MAS8403 - Statistical Foundations of Data Science

Rutuja Jagtap - 230118658

2023-10-20

1. Overview

The dataset comprises 200 rows and 8 columns, which represent various variables. Out of these 8 variables: Species, Island, and Sex are qualitative or categorical variables, and they fall under the nominal data category. They describe characteristics such as species, location, and gender. Year is a qualitative variable as well but is of the ordinal type, indicating data in chronological order. Bill length in mm, bill depth in mm, flipper length in mm, and body mass in grams are quantitative or continuous variables. They represent measurable quantities like physical dimensions and weight.

Below is a flowchart that illustrates different categories in data. I have extracted 200 rows of data from the original penguin data set, which originally consisted of 333 entries. This subset of data has been selected for the purpose of conducting data analysis and exploration. Penguins are categorized into two sexes, male and female, and further classified based on three different islands and three distinct species.

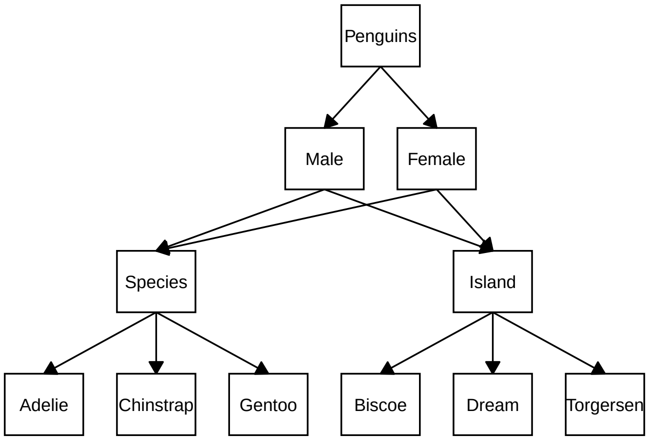


Figure 1: Data Representaion

2. Exploratory Data Analysis

Study of palmer penguin species

Since the penguin species are one of a significant aspect of this dataset, we will proceed to conduct an in-depth exploration and analysis of each species. Adelie penguins are present on all three islands, with an average population count of around 30. On the other hand, Chinstrap penguins are exclusively found on Dream Island, while Gentoo penguins inhabit Biscoe Island.

```
##
##             Biscoe Dream Torgersen
##   Adelie        32    30        33
##   Chinstrap      0    38         0
##   Gentoo        67     0         0
```

From 2007 to 2009, an equal number of male and female Chinstrap penguins were observed. Nonetheless, it's important to note that the assumption of an equal male-female ratio at any given point in time may not be accurate because this dataset lacks detailed information about the birth and death of penguins. Female penguins are prevalent on Biscoe and Torgersen islands. While they are approximatly equal to male penguins on Dreams island.

```
##
##             female male
##   Adelie        52   43
##   Chinstrap     19   19
##   Gentoo        34   33
```

```
##
##             female male
##   Biscoe        53   46
##   Dream         33   35
##   Torgersen     19   14
```

The Gentoo specie population show a gradual increase from 2007 to 2009 according to research data, on the other hand there was a significant surge in the number of Chinstrap penguins in the year 2008.

