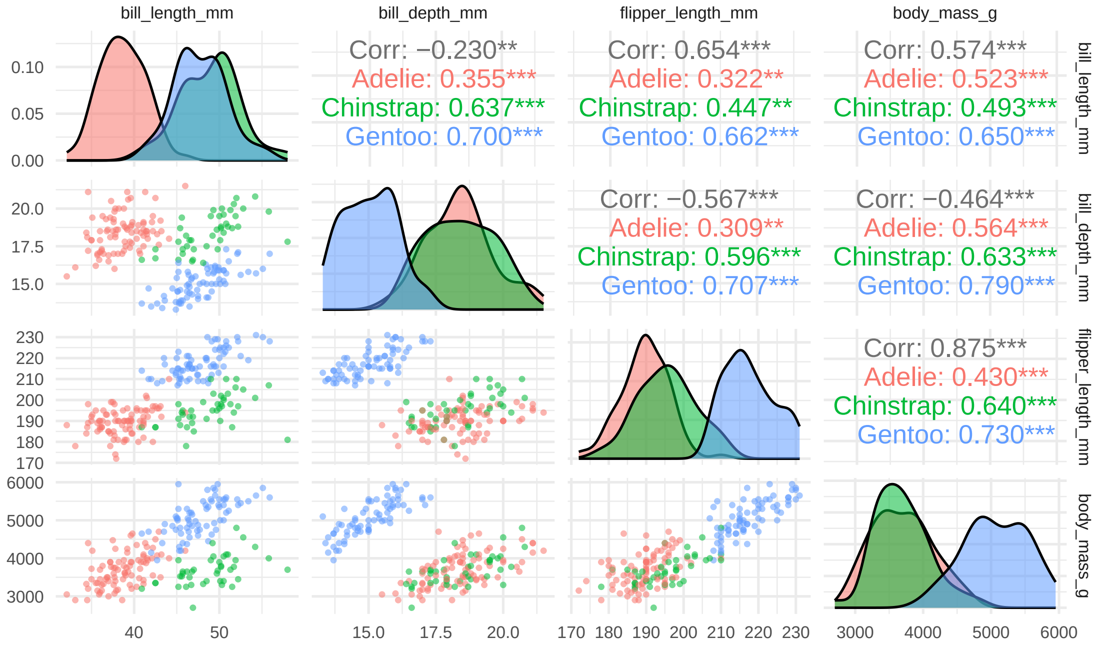
To confirm the first hypothesis, the average flipper lengths of the three penguin species should be considered, as species with longer flippers may have stronger swimming abilities. Additionally, the ratio of flipper length to body mass should also be examined, as there are weight differences among the three species, and body mass is positively correlated with flipper length (which will be discussed further below). Therefore, in this context, the ratio of flipper length to body mass may be more persuasive.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## # A tibble: 3 x 4
##   species   average_body_mass average_flipper_length flipper_length_to_body_ma~1
##   <fct>     <chr>             <chr>                  <chr>
## 1 Adelie    3695.06           189.67                 0.0520
## 2 Chinstrap 3686.90           195.52                 0.0535
## 3 Gentoo    5071.53           217.60                 0.0432
## # i abbreviated name: 1: flipper_length_to_body_mass_ratio
```

From the data above, we can see that the average flipper length and the average ratio of flipper length to body mass of Adelie penguins are not higher than those of the other two species. Therefore, up to this point, we cannot conclude that the distribution of Adelie penguins across these three islands is due to their stronger swimming abilities. As for the second hypothesis, more geological or meteorological evidence is needed.

1.3 Correlations Among Key Morphometric Traits of Penguins

To investigate the relationships among the four variables: bill length, bill depth, flipper length, and body mass, the following scatterplot matrix was generated based on three different species.



From the graph, it can be seen that for these three species of penguins, the four variables are all positively correlated with each other. However, the strength of these positive correlations varies among the different species. The bill depth of Adelie penguins is moderately positively correlated with body mass, the flipper length of Chinstrap penguins is strongly positively correlated with body mass, the bill length and bill depth of Gentoo penguins are strongly positively correlated, and the flipper length and body mass of Gentoo penguins are also strongly positively correlated.

The variable distribution plot on the diagonal also shows that the bill length of Adelie penguins is shorter than that of the other two penguin species, the bill depth of Gentoo penguins is shorter than that of the other two species, while their flipper length is longer than that of the other two species, and their body mass is larger than that of the other two penguins. Combined with the scatter plot, it can be concluded that there are significant species differences in the physical characteristics of these three penguin species. Finally, when considering all three penguin species as a whole, there is a strong positive correlation between flipper length and body mass for all penguins, with a correlation coefficient of 0.875.

2 Fitting a Normal Distribution to the Bill Length Data of Adelie Penguins

Based on the scatterplot matrix in Section 1.3, we can observe that the bill length data for Adelie penguins appears to follow a normal distribution. Next, we will fit a normal distribution to this dataset.

2.1 Calculate the maximum likelihood estimate of the parameters

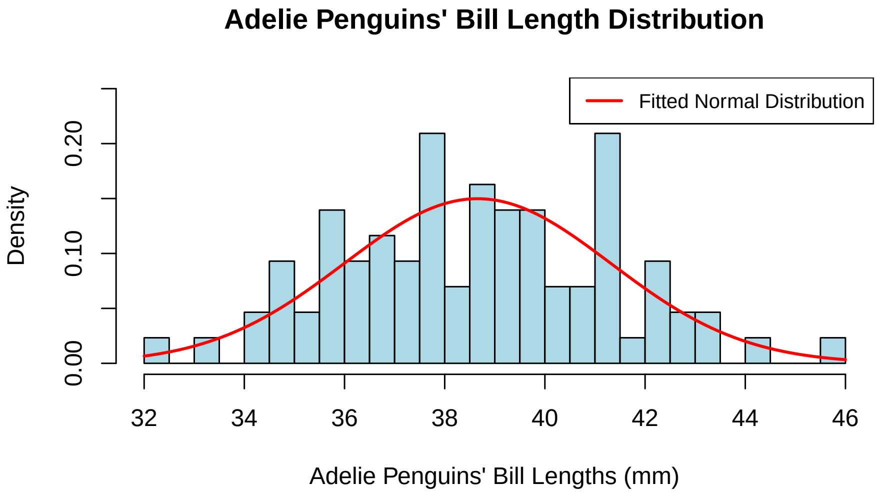
The parameters calculated using R for the maximum likelihood estimate are as follows:

