# Data Analysis of Palmer Penguins

## Introduction

This paper is about the exploratory data analysis of Palmer Penguins dataset. Palmer Penguins dataset is collected by the Palmer Station located in the Palmer Archipelago in Antarctica (Palmer Station, 2021). So the dataset is consisting of 8 different variables measured on 333 penguins, such as species, island that penguin lives, bill length, bill depth, flipper length, body mass, sex and year that measurements were taken. In accordance with those data, this study focuses on the estimation of population proportions, sex determination and the association with island.

## Summary of Data

Before estimating the population, the scatter plot is plotted to find the appropriate measurement variables for fitting the distribution model. Figure 1 represents the pair of scatter plots with ellipse between three different species – Adelie, Chinstrap and Gentoo. As it illustrates, bill length versus bill depth graph features the excellent separation between three species, on the other hand, body mass versus bill depth graph has the most overlapping area between Adelie and Chinstrap. Therefore, the measurement variables are set as bill length and bill depth for population estimation.

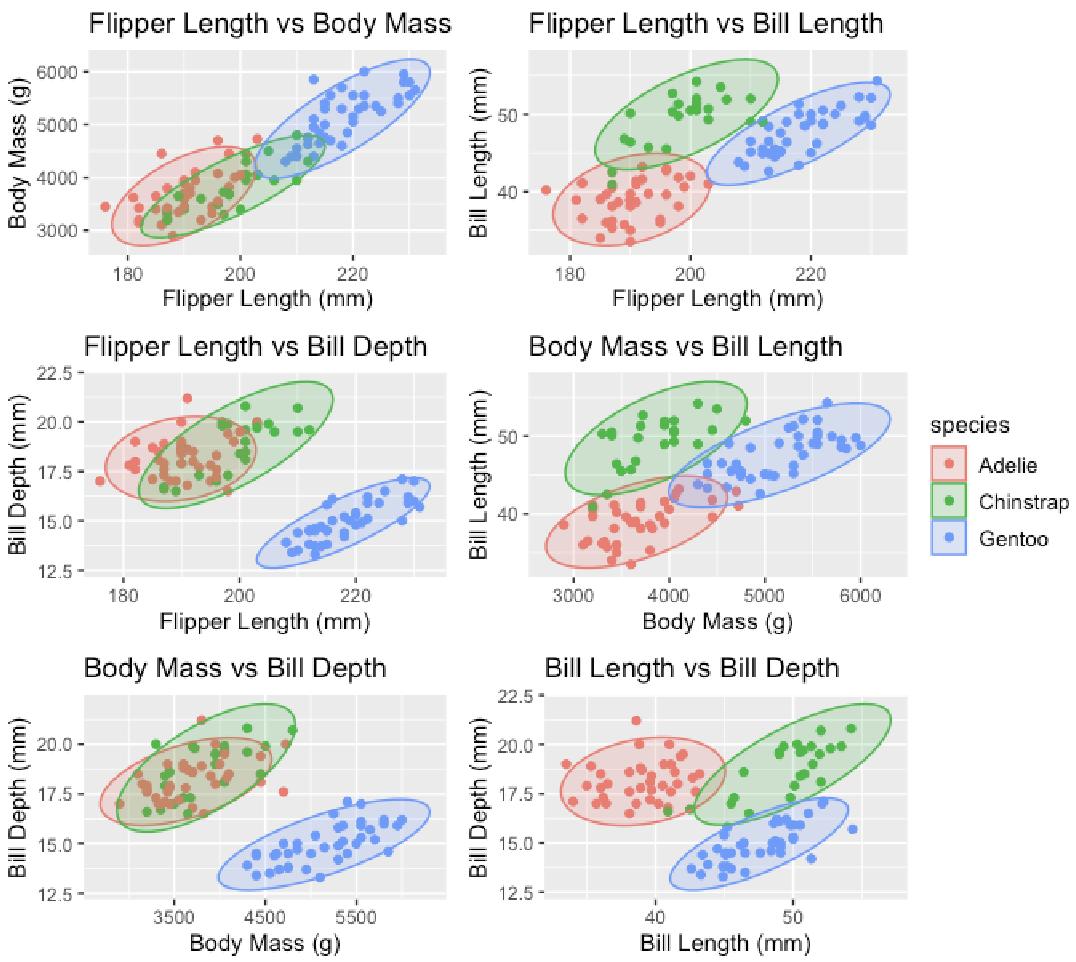

Figure 1: Ellipse plots with each pair of variables between species

Likewise, for distinguishing male and female, ellipse feature plot is plotted with each pair of variables as seen in Figure 2. Even though most of graphs appear to have high similarity between male and female, body mass versus bill depth graph turns out to be the most distinctive features compare to others. Thus, body mass and bill depth are selected as variables for determining the sex of penguin.