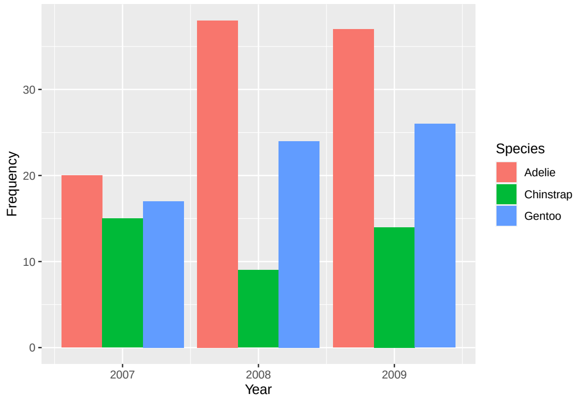


Figure 2: Frequency of species between 2007 and 2009

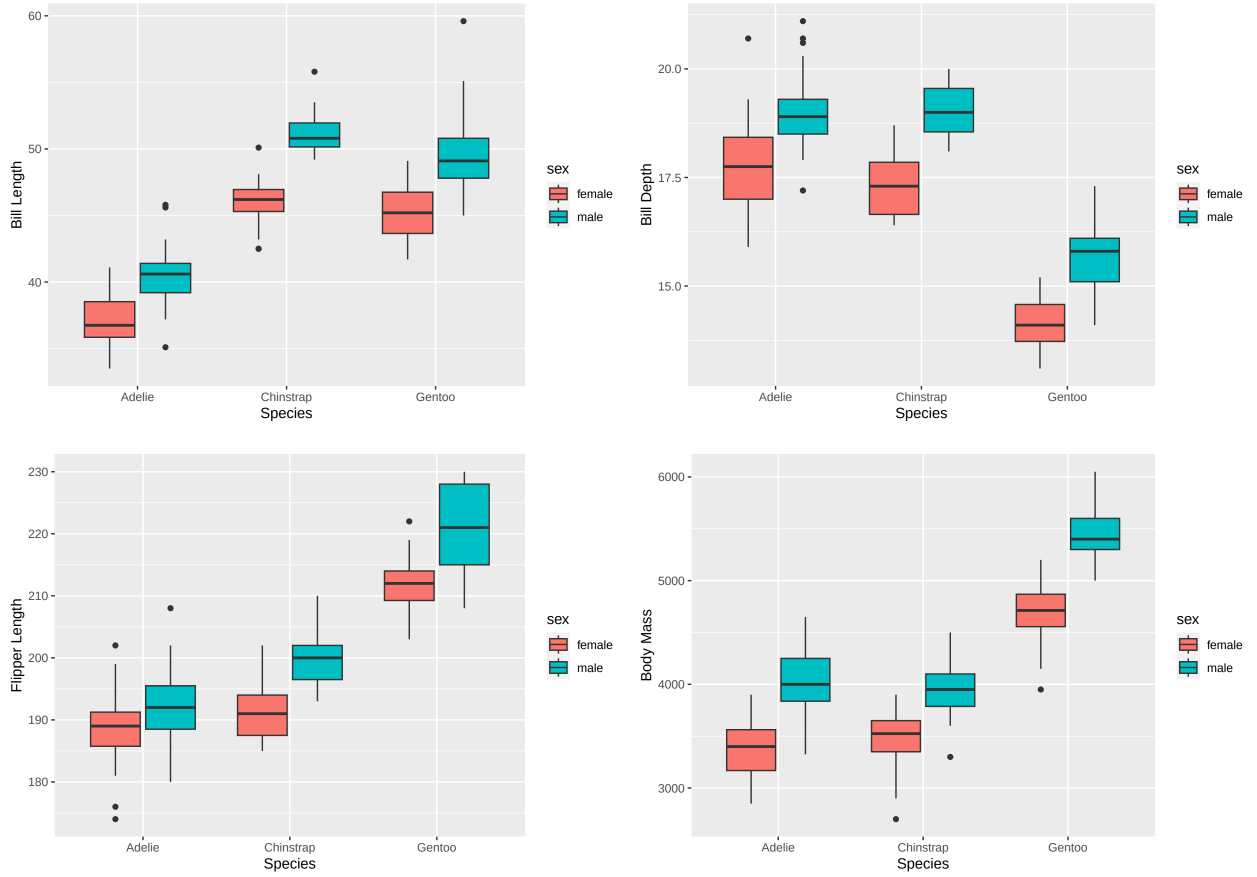
Let's compare all the penguin species based on their physical characteristics using one of the most standardized methods of data distribution—box plots.

Bill Length : There is a general trend among all three species: female penguins tend to have shorter bill lengths than male penguins. The Chinstrap male species exhibits longer bills, followed by the Gentoo male species. However, the highest bill length, around 60mm, is observed in one of the Gentoo penguins. The gap between Chinstrap male and female species is larger than other two species. Female penguins of the Adelie and Gentoo species exhibit a larger spread of values in bill length and fewer outliers compared to males or other species.

