Data Analysis of Palmer Penguins using Statistical Methods

Vaishnavi Jeurkar
Student No.: 230595217 Newcastle University, Newcastle upon Tyne
Project Supervisor: Dr. Clement Lee

Introduction

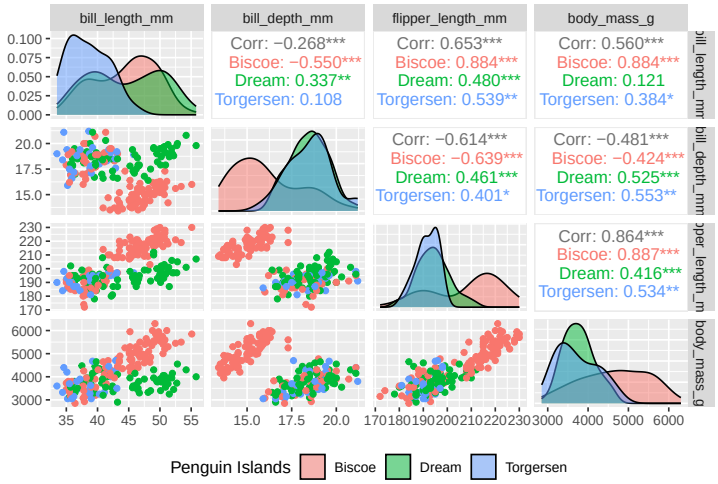
This research project aims to apply diverse statistical methods to analyze the Palmer Archipelago penguin dataset from Antarctica, particularly examining the relationship between penguin populations on different islands and their gender in conjunction with various dataset features. The methods utilized include exploratory data analysis, population proportion estimation, and hypothesis testing, providing essential tools for scientists to draw meaningful inferences and make predictive assessments.

Methodology

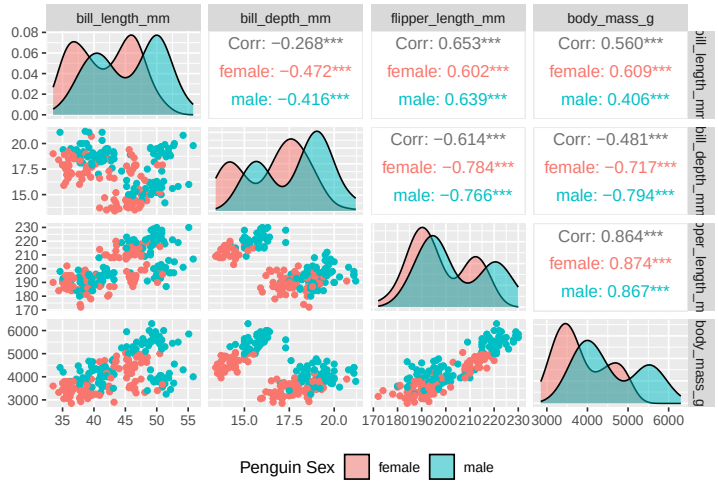
1. Exploratory Data Analysis

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :90   Biscoe   :104   Min.   :33.50   Min.   :13.40
##  Chinstrap:40   Dream    : 69   1st Qu.:39.00   1st Qu.:15.68
##  Gentoo   :70   Torgersen: 27   Median :43.50   Median :17.55
##                                 Mean   :43.72   Mean   :17.24
##                                 3rd Qu.:48.42   3rd Qu.:18.82
##                                 Max.   :55.80   Max.   :21.20
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172       Min.   :2850   female: 96   Min.   :2007
##  1st Qu.:190       1st Qu.:3594   male  :104   1st Qu.:2007
##  Median :197       Median :4050                Median :2008
##  Mean   :201       Mean   :4219                Mean   :2008
##  3rd Qu.:214       3rd Qu.:4750                3rd Qu.:2009
##  Max.   :230       Max.   :6300                Max.   :2009
```

The summary table tells us that Adelie penguins have highest sample size in the data set and Chinstrap has the lowest size.
The highest sample is taken from Biscoe island and lowest sample is from Torgersen island. There are 96 female and 104 male penguins in the dataset.



Using this plot we can visually explore the correlations and distributions of all the continuous variables for penguins living on different islands.
By analyzing the scatter plots we can see three distinctive groups.
We can observe that body mass of penguins from dream island seem to have a normally distributed data. This will be analyzed further.
Penguins living on Biscoe island have longer flipper length and high body mass but lowest bill depth.
Bill length and depth have gotten a weak to moderate correlation.
Body mass has a strong positive correlation with flipper length but weak correlation with bill depth.

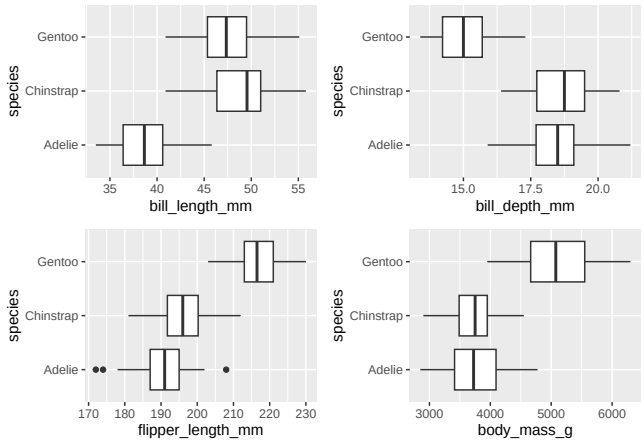


By examining the plots side by side on the same scale we can compare the distributions of measurement variables for both genders.
From the above plot we can observe that male penguins have longer flippers and higher body mass.
The data distribution of all continuous variables across genders is asymmetric and the data is multi-modal.

