Palmer Penguins Analysis

Harish Sainath

Introduction

In this report, we aim to analyse a subset of the famous Palmer Penguins data set. The subset data used for analysis has 200 entries and 8 different variables describing the penguins. These variables are species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex, year. Before diving deep into the data, let's try to get a better understanding about the penguins given in the data set by exploring all the variables, one at a time.

Species and Islands

- There are 3 different islands and these are Biscoe, Dream, Torgersen.
- We can observe from the below plot that, Biscoe is the most populated island followed by Dream and Torgersen.
- There are 3 species of penguins and they are Adelie, Chinstrap, Gentoo respectively.
- Population of each species in these island is given in the below plot.

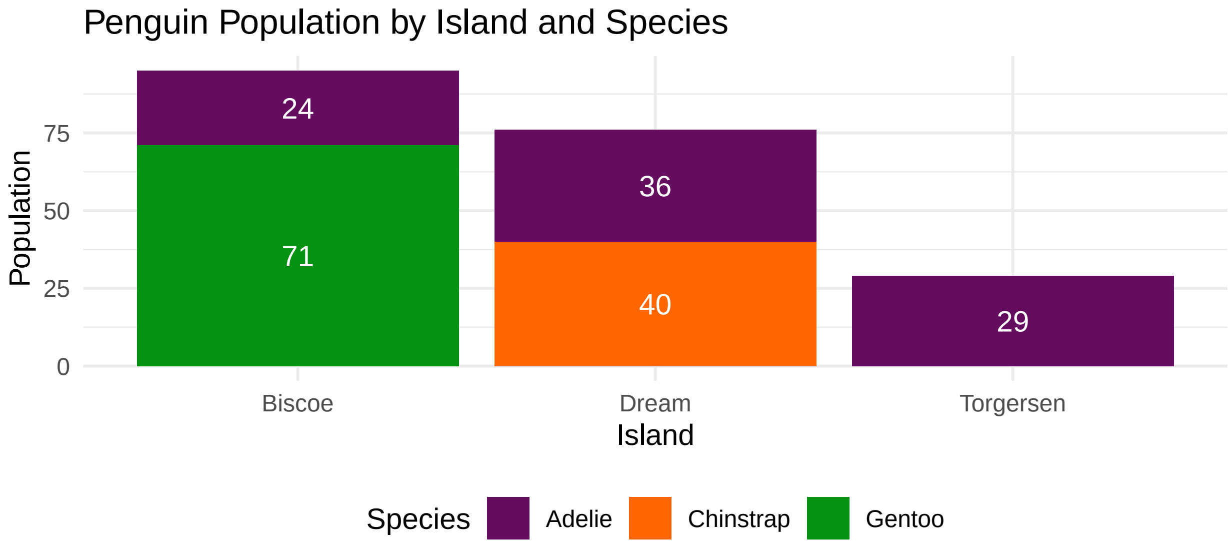

Figure 1: Penguin distribution by Species and Islands

Bill Length, Body Mass, Bill Depth and Flipper Length

- To get an idea about the features of penguins, we will consider median as a the measure of central tendency because it is not affected if the data is skewed.

- The penguins have a median bill length of 43.5 mm.
- From the below box plots we can observe that Chinstrap Penguins has the highest median Bill length. This is closely followed by Gentoo penguins.
- Apart from the median bill lengths, a few interesting points can also be observed. Adelie is the most common type of penguin that is present in all three islands. Chinstrap is present only in Dream island, and Gentoo is present only present in Biscoe island.
- All penguins have a median bill depth of 17.3 mm. Median bill depth for each species can be observed from the below box plots.
- Gentoo Penguins have the highest median flipper length and Adelie has the overall lowest median flipper length.
- In general, all the penguins have a median body mass of 3950 g.
- Gentoo penguins have the highest body mass in the dataset. Adelie penguins have the the lowest body mass of the group and Chinstrap penguins are comparable to Adelie penguins.