Bill Depth : Similar trend is observed in male and female penguins in terms of bill depth as in bill length that males exibit more bill depth than female penguins. Male Adelie and Chinstrap species show equal median plot. Bill depth of Gentoo species are very less when compared to other two species. There is minimum gap between box plot of Adelie male and female species. There are few outliers in this box-plot which are from Adelie species otherwise rest two species are withing strict range of values.

Flipper Length : The trend still continues, females have short flippers than male penguins. Highest flipper length, exceeding 230mm, is found in male Gentoo penguins. Both male and female Gentoo penguins have the longest flipper length compared to other species, and there is a significant separation between them. In contrast, there is no gap, and the distribution of values overlap for male and female Adelie penguins. Few outlier are again noticed in Adelie species only.

Body Mass : Gentoo species have the largest body mass, including some male penguins weighing over 6000 grams. There is large difference between female and male Gentoo penguin body mass. Median of male Adelie and Chinstrap penguins is almost equal. Adlie female penguins have least body mass of all.



According to below plot, Gentoo penguins stand out as significantly larger than the other two species, particularly when considering body mass and flipper length. These two characteristics seem to be closely related, as it can be seen an increase in flipper size might contribute to an overall growth in a penguin's body mass. Nonetheless, this correlation should not always hold, as various other physical characteristics of penguins (which might not be given in this dataset) can also play a role in influencing their body mass. For instance, according to scatter plot in Figure. 3, some Chinstrap penguins, despite having long flippers, exhibit lower body mass compared to Adelie species, which have shorter flippers but higher body mass. Hence, there are some noticeable outliers in boxplot illustration of the same data.

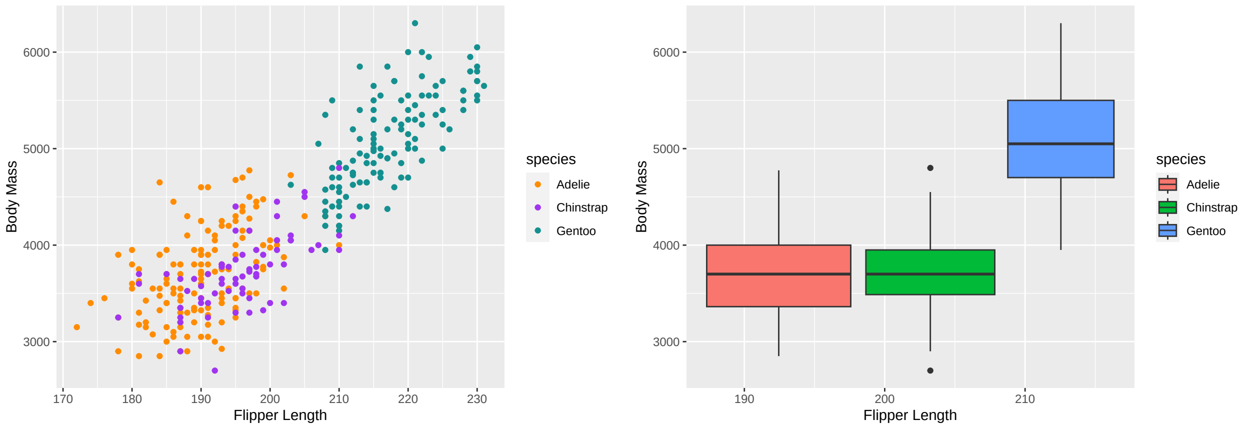


Figure 3: Scatter Plot (Left plot) and Box plot (Right plot)

3. Distribution

We will now explore the distribution of bill length within the palmerpenguins dataset. Since, the observation in bill length column are continuous and finite, we can consider using normal distribution for this particular column. To get graphical understanding of the values I have plotted a histogram, the distribution of bill lengths exhibits a departure from a uniform distribution, and it does not conform to a bell-shaped curve. This deviation is primarily due to a decrease in observations falling within the 40mm to 46mm range on the scale. The mean of the observations, at 43.407, is lower than the upper range of some bill lengths. Hence, we cannot assert with confidence that it adheres to a normal distribution. To present this more effectively, I have plotted the probability density on bill length observations in red.