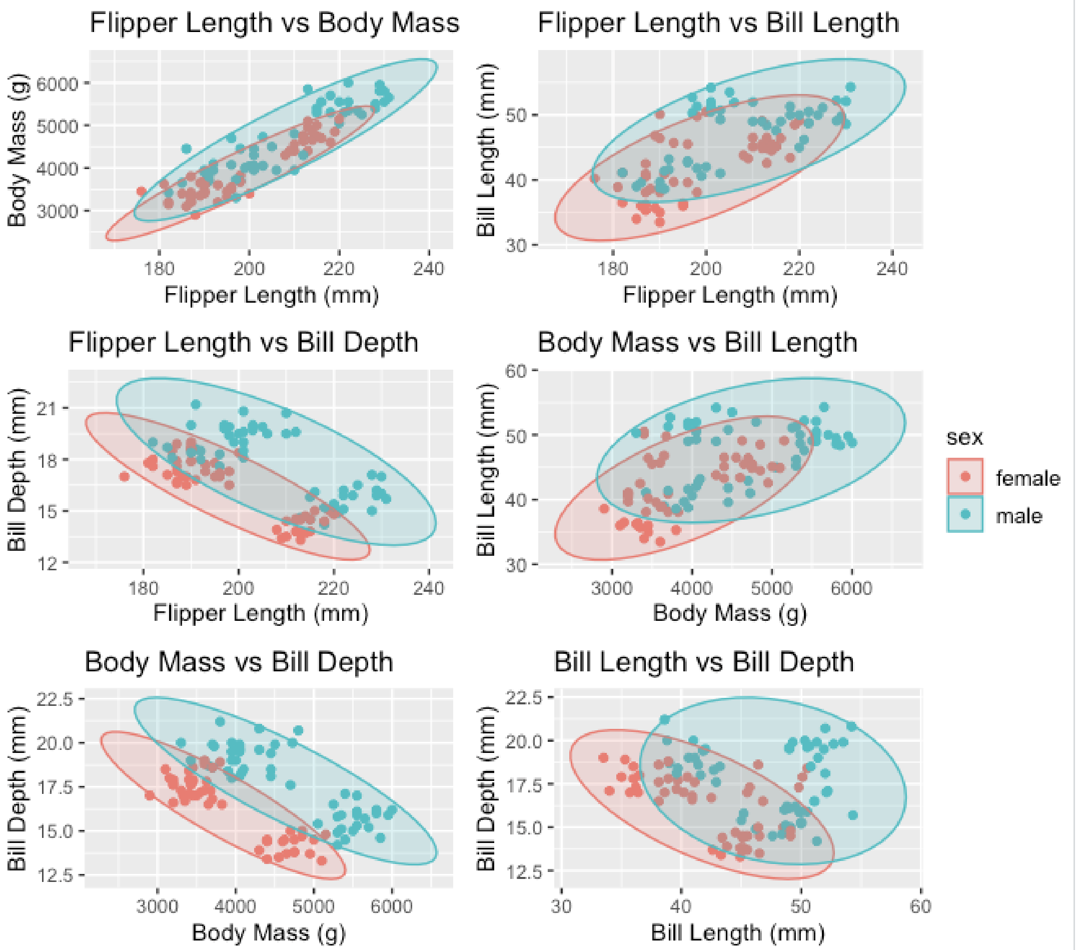

Figure 2: Ellipse plots with each pair of variables between male and female

## Population of Species

### Distribution Model

As bill length and bill depth are uncountable, not like discrete variable, both are continuous variables. Additionally, since it was unable to plot any of discrete distributions – such as Poisson and Binomial distribution – with those two variables, it is obvious that they are continuous variables.

Thus, to identify the proper distribution among continuous distributions, several graphical tests – histogram, Q-Q plot, CDF plot and P-P plot (Delignette-Muller and Dutang, 2020) – are implemented to fit distribution among normal distribution and exponential distribution. Figure 3 includes the four graphical tests for bill length and Figure 4 for bill depth. As Figure 3 and Figure 4 features, normal distribution follows the reference line the most, hence, it displays the best matches compared to exponential distribution.