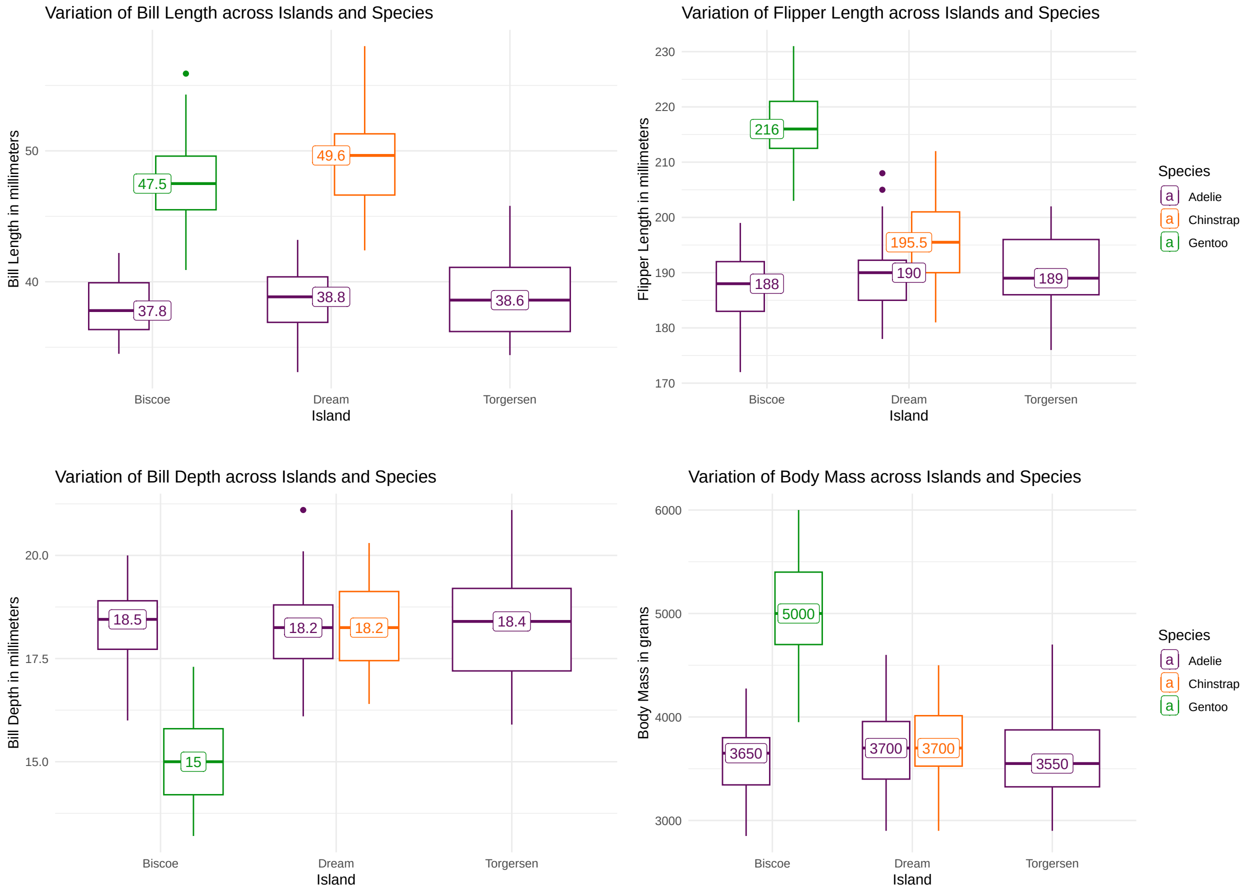

Figure 2: Distribution of Physical Characteristics by Islands and Species

Year and Sex

- The data is collected for 3 years and these are 2007, 2008, 2009 respectively. There are 98 female penguins and 102 male penguins in the dataset.