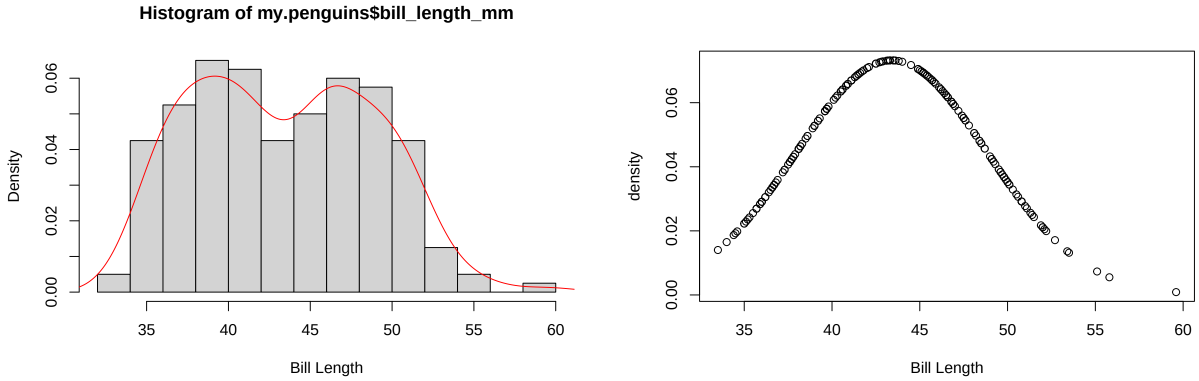


Figure 4: Normal Distribution

Figure on right side above, is an ideal example of a normal distribution, as illustrated using an actual mean and standard deviation. The graph exhibits symmetry around the mean and showcases a roughly uniform dispersion of data points within one standard deviation.

To determine maximum likelihood estimate (MLE) using normal distribution: MLE takes two arguments, mean and standard deviation, and x, which is a vector of data (in this case, bill lengths). The function calculates the log-likelihood of the data assuming a normal distribution with parameters mu (mean) and sigma (standard deviation). The optim function is used to find the maximum likelihood estimates of the parameters that maximize the log-likelihood. As a result we have found values for parameters,Estimated Mean = 43.41336 , Estimated Standard Deviation = 5.44205 and the maximum log-likelihood vale as 622.3825 of a normal distribution that best describe the distribution of bill lengths and it appears to have converged successfully.

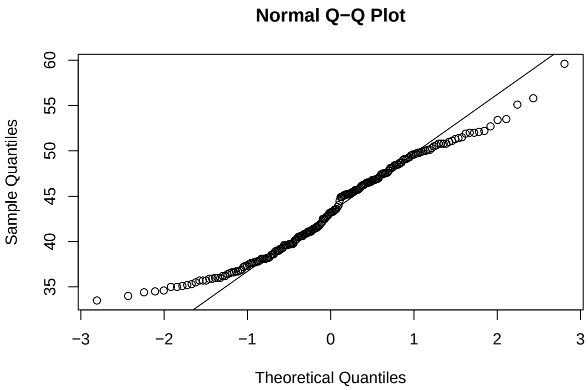
```
## $par
## [1] 2.587016e-01 1.542165e-37
##
## $value
## [1] -7.952497e+78
##
## $counts
## function gradient
##      501       NA
##
## $convergence
## [1] 1
##
## $message
## NULL
```

The MLE (Maximum Likelihood Estimation) output suggests that the estimated parameters for a normal distribution (i.e., the mean and standard deviation) are approximately 43.41336 and 5.44205, respectively. This indicates that a normal distribution may be a good fit for the data.

A normal distribution is often used when data is expected to be symmetrically distributed around a central value, with the majority of data points clustered around the mean. The standard deviation parameter measures the spread or variability of the data. In this case, the estimated mean and standard deviation correspond to the central tendency and spread of the data.