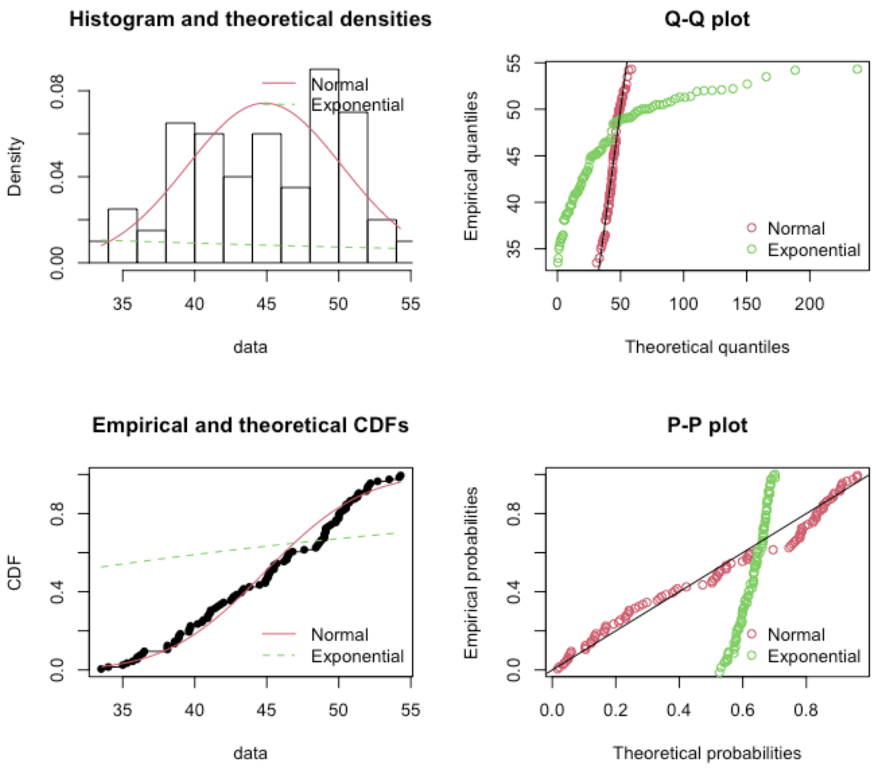

Figure 3: Fitting Distributions for Bill Length

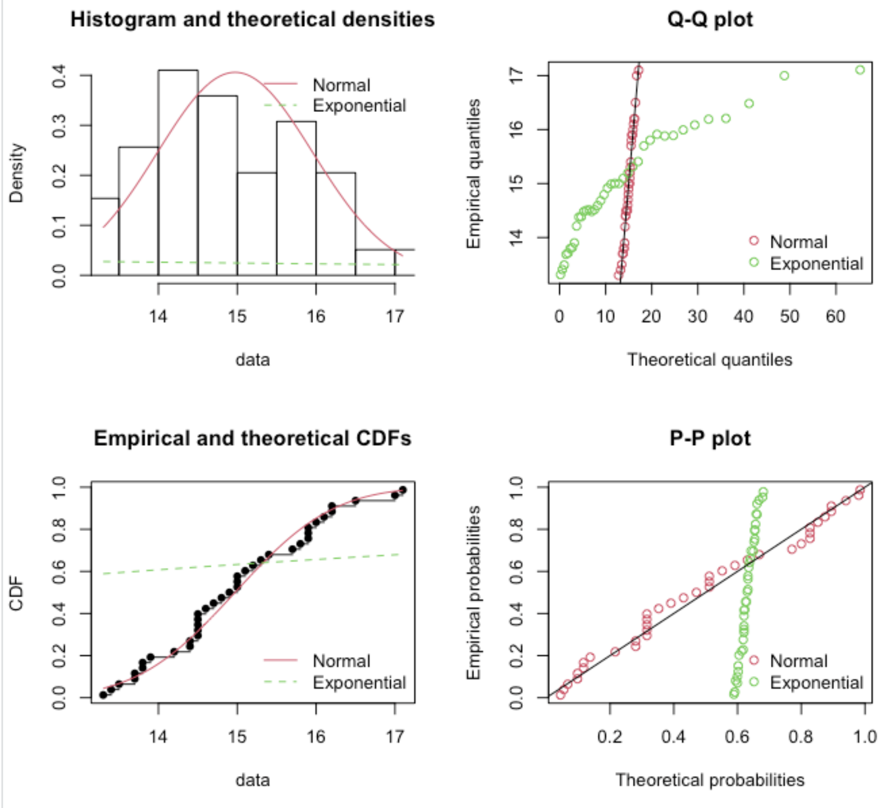

Figure 4: Fitting Distributions for Bill Depth

## Normal Distribution

In the previous paragraph, the normal distribution is revealed to be the most suitable model to find out the probability. Furthermore, the mean and standard deviation are required to plot a normal distribution so those two are settled as parameters as seen in Table 1.

TABLE I: Mean and Standard Deviation for Bill Length and Bill Depth

| Bill Length | Adelie | Chinstrap | Gentoo | Bill Depth | Adelie | Chinstrap | Gentoo |
|---|---|---|---|---|---|---|---|
| mean $\mu$ | 38.95 | 49.45 | 47.62 | mean $\mu$ | 18.20 | 18.78 | 14.97 |
| standard deviation $\sigma$ | 2.53 | 3.23 | 2.80 | standard deviation $\sigma$ | 1.01 | 1.30 | 0.98 |

Based on the bill length distribution in Figure 5, Gentoo and Chinstrap are partially overlap because their mean $\mu$ is only slightly different, 47.62 and 49.45 but Adelie can identify. On the contrast, for bill depth distribution, Adelie and Chinstrap are the similar cases due to their mean $\mu$ of 18.20 and 18.78 but not Gentoo.

As a consequence, since data is independently distributed and identically distributed, it is possible to get likelihood with $\mu$ and $\sigma$ with maximum likelihood estimates by following (1) (Eppes,2019).