Distribution Fitting and Population Proportion Estimation

Let's take body mass of penguins and fit a distribution for proportion estimation. For getting initial impressions of how the data is distributed, let's plot a density plot. Also a lot of things in nature are observed to be falling under a normal distribution. So Let's use a Q-Q plot to check if body mass can be fitted using a Normal distribution.

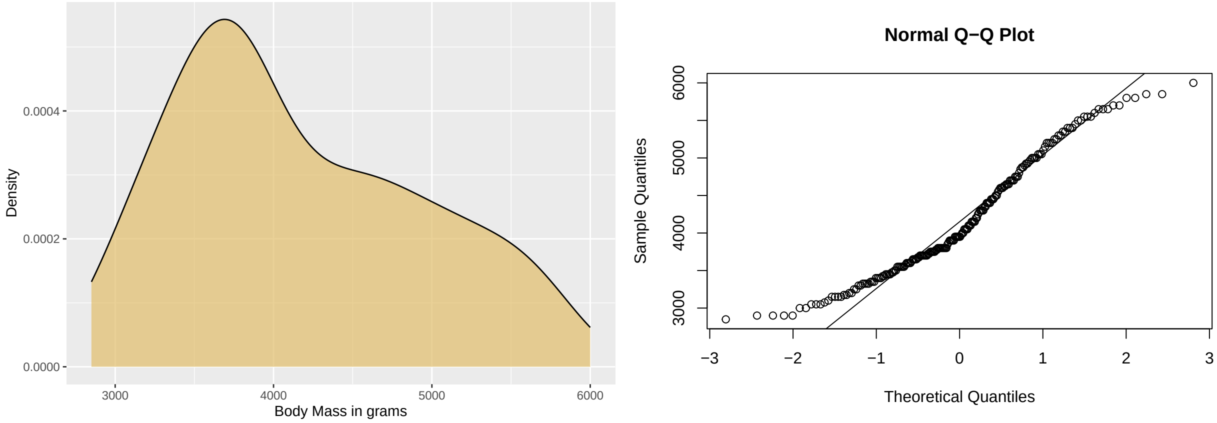

Figure 3: Data Distribution and Q-Q Plot

Two major things can be observed from the above plots.

- Q-Q plot shows that fitting a normal distribution might not be the best choice for good parameter estimations.
- Distribution of body mass seems to right skewed.

Now we have to look for some other distribution to fit the data. The body mass is a positive value and the data is right skewed, Log-Normal Distribution might be a good choice to fit the data.

Fitting a Distribution

Probability density function for a Log-Normal distribution is given by

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

- Mean is given by μ.
- Standard deviation is given by σ
- n is equal to 200.

Likelihood function:

$$L(\mu, \sigma | x_n) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma x_i} e^{-\frac{1}{2\sigma^2}(lnx_i-\mu)^2} \right]$$

Rearranging the above equation to make the calculations convenient

$$L(\mu, \sigma|x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \frac{1}{(\sigma^2)^{\frac{n}{2}}}} \prod_{i=1}^{n} \left[(\frac{1}{x_i}) e^{-\frac{1}{2\sigma^2} \Sigma (lnx_i - \mu)^2} \right]$$

taking ln on both sides

$$\ln L(\mu, \sigma^2|x_n) = \ln 1 - \frac{n}{2} \ln \pi + \ln 1 - \frac{n}{2} \ln \sigma^2 + \ln(x_1.x_2....x_n) - \frac{1}{2\sigma^2} \Sigma (lnx_i - \mu)^2 \ln e$$

$$\ln L(\mu, \sigma^2|x_n) = -\frac{n}{2} \ln \pi - \frac{n}{2} \ln \ \sigma^2 + \sum lnx_i - \frac{1}{2\sigma^2} \sum (lnx_i - \mu)^2 \qquad ...(1)$$

Where (1) is the log likelihood function

Now, by partially differentiating (1), we find equations for μ and σ.