To assess the goodness of fit, it's important to conduct additional diagnostic tests, such as visual inspection of a histogram which is already illustrated in figure above to see if the data closely resembles a bell-shaped curve. Additionally, we can perform statistical tests to evaluate the normality assumption. If the data doesn't conform to a normal distribution, alternative distributions or models may be more appropriate.

To check normality of our penguin data lets perform a normality test: A Q-Q (quantile-quantile) plot is a graphical method to assess normality. It compares the quantiles of data against the quantiles of a theoretical normal distribution. If the points align to follow the straight line, it confirms normality. Following plot shows that straight line is not maintained throughout and hence normal distribution does not fit bill length column of data set.



In practice, the choice of distribution should be based on the data's characteristics and the specific context of the analysis. While the MLE estimation provides parameter estimates for a normal distribution, it doesn't guarantee that a normal distribution is the best fit. Further model validation and analysis are often necessary to make a definitive determination about the data's distribution.

4. Sex Determination

As per our data, out of 200 penguins, there are 105 female and 95 male penguins observed between 2007 and 2009.

```
## female   male
##    105     95
```

To identify patterns that can determine the sex of penguins without causing them distress, let's study all physical attributes of male and female penguins collectively. In the four box-plots below, male penguins exhibit dominant physical characteristics i.e male penguins posses higher value for bill length, bill depth, flipper length and body mass as compared to female penguins. The overlapping distribution in the box-plot of data makes it challenging to identify characteristics that can distinguish male from female penguins. However, there is certain trend followed which can be considered for making this distinction.

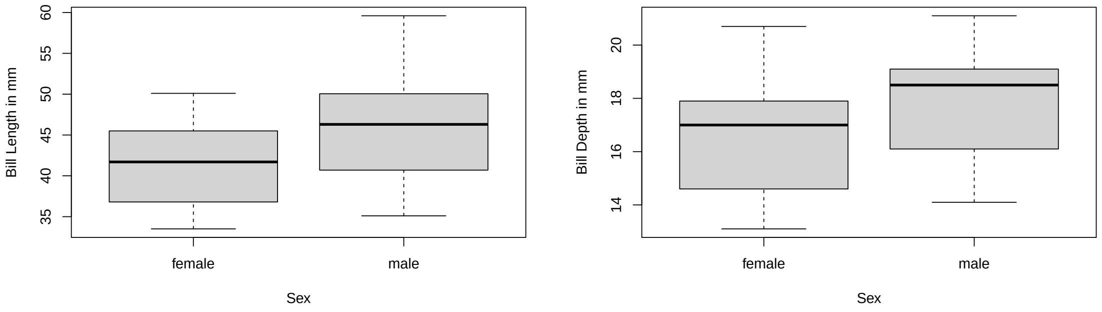


Figure 5: Bill Length Vs Sex (Left plot) and Bill Depth Vs Sex (Right plot)

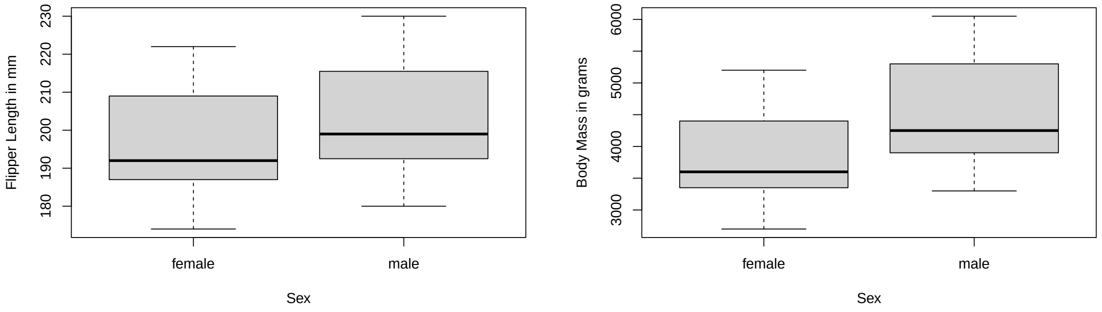


Figure 6: Flipper Length Vs Sex (Left plot) Body Mass Vs Island (Right plot)

To determine which features are the most informative for distinguishing male and female penguins, we can perform statistical tests such as t-test. This test and algorithms can help us identify which variables have the most significant impact on gender classification. Remember that the choice of features may also depend on the specific penguin species included in our dataset, as different species may exhibit different characteristics.