```
## , , island = Biscoe
##
##       species
## year   Adelie Chinstrap Gentoo
##   2007      6         0     20
```

```
##   2008     14         0     25
##   2009     14         0     25
##
## , , island = Dream
##
##       species
## year   Adelie Chinstrap Gentoo
##   2007     11        16      0
##   2008      8         9      0
##   2009     10        15      0
##
## , , island = Torgersen
##
##       species
## year   Adelie Chinstrap Gentoo
##   2007      8         0      0
##   2008     11         0      0
##   2009      8         0      0
```

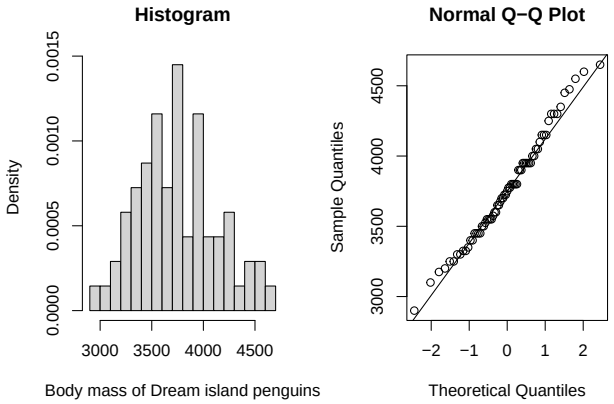
It is observed that Adelie species can be found on every island, Gentoo species are only found on Biscoe island and Chinstrap could be found on Dream island.
The number of penguins on Biscoe island have increased over the years.



The box plot of different species and their measurement variables is observed.
It is seen that Gentoo penguins have higher body mass and flipper length but lowest bill depth. As Gentoo penguins are only found on Biscoe island, we can possibly infer that differences in measurement variables over islands is because of certain species residing on that island.
Flipper length of Adelie penguins have a few outliers.
Adelie penguins have the shortest bill length.

2. Estimating probability/proportions for Penguin population.

The data sample is used to estimate the parameters of body mass for the population of penguins inhabiting Dream Island.



The histogram reveals a near-normal distribution of body mass for Dream Island's penguins. To delve deeper, a normal Q-Q plot was generated, displaying slight deviations at the tails, yet the majority of data points closely align with the expected line. To find the mean and standard deviation of the sample we use most likelihood estimation. **Equation for normal distribution**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Calculating log likelihood function

$$\log(f(x)) = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$\log(f(x)) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

Differentiate the log-likelihood function with respect to μ

$$\frac{\partial}{\partial\mu}\log(\mathcal{L}(\mu,\sigma)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Differentiate the log-likelihood function with respect to σ

$$\frac{\partial}{\partial\sigma}\log(\mathcal{L}(\mu,\sigma)) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Setting the result to 0 and solving for MLE of μ and σ Step 1: Set the derivative with respect to μ to zero:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

Step 2: Solve for μ:

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

Step 3: Rearrange the equation and solve for μ:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Step 4: Now, set the derivative with respect to σ to zero:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Step 5: Solve for σ:

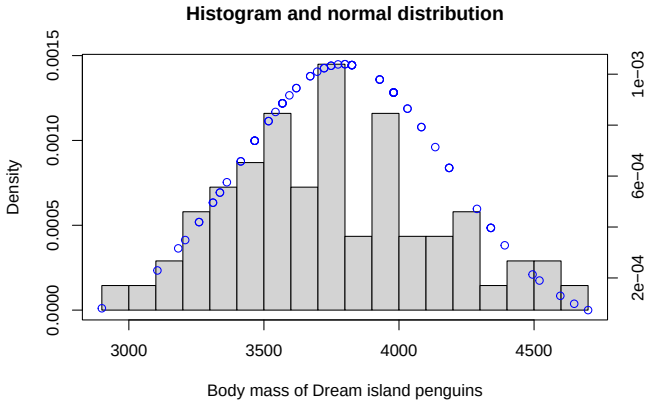
$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{\sigma}$$

Step 6: Rearrange the equation and solve for σ:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

further we can compute the parameters in R using dnorm function