- To estimate parameter μ (Mean) we differentiate the log-likelihood with respect to μ

$$\frac{\partial}{\partial \mu} lnL(\mu, \sigma^2|x_n) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (ln(x_i) - \mu)$$

$$\frac{\partial}{\partial \mu} lnL(\mu, \sigma^2) = 0$$

Solving this equation we get,

Mean is given by, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln(x_i)$

- To estimate parameter for σ^2 (Variance) we differentiate the log-likelihood with respect to σ^2

$$\frac{\partial}{\partial \sigma^2} lnL(\mu, \sigma^2|x_n) = -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{n} (ln(x_i) - \mu)^2$$

by solving this equation, we get

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

Standard Deviation is given by, $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\ln(x_i) - \hat{\mu})^2}$

- MLE for μ (mean of log body mass): 8.3177162
- MLE for σ (standard deviation of log body mass): 0.1857341

Observations

- The fitted distribution (Figure 4) approximately matches the actual data and this Log-Normal distribution represents skewness.
- While Log-Normal represents skewness, it is still not equal to actual data. But this is far better than normal distribution where skewness is zero.
- A more sophisticated model would better represent the data that is presented.

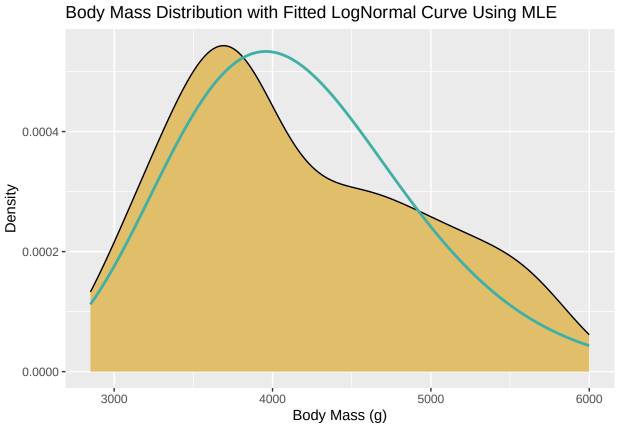

Figure 4: Body Mass Distribution with Fitted Log-Normal Curve

Determination of Sex

Determination of sex among penguins is often a difficult process where a lot of distress is caused to the penguins. Here, we look to use the available data to determine the sex of the penguins, thereby avoiding any invasive procedures.