If the p-value is less than 0.05, we can reject the null hypothesis and conclude that the chosen feature is informative for distinguishing between male and female penguins.

After conducting t-tests on the variables bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g, the resulting p-values are as follows:

Table 1: T-Test Results for Gender Determination

feature	p_Value
Bill Length	0.00000
Bill Depth	0.00000
Flipper Length	0.00024
Body Mass	0.00000

P_value of all features are less than 0.05, so we reject null hypothesis. Therefore according to t-test, all these feature can contribute to provide sufficient information to determine gender of penguins. We cannot 100% rely on this test and one should perform few more test to learn the accuracy of p-value method.

5. Impact of island on physical feature of Penguins

To study the impact of islands on the physical characteristics of penguins, I applied three t-tests to determine p-values. I have chosen this test because we have more than two categories to compare and so this cannot be tested with normal t-test method. Subsequently, these p-values are compared to the threshold value of 0.05. If a p-value is greater than 0.05, we reject the alternative hypothesis (H1), and if a p-value is less than 0.05, we reject the null hypothesis (H0).

The tables below shows p-values for all four features across the three islands:

Table 2: T-Test Results for Bill Length by Island

islands	p_Value
Biscoe	0.7216366
Dream	0.0000000
Torgersen	0.0000000

Table 3: T-Test Results for Bill Depth by Island

islands	p_Value
Biscoe	0.0000000
Dream	0.0000000
Torgersen	0.7542984

Table 4: T-Test Results for Flipper Length by Island

islands	p_Value
Biscoe	0.0000000
Dream	0.0000000
Torgersen	0.0405556

Table 5: T-Test Results for Body Mass by Island

islands	p_Value
Biscoe	0.0000000
Dream	0.0000000
Torgersen	0.0993913

Bill Length - The p-value for Bill length on Dream and Torgersen islands are very low, indicating a strong statistical difference from Biscoe island.

Bill Depth - The p-values for Bill Depth on Dream and Torgersen islands are very low, suggesting significant differences from Biscoe island.

Flipper Length - The p-value for Flipper Length on Torgersen is again less than 0.05, while the latter two have p-values of 0. This suggests there is still some statistical significance, making it a viable differentiating parameter.

Body Mass- Torgersen exhibits a higher p-value for the body mass feature, indicating very little statistical difference from the other two islands.

From above test, we can say that physical features vary from island to island. However, the level of significant statistical variance for each feature is still not known. I have used ANOVA-Analysis of Variance test on all physical features to assess the significance of variance and obtained the following p-values:

```
## P-value of ANOVA Test for Bill Length      : 1.856e-07

## P-value of ANOVA test for Bill Depth       : 1.446e-19

## P-value of ANOVA test for Flipper Length   : 1.924e-16

## P-value of ANOVA test for Body Mass        : 1.786e-17
```

All physical characteristics are giving negative exponential p-values i.e they are showing p-values of nearly 0. This means we can reject the null hypothesis (H0) and conclude that penguins from different islands possess different physical characteristics. Numerous statistical tests like ANOVA make the assumption of equal variances across samples. The Bartlett test can be employed to validate this assumption.

```
## P-value of Bartlett Test for Bill Length      : 0.00046

## P-value of Bartlett Test for Bill Depth       : 0

## P-value of Bartlett Test for Flipper Length   : 0

## P-value of Bartlett Test for Body Mass        : 0
```

All the p-values are close to 0, indicating a considerable difference in the variance of physical features among penguins from different islands. Therefore, the island of origin has a notable impact on the physical characteristics of the penguins. The three t-tests produced three different p-values, making it challenging to assess the impact of the island on the physical features of penguins. Furthermore, this method is prone to errors when tested multiple times. In contrast, ANOVA provides p-values based on the mean of the data, and we also required a measure of variance to draw a conclusive decision. Hence, the Bartlett test was conducted to arrive at a final conclusion.