```
## $par
## [1] 3767.1958  383.8225
```



The computed mean is 3767.1958 and standard deviation is 383.8225.
Following the construction of a normal distribution graph using the Maximum Likelihood Estimates for the mean and standard deviation over the histogram, it becomes evident that the body mass of penguins residing on Dream Island conforms to a normal distribution. This observation is further affirmed by the Q-Q plot.
When the model is accurately assumed, the Maximum Likelihood Estimator (MLE) stands out as the most efficient estimator, as it leverages the available data optimally to estimate the model parameters. This makes it a preferred choice for a wide range of applications, even in situations where assumptions of other models are not met. Moreover, in larger samples, the MLE tends to yield unbiased estimates, further enhancing its reliability and utility in statistical analysis.
A major disadvantage of MLE is that it tends to be biased for small samples of data.
To find the parameters of other measurement variables we need to first convert the data to a normal distribution by applying appropriate transformations.

Calculating the confidence interval of mean

```
## [1] "mean values of body mass for penguins from Dream island lie in the range of 3859.4 and 3675"
```

3. To identify the variables that are useful for determining sex of the penguin.

For this we use hypothesis testing. As we have observed from previous plots, the data distribution of continuous variables is multimodal and asymmetric. Therefore we cannot perform a parametric test(t-test) on the data.
Non-parametric tests do not assume distribution of data therefore we will perform multiple non parametric tests to confirm our hypothesis.

Tests to compare numerical features with sex of the penguin
ANOSIM test
ANOSIM test lets us compare a categorical variable with more than two groups to more than one numerical variable at a time.
Null Hypothesis (H0): There is no significant difference between numerical variable and different sexes.
Alternative Hypothesis (H1): There is a significant difference between numerical variable and different sexes.

```
## [1] "significance value is 0.001"
```

Since the significance value(p-value) is less than 0.05 we can reject the Null hypothesis. a small p-value suggests that there is a significant dissimilarity between groups, indicating that the groups are more different from each other than would be expected by chance.

Wilcox test
Wilcoxon test can be used when the predictor variable has 2 groups.
Null Hypothesis (H0): There is no significant difference between numerical variable and different sexes.
Alternative Hypothesis (H1): There is a significant difference between numerical variable and different sexes.

```
## [1] "The p-value of flipper length and sex is 0.000193056476503043"

## [1] "The p-value of body mass and sex is 3.38985723311065e-10"

## [1] "The p-value of bill length and sex is 8.58621217030712e-07"

## [1] "The p-value of bill depth and sex is 1.8654019132567e-07"

## [1] "The p-value of year and sex is 0.591030887822461"
```

From the hypothesis testing we can observe that flipper length, body mass, bill length and bill depth have a very small p-value. A very small p-value in a Wilcoxon test for the correlation between the variables and sex features suggests strong evidence to reject the null hypothesis. Therefore we have strong evidence to support the alternative hypothesis, suggesting that there is a significant difference in physical characteristics of penguins based on sex.
The p-value for correlation between year and sex is 0.059 which is greater than 0.05. Hence there is not enough evidence to reject the null hypothesis. We can also perform Kruskal Wallis H test for the comparison.

Test to compare categorical features i.e. island and species with sex
Chi-Squared test
Chi-squared test is used to find the correlation between two categorical variables.

```
## [1] "The p-value of species and sex is 0.973798946405567"

## [1] "The p-value of island and sex is 0.640062825523535"
```

The high p-value suggests that there is no significant association or dependency between species, island and sex variables. Therefore, it is appropriate to state that the features are independent of each other.

4. Impact of Island of on Penguin's Physical Characteristics

Based on our previous data visualization, it is evident that Adelie penguins are distributed across all three islands, whereas Chinstrap penguins are exclusively found on Dream Island, and Gentoo penguins are primarily located on Biscoe Island. These distinct species exhibit variations in their physical characteristics, as revealed by the box plots. Consequently, it is reasonable to infer that there exist significant differences in the physical attributes of penguins residing on different islands. This inference is substantiated through hypothesis testing. Given the non-normal distribution of these features, employing non-parametric tests is more appropriate to establish any existing correlations.

ANOSIM test
Null Hypothesis (H0): There is no significant difference between physical characteristics and different islands.
Alternative Hypothesis (H1): There is a significant difference between physical characteristics and different islands.