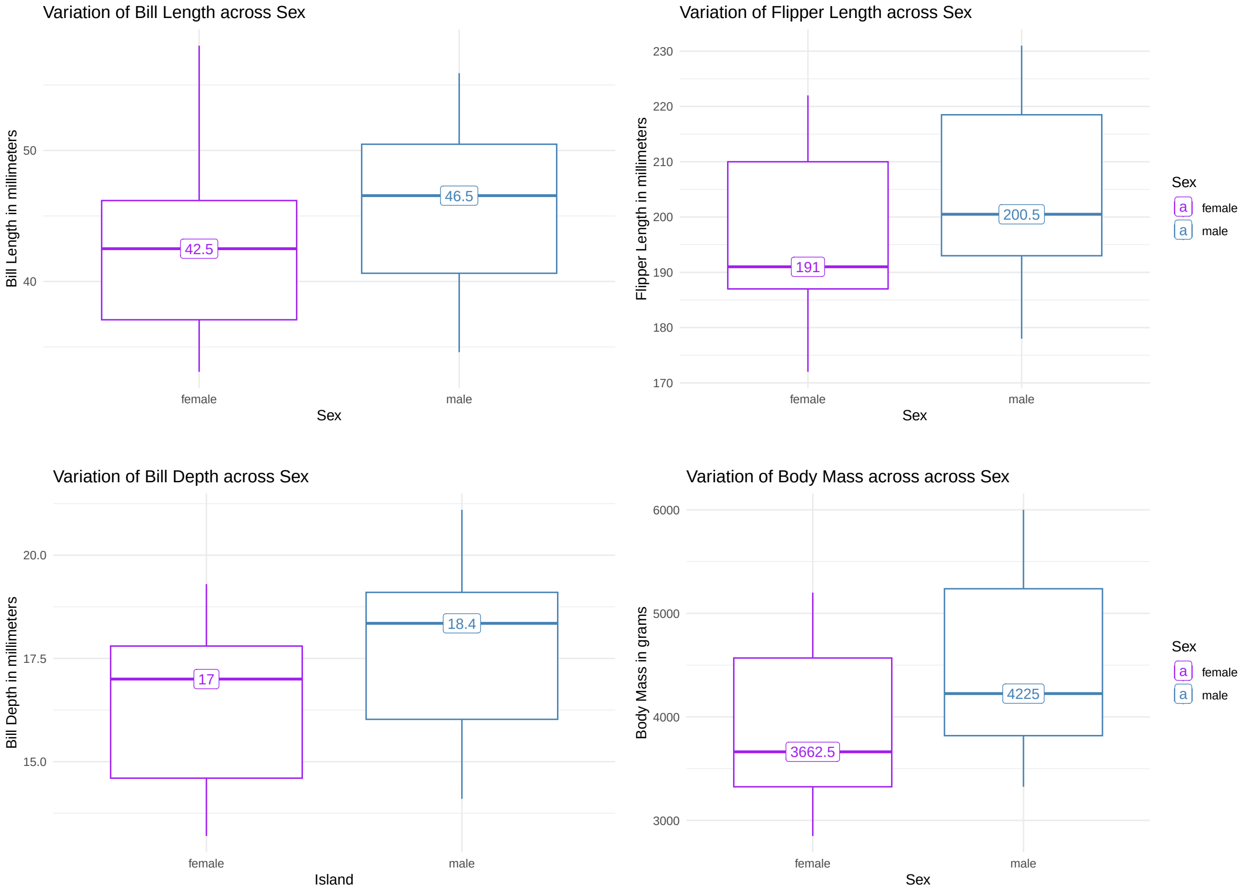

Figure 5: Distribution of Physical Characteristics by Sex

The above figure gives us a rough idea of how the data is distributed. But we need a mathematical way to determine if the differences are significant.

This task can be performed using the concept of hypothesis testing where we will assess physical characteristics and compare the means of each physical characteristics between two sexes.

For achieving this task, we need to frame our null and alternate hypothesis.

We can consider

- H_0: No difference between males and females for the considered variable
- H_1: Significant difference between males and females is present for the considered variable

When we use a t-test to check our hypothesis we will get something called as a P-value. If it is less than 0.05 then we can reject H_0 and say that there is significant difference between males and females for the given variable. Let's use the `t.test` function to test our hypothesis.

Note: We are assuming variances of two groups are equal

Hypothesis Testing

```
bill_length_t_test = t.test(bill_length_mm ~ sex, data = my.penguins,var.equal=TRUE)
bill_depth_t_test = t.test(bill_depth_mm ~ sex, data = my.penguins,var.equal=TRUE)
flipper_length_t_test = t.test(flipper_length_mm ~ sex, data = my.penguins,var.equal=TRUE)
body_mass_t_test = t.test(body_mass_g ~ sex, data = my.penguins,var.equal=TRUE)
```

- **Bill Length**

  P value : 0.0000030928386

  Mean of Female : 41.9418367

  Mean of Male : 45.5686275

  T value : -4.8020248

  Degree of Freedom: 198

  The p value is smaller than 0.05. So, there is evidence that the bill length for female penguins is smaller than the average male penguin.

- **Bill Depth**

  P value : 0.000000040840847

  Mean of Female : 16.4091837

  Mean of Male :17.8156863

  T value : -5.7102107

  Degree of Freedom:198

  The p value is smaller than 0.05. So, there is evidence that a female penguins' bill depth will be smaller than an average male penguin.