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

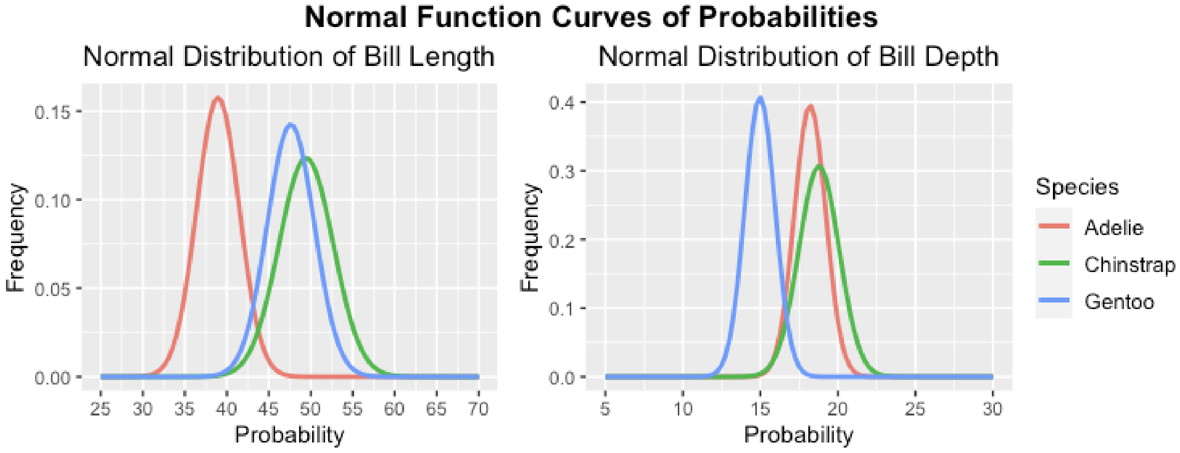

Figure 5: Normal Distribution of Bill Length and Bill Depth by Species

## Accuracy

To identify the accuracy of probability, confidence interval is calculated. Confidence interval can be figured out by (2) (Kelly, 2020).

$$\overline{X} \pm t_{\alpha/2,N-1} S_{\overline{X}} \tag{2}$$

where $S_{\overline{X}} = \frac{s}{\sqrt{N}}$ and $t_{\alpha/2,N-1}$ is for an area of $\alpha/2$ in each tail of a t-distribution with n-1 degrees of freedom.

In Table 2 and Table 3, the standard error of the mean is obtained by $S_{\overline{X}} = \frac{s}{\sqrt{N}}$, t-score $t_{\alpha/2,N-1}$ and margin of error $t_{\alpha/2,N-1} S_{\overline{X}}$. It means the confidence interval here is based on the t-distribution which is more conservative interval (ibid). As a result, according to the confidence intervals from Table 2 and Table 3, the range between three species are not overlapped so it is possible to classify the penguins with bill length and bill depth.

TABLE II: Confidence Interval of Bill Length

| | Adelie | Chinstrap | Gentoo |
|---|---|---|---|
| Mean | 38.95 | 49.45 | 47.62 |
| Standard error of mean | 0.42 | 0.67 | 0.45 |
| T-score | 2.03 | 2.07 | 2.02 |
| Margin error | 0.86 | 1.39 | 0.92 |
| Confidence Interval | 38.09 < x < 39.81 | 48.06 < x < 50.84 | 46.70 < x < 48.54 |

TABLE III: Confidence Interval of Bill Depth

| | Adelie | Chinstrap | Gentoo |
|---|---|---|---|
| Mean | 18.20 | 18.78 | 14.97 |
| Standard error of mean | 0.17 | 0.27 | 0.16 |
| T-score | 2.03 | 2.07 | 2.02 |
| Margin error | 0.34 | 0.56 | 0.32 |
| Confidence Interval | 17.86 < x < 18.54 | 18.22 < x < 19.34 | 14.65 < x < 15.29 |

In conclusion, according to Figure 3 and Figure 4, normal distribution is the model which fits the most since Q-Q plot, CDF and P-P plot follows the reference line well. However, as Figure 5 describes, it is hard to distinguish Chinstrap only with one variable. Nevertheless, this can be overcome by taking the confidence interval considering 95% ratio since the range is coinciding less than 1.00.

## Sex Determination

### Two-sample t-test

Two-sample t-test is implemented since it is required to distinguish between two groups, male and female penguins. Based on the ***Summary of Data*** paragraph, variables for the test is picked as body mass and bill depth.