```
## [1] "significance value is 0.001"
```

A p-value of 0.001 in an ANOSIM test suggests that there is strong evidence to reject the null hypothesis. This indicates a significant dissimilarity between the groups being compared. We will conduct additional tests to examine the relationships between individual characteristics and the specific islands.

Kruskal-Wallis H test
The predictor variable of this test can have 2 or more groups, there can only be one outcome variable.
Null Hypothesis (H0): There is no significant difference between physical characteristics and different islands.
Alternative Hypothesis (H1): There is a significant difference between physical characteristics and different islands.

```
## [1] "The p-value of flipper length and island is 5.6788113532434e-11"
```

```
## [1] "The p-value of body mass and island is 3.40246607045495e-13"
```

```
## [1] "The p-value of bill length and island is 8.44731384897022e-07"
```

```
## [1] "The p-value of bill depth and island is 6.12974502262742e-16"
```

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the physical characteristics and island of the penguins.

Multiple Pairwise comparison between groups
Based on the results of the Kruskal-Wallis test, we have determined that there is a significant difference between groups. However, to identify which specific pairs of groups exhibit differences, we can utilize the function pairwise.wilcox.test(). This function allows for pairwise comparisons between different group levels while applying appropriate corrections for multiple testing.

```
## [1] "The p-value of flipper length and island is"
```

```
## $p.value
##                  Biscoe      Dream
## Dream     7.154024e-09         NA
## Torgersen 2.129865e-06 0.09236236
```

```
## [1] "The p-value of body mass and island is"
```

```
## $p.value
##                  Biscoe     Dream
## Dream     1.656688e-11        NA
## Torgersen 1.043559e-06 0.5354192
```

```
## [1] "The p-value of bill length and island is"
```

```
## $p.value
##                  Biscoe        Dream
## Dream     7.814110e-01           NA
## Torgersen 5.179738e-07 7.835477e-06
```

```
## [1] "The p-value of bill depth and island is"
```

```
## $p.value
##                  Biscoe     Dream
## Dream     7.052549e-14        NA
## Torgersen 6.877864e-08 0.9804799
```

The pairwise comparison shows that the flipper length, body mass and bill depth of Dream and Torgersen island penguins is not significantly different. i.e.($p>0.05$).
The bill length of biscoe and dream island penguins is not different. as the p-value is 0.78 we fail to reject the null hypothesis in this case.

Conclusion

Successfully applied various statistical methods to analyze the dataset. The exploratory data analysis provided valuable insights into the relationships between different variables. Maximum likelihood estimation allowed us to determine the mean and standard deviation of specific variables. Several non-parametric tests were conducted to investigate how different variables relate to the sex and island of the penguins.

The results indicate that certain physical characteristics, including bill length, bill depth, flipper length, and body mass, exhibit a higher correlation with the sex of the penguins. This suggests that these variables can be utilized by scientists for sex determination in penguins. However, our chi-squared test suggests that there is no statistically significant relationship between island/species and penguin sex. Additionally, the year of measurement does not appear to be related to the sex of the penguin. Hence island, species and year would not be useful for sex determination.

From our exploratory data analysis and the pairwise Wilcoxon test, we can infer that penguins from Dream and Torgersen islands possess distinguishing physical characteristics. This claim does not hold for differentiating penguins from other islands, as Dream and Torgerson share similarities in flipper length, body mass, and bill depth characteristics. Moreover, the bill length of penguins from Biscoe and Dream islands also shows similarity.

Biblography

1. Pawar, Aishwarya. (2022). Data Analysis Using Statistical Methods: Case Study of Categorizing the Species of Penguin.
2. Predictive analysis the study of different characteristics of Palmer penguins using R-programming Dr. Ramesh D Jadhav, Miss. Vaishali Bhujbal
3. Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises by Allison M. Horst, Alison Presmanes Hill, and Kristen B. Gorman

4. https://github.com/MaitriMZ/Penguins-Exploratory-Data-Analysis/blob/master/EDA_Penguins.ipynb
5. https://allysonf.medium.com/exploratory-data-analysis-on-palmer-archipelago-antarctica-penguin-data-41ff3e6efeda
6. https://www.w3schools.com/statistics/statistics_introduction.php
7. https://www.kaggle.com/code/florianspire/palmer-penguins-data-preprocessing-and-analysis
8. https://www.scribbr.com/statistics/statistical-tests/