- **Flipper Length**

  P value : 0.00008493954

  Mean of Female : 196.3571429

  Mean of Male : 204.245098

  T value : -4.0131483

Degree of Freedom: 198

The p value is smaller than 0.05. So, there is evidence that the flipper length for female penguins is smaller than the average male penguin.

- **Body Mass**

  P value : 0.0000000038381803

  Mean of Female : 3845.9183673

  Mean of Male :4476.7156863

  T value : -6.1671935

  The p value is smaller than 0.05. So, there is evidence that female penguins will weigh much lesser than an average male penguin.

Conclusion

So the scientists can take the above calculated means of female/male penguins, create a threshold and then determine the sex based on the threshold set. In this way, no invasive procedure is needed. More accuracy can be achieved if we use more number of observations and a good machine learning model. One thing to note is that there will always be outliers and we should always consider that in our analysis.

Difference in Physical Characterstics of Penguins between Islands

We will now investigate the difference in physical characteristics of penguins between Adelie, Chinstrap, Gentoo islands.

From Figure 2 we can observe that physical characteristics of penguins across islands is different, but to be sure let's calculate if the means of the below variables differ significantly between islands or if they are similar enough that any observed difference are just due to random change.

- Bill Length
- Bill Depth
- Flipper Length
- Body Mass

We can use the regular t-test here, but there is a slight problem, we will need to perform 3 individual t-tests for each variable as we need to compare penguins between 3 islands. This will lead to rise in the risk of finding false positives or Type 1 errors.

This can be avoided if we use a method called Analysis of Variance or ANOVA. ANOVA enables us to compare means of two or more groups. We can then determine if the effect is due to random chance or if they mean something. Since we would be comparing each variable with island, we would be using *One Way ANOVA*. This is achieved by analyzing difference between group means and determining if the values are statistically significant.

One Way Analysis of Variances (ANOVA)

Let's consider

- H_0: Mean of the variable that we are considering is same for penguins across all islands.
- H_1: At least one island has significant difference in the variable that we are considering.

We will be using `aov` function to compute the P value and F value. P value is the probability of observing the result assuming Null Hypothesis is true. F value measures the ratio betweeen-group variance and within-group variance.

If p value is less than the statistical threshold value (0.05), we can reject the null hypothesis and say that there is significant difference between variables(e.g body mass) between penguins of each Islands.

- **Bill Length**:

```
##              Df Sum Sq Mean Sq F value       Pr(>F)    
## island        2    979   489.4   18.11 0.0000000603 ***
## Residuals   197   5324    27.0                         
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Bill length is different for penguins from different islands. Since P value is less than the threshold, we can reject the null hypothesis and go with alternate hypothesis.

- **Bill Depth**:

```
##              Df Sum Sq Mean Sq F value              Pr(>F)    
## island        2  294.1  147.07   71.52 <0.0000000000000002 ***
## Residuals   197  405.1    2.06                                
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Bill depth is different for penguins from different islands. Since P value is less than the threshold, we can reject the null hypothesis and go with alternate hypothesis.

- **Flipper Length**:

```
##              Df Sum Sq Mean Sq F value              Pr(>F)    
## island        2  15243    7622   57.53 <0.0000000000000002 ***
## Residuals   197  26098     132                                
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Flipper length is different for penguins from different islands. Since P value is less than the threshold, we can reject the null hypothesis and go with alternate hypothesis.

- **Body Mass**:

```
##              Df   Sum Sq  Mean Sq F value              Pr(>F)    
## island        2 48985030 24492515   64.82 <0.0000000000000002 ***
## Residuals   197 74432216   377829                                
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Body mass is different for penguins from different islands. Since P value is less than the threshold, we can reject the null hypothesis and go with alternate hypothesis.

Conclusion

From the One way ANOVA results, we can say that there are significant differences in physical characteristics (bill length, bill depth, flipper length, and body mass) among penguins from the 3 islands. So the island of origin has an impact on the physical characteristics of penguins.