Since the body mass has female group with mean $\mu_f$ and variance $\sigma_f^2$; and male group with mean $\mu_m$ and variance $\sigma_m^2$, the hypothesis will test if there is a significant difference between female group $\mu_f$ and male group $\mu_m$. Therefore, the hypothesis is:

$$H_0: \mu_f = \mu_m \,,$$
$$H_1: \mu_f \neq \mu_m \,.$$

Before conducting the two-sample t-test, it is acquired to verify if the equal variance assumption is valid. According to the result of Bartlett's test (Arsham and Lovric, 2011), p-value = 0.20 which is bigger than 0.1 so it is valid for t-test. Then, conducting t-test follows the form of (3) (JMP, 2022).

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \tag{3}$$

, where pooled standard deviation $s_p^2 = \frac{\left((n_1-1)s_1^2\right)+\left((n_2-1)s_2^2\right)}{n_1+n_2-2}$ and the average for each group $\bar{x}_1$ and $\bar{x}_2$. As a result, p-value = $9.29 \times 10^{-7}$ which is smaller than 0.001 so null is rejected and conclude that there is very strong evidence of a significant difference between the population mean body mass in the female and male. In other words, it is possible to distinguish penguin's sex by their body mass.

In the same context, when the same attempt is applied for penguin's bill depth, p-value results in 0.0002 which is smaller than 0.001. Thus, bill depth is also a proper variable to identify penguin's sex.

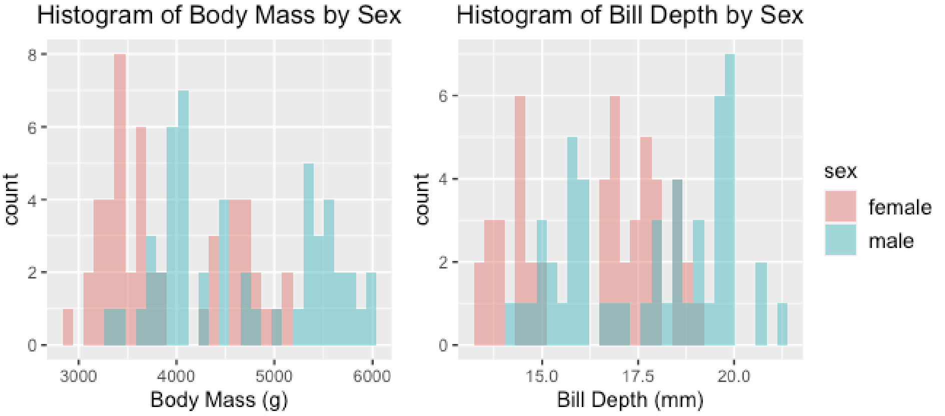

Figure 6: Histogram of Body Mass and Bill Depth by Sex

Similarly, the histograms of body mass and bill depth also tell that male penguins have larger figures then female in both area (Figure 6). Hence, it is verified that distinguishing between male and female is possible with body mass and bill depth.

## Island & Physical Characteristics

To identify the correlation of island and penguin's physical characteristic, the control variable is crucial for fair experiment. As Figure 7 describes, Chinstrap and Gentoo penguins only inhabit in one island so it is difficult to compare the physical features between three islands. On the contrary, Adelie penguin lives in every three island quite evenly so that it is proper to be an object for analysing.

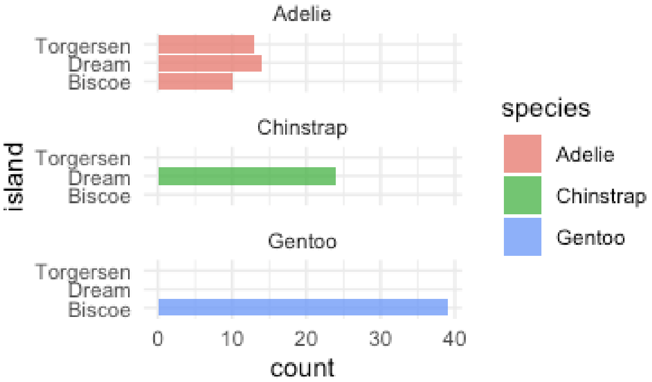

Figure 7: Bar chart of Species and Island

As Figure 8 demonstrates, Adelie penguin from Biscoe Island tends to have the highest figures in flipper length, body mass and bill length. However, although the Adelie penguin from Dream Island have the lowest quantities in almost every characteristic, only bill depth has the highest figure than others.