```
## Mean bill length for Adelie penguins: 38.65581
```

```
## Standard deviation of bill length for Adelie penguins: 2.662136
```

2.2 Plot the histogram and overlay the normal distribution curve

Generate a histogram of the bill length data for Adelie penguins using R, and overlay a normal distribution curve on the histogram based on the data from Section 2.1.



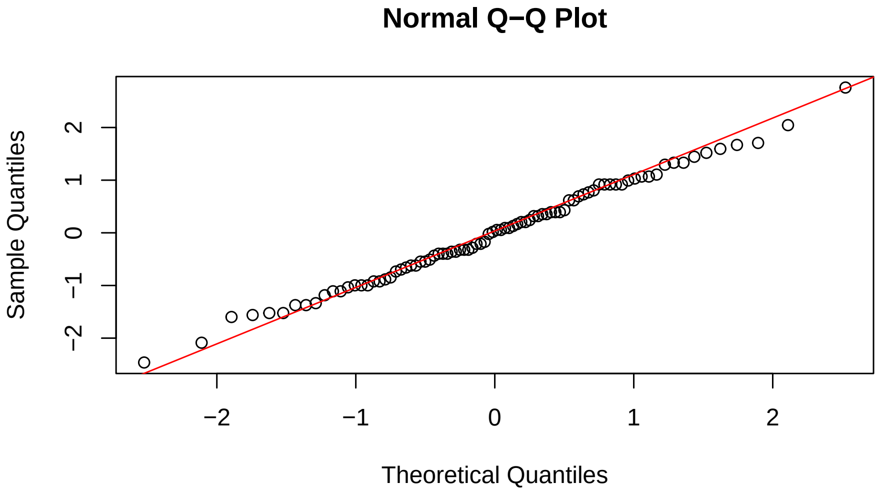
Visually, the fitted normal distribution curve aligns well with the actual data. To further assess the accuracy of this distribution, we conduct a Shapiro-Wilk test on the bill length data of Adelie penguins and generate a Q-Q plot related to this normal distribution.

2.3 Accuracy test of the distribution

The p-value from the Shapiro-Wilk test and the generated Q-Q plot are displayed below using R:

```
## Shapiro-Wilk test W value: 0.994
```

```
## Shapiro-Wilk test p-value: 0.974
```



In this Q-Q plot, the data points in the upper right corner are slightly below the red reference line, indicating that the actual data in the right tail of the distribution is slightly less than what is expected for a normal distribution. Similarly, there are some data points above the reference line in the lower left corner, suggesting a slight deviation in the lower values as well. Nevertheless, in the Shapiro-Wilk test, the p-value is 0.827, which indicates that the data does not significantly deviate from normality. Although there are minor deviations in the tails of the Q-Q plot, statistically, these deviations are not sufficient to conclude that the data is not normally distributed. In summary, we can conclude that the bill length of Adelie penguins follows a normal distribution.

2.4 Calculating Probability from the Normal Distribution of Adelie Penguins' Bill Length

The bill length data of Adelie penguins follows a normal distribution, which allows us to calculate the probability that the bill length of any given Adelie penguin is less than or equal to a specific value using the normal distribution model for Adelie penguin bill length. In section 2.1, it was determined that the mean and standard deviation of the Adelie penguin bill length are 38.65581 mm and 2.662136 mm, respectively. Using these two values, print out two example probabilities in R as follows:

```
## The probability that the bill length is less than or equal to 40 mm is: 0.693
```

```
## The probability that the bill length is less than or equal to 36 mm is: 0.159
```

3 Estimating Penguin Sex Using Measurement Data

In order to estimate the sex of penguins from measurement data and avoid direct harm to the animals, we need to determine whether there are significant differences in body measurements between male and female penguins. To achieve this, we can perform T-tests on all of the penguins' measurement data.

3.1 Performing T-Test on the Penguin Dataset

As noted in Section 1.3, there are differences in physical characteristics among the three penguin species. To avoid any interaction effects between the data, we should perform T-tests separately for each species to compare sexes, examining whether bill length, bill depth, flipper length, and body mass show significant statistical differences between males and females. The following are the results of the relevant T-tests performed on the penguin dataset using R:

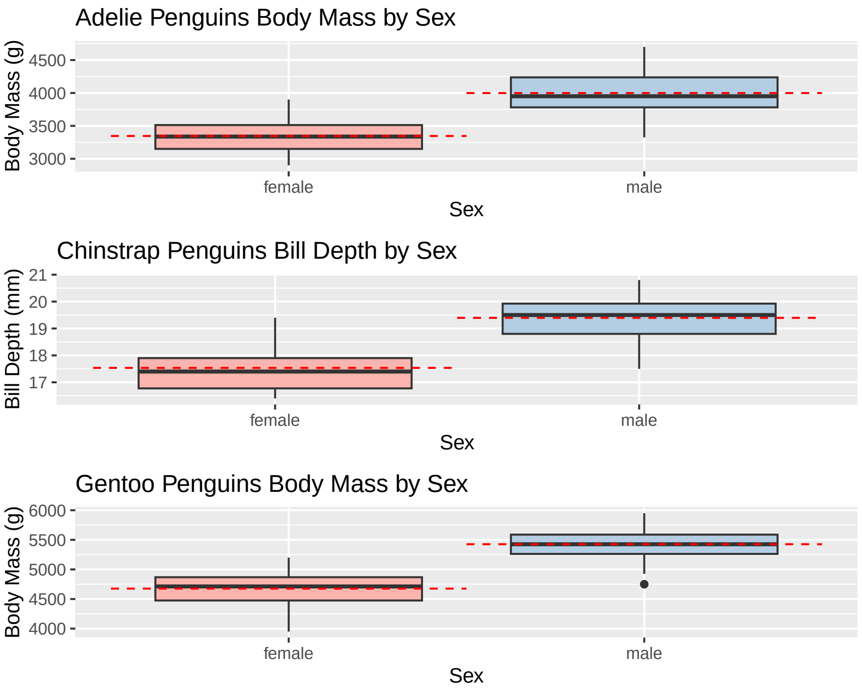
```
##       Species          Variable  t_value    p_value  conf_low conf_high
## t      Adelie    bill_length_mm  -6.2916 1.3577e-08   -3.9265   -2.0405
## t1     Adelie     bill_depth_mm  -5.6737 1.9752e-07   -1.7721   -0.8522
## t2     Adelie flipper_length_mm  -3.8142 2.6015e-04   -7.3936   -2.3260
## t3     Adelie       body_mass_g -10.2839 1.8661e-16 -778.2493 -525.9898
## t4  Chinstrap    bill_length_mm  -4.9753 2.1417e-05   -5.9254   -2.4828
## t5  Chinstrap     bill_depth_mm  -7.2556  8.981e-09   -2.3766   -1.3407
## t6  Chinstrap flipper_length_mm  -4.8525 2.0523e-05  -13.1912   -5.4270
## t7  Chinstrap       body_mass_g  -4.1810 2.1135e-04 -665.3838 -229.3890
## t8     Gentoo    bill_length_mm  -6.9500 1.5875e-09   -5.0215   -2.7819
## t9     Gentoo     bill_depth_mm -10.8597 1.2109e-16   -1.8451   -1.2725
## t10    Gentoo flipper_length_mm  -7.8620 8.4376e-11  -11.2063   -6.6605
## t11    Gentoo       body_mass_g -10.8739 1.4047e-16 -887.5170 -612.3282
```

Based on the above results, for Adelie penguins, body mass and bill depth are the most effective variables for distinguishing their sex, showing very high reliability. Bill length also has a good distinguishing effect, but the ability of flipper length to differentiate between sexes is relatively weaker. For Chinstrap penguins, bill depth and bill length are the most effective variables for distinguishing their sex, showing the highest reliability. Flipper length also shows some differences between sexes, while body mass exhibits relatively smaller differences. For Gentoo penguins, body mass is the most significant variable for distinguishing sex, followed by bill depth. Both variables demonstrate high reliability. In contrast, flipper length and bill length show relatively weaker differences between male and female Gentoo penguins.

Overall, the differences in body mass between male and female Adelie and Gentoo penguins are significant, while the differences in bill depth between sexes are notable across all species. Therefore, body mass and bill depth can serve as reliable indicators for distinguishing between male and female penguins, allowing for effective sex estimation while minimizing disturbance and harm to the animals.

3.2 Visualizing Key Variables for Penguin Sex Differentiation

Based on the T-test results, the variable that best distinguishes the sex of Adelie penguins is body mass, the best variable for distinguishing Chinstrap penguins is bill depth, and for Gentoo penguins, body mass is the most effective variable for sex differentiation. The body mass of Adelie penguins, the bill depth of Chinstrap penguins, and the body mass of Gentoo penguins will be visualized by sex to provide further reference for distinguishing the sex of penguins across different species.



4 Does Island Habitat Affect Penguin Physical Characteristics

To determine whether there are significant differences in the physical characteristics of penguins living on different islands, an ANOVA can be performed on penguins from different islands. First, perform an ANOVA on all penguins from the three islands, and then conduct an ANOVA specifically on the Adelie species across the three islands.

4.1 ANOVA on Penguins from Three Islands

Using R language, perform an ANOVA on the four variables: bill length, bill depth, flipper length, and body mass for penguins from three different islands: Biscoe, Dream, and Torgersen. The results are as follows:

```
## 
## --- Bill Length ANOVA ---

##              Df Sum Sq Mean Sq F value     Pr(>F)    
## island        2    875   437.3   15.98 0.00000037 ***
## Residuals   197   5390    27.4                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## 
## --- Bill Depth ANOVA ---

##              Df Sum Sq Mean Sq F value              Pr(>F)    
## island        2  319.9  159.93   68.39 <0.0000000000000002 ***
## Residuals   197  460.7    2.34                                
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## --- Flipper Length ANOVA ---

##              Df Sum Sq Mean Sq F value              Pr(>F)
## island        2  15027    7514   56.55 <0.0000000000000002 ***
## Residuals   197  26173     133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## --- Body Mass ANOVA ---

##              Df   Sum Sq  Mean Sq F value              Pr(>F)
## island        2 51069927 25534963   66.25 <0.0000000000000002 ***
## Residuals   197 75927195   385417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA results, the p-values for all four variables are very small, which seems to indicate that the different islands have a significant effect on the body characteristics of the penguins. So, can we conclude that the island where the penguins live affects their body traits? Next, an ANOVA will be conducted on the Adelie penguins from the three islands.

4.2 ANOVA on Adelie Penguins from Three Islands

From section 1.2, we know that Gentoo penguins are only found on Biscoe Island, Chinstrap penguins are only found on Dream Island, while Adelie penguins are distributed across all three islands. To determine whether the Adelie penguins from different islands differ in their body characteristics, an ANOVA is conducted on four body-related variables for Adelie penguins from the three islands. The results are as follows:

```
##
## --- Adelie Penguins Bill Length ANOVA ---

##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2    8.8   4.393   0.614  0.544
## Residuals   83  593.6   7.152

##
## --- Adelie Penguins Bill Depth ANOVA ---

##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2   4.99   2.494   1.609  0.206
## Residuals   83 128.62   1.550

##
## --- Adelie Penguins Flipper Length ANOVA ---

##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2    313  156.73   4.091 0.0202 *
## Residuals   83   3179   38.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## --- Adelie Penguins Body Mass ANOVA ---

##             Df   Sum Sq Mean Sq F value Pr(>F)
## island       2    31595   15797   0.079  0.924
## Residuals   83 16565680  199587
```

In the ANOVA results for Adelie penguins, the p-values for the four variables related to the penguins' body characteristics are all greater than 0.05, indicating that there are no significant differences in body characteristics among Adelie penguins from different islands. Therefore, it can be further inferred that the differences in body characteristics among penguins on the three islands do indeed exist, but these differences are more likely due to the inherent differences among the three species themselves, rather than being caused by environmental factors, such as the islands where they live.