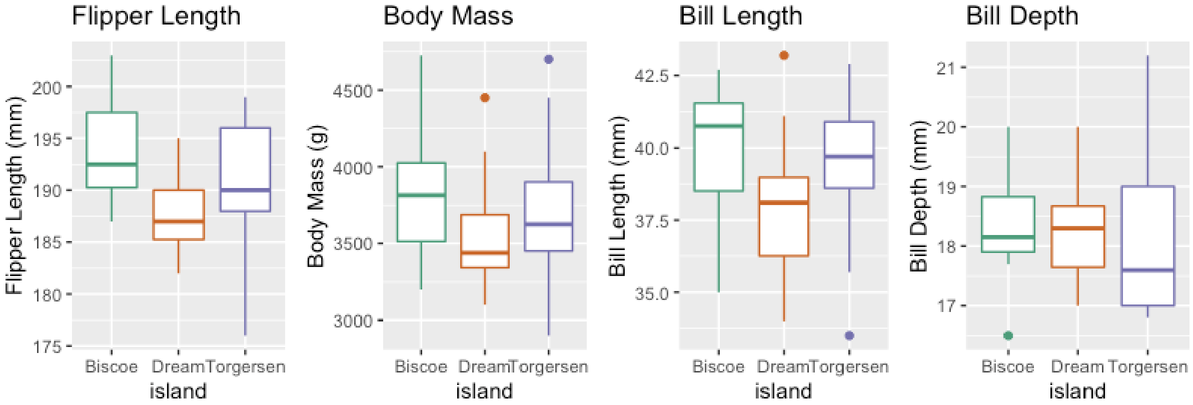

Figure 8: Adelie's physical characteristics by island

## Conclusion

To summarise, the paper mainly studied about three criteria, population of species, sex determination and island & physical characteristics.

The normal distribution of bill length and bill depth has conducted to identify the species, then 95% confidence interval has confirmed that it is possible to distinguish the species with those two variables. Furthermore, two-sample t-test hypothesis is applied to body mass and bill depth for identifying penguin's sex and the results is verified with histogram that male penguin tends to have larger figures for both of variables.

Finally, Adelie penguin's data is used to study about the correlation between the islands and penguin's physical characteristics. It turns out that penguins from Biscoe Island have the largest figures from almost every feature but not from bill depth. However, penguins from Dream Island usually have the smallest figures from every physical characteristic but have the largest bill depth.

# References

- Arsham, H., & Lovric, M. (2011). ‘*Bartlett's Test*’, International Encyclopedia of Statistical Science.
- Delignette-Muller, M. L. and Dutang, C. (2020) ‘*fitdistrplus: An R Package for Fitting Distributions*’, Journal of Statistical Software, 64 (4), pp. 1–34
- Eppes, M. (2019) *Maximum Likelihood Estimation Explained - Normal Distribution.* Available at: https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f [Accessed 18 Oct 2022]
- JMP (2022) *The Two-Sample t-Test.* Available at: https://www.jmp.com/en_gb/statistics-knowledge-portal/t-test/two-sample-t-test.html [Accessed 19 Oct 2022]
- Kelly, L. (2020) *Calculating Confidence Intervals in R.* Available at: https://bookdown.org/logan_kelly/r_practice/p09.html [Accessed 18 Oct 2022]
- Palmer Station (2021) *Palmer Station.* Available at: https://www.palmerstation.com/ [Accessed 15 Oct 2022]