Exploratory Analysis of Palmer Penguin Dataset

Aulia Amirullah Zulkarneidi
School of Computing
Newcastle University
Newcastle Upon Tyne, UK
c2606034@newcatstle.ac.uk

***Introduction* — In this analysis, we limit 100 out of 333 palmer penguins from the dataset. We analyze physical measurements and the gender of 3 different species (Gentoo, Chinstrap and Adelie) in the 3 year-time (2007-2009).**

Keywords—exploratory analysis, data science, statistics, statistical foundation of data science, palmer penguins.

I. Background and Facts

K. Gorman, the researcher at Palmer Long-Term Ecological Research (LTER) Station, claimed that the proportion in the dataset doesn't constitute the global population [1]. Parameters such as food supply, habitat availability and body mass were influenced by climate change over the 3-year observation period. Additionally, Wells said that it's cumbersome to tell penguin genders apart because they don't have external genitalia (it requires tools to examine internal genitalia) [2]. Moreover, location shouldn't have an impact to our species prediction model given that all three locations are relatively close to one another.

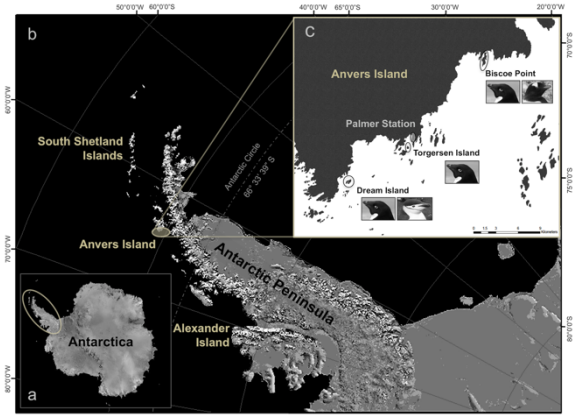

Fig. 1. 3 Antarctic islands where the observation was conducted [1]

II. Findings

There are some general findings we can talk about from Figure 2 and 3 such as: sample gathering was less equal between islands (Fig 3) and the population of palmer penguins are relatively stable over the 3-year time in 3 Antarctic islands even though there's an increase in Gentoo population in 2008 and a significant decline in Chinstrap population (Fig 2 and 3).

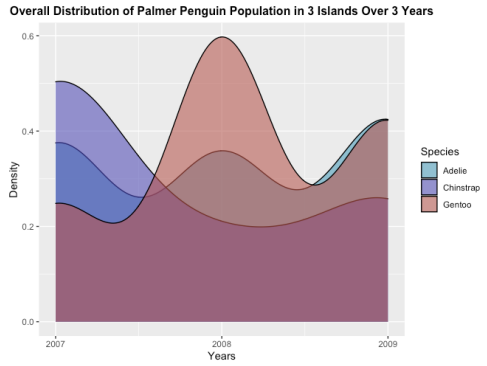

Fig. 2. Overall distribution of palmer penguin population over 3 years

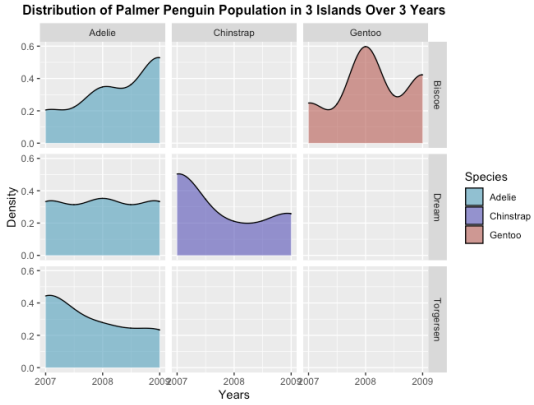

Fig. 3. Distribution of palmer penguin population in 3 islands over 3 years

In terms of general statistics of the dataset, we don't have an equal proportion between species and penguins which were from 3 different islands. Some data dominate others while we have almost a balanced proportion in sex gatherings of the data.

Species		Island		Sex		Year	
Adelie	46	Biscoe	52	Female	53	Min.	2007
Chinstrap	18	Dream	36	Male	47	1st Qu.	2007
Gentoo	36	Torgersen	12			Median	2008
						Mean	2008
						3rd Qu.	2009
						Max.	2009

Fig. 4. Overal summary of the dataset

Now, we look at the bigger picture of the physical measurement data. There's a significant difference between the minimum of body mass of the palmer penguins with the max data (around 3000 g difference). The flipper length variance also ranges for more than 50 mm from the max range. Nonetheless the bill depth and length might have some difference but not as much as the two body features described.

	Bill Length (mm)	Bill Depth (mm)	Flipper Length (mm)	Body Mass (g)
Min.	33.10	13.70	178.0	2850
1st Qu.	38.80	15.30	189.0	3544
Median	43.00	17.20	197.0	4050
Mean	43.29	17.10	200.3	4186
3rd Qu.	47.45	18.52	212.2	4731
Max.	55.90	21.50	230.0	6000

Fig. 5. Overall summary of body feature measurement of Palmer penguin dataset

III. Palmer Penguins Body Features in 3 Different Islands Over 3-year Time

We do some analysis over the 3-year observation period upon body mass, bill length, bill depth and flipper length of

palmer penguins of 3 Antarctic islands. In this part of analysis, we exclude sexing to find the pattern between 3 different islands about the body features which might be affected by climate variation in 3 Antarctic islands. Sexing will be discussed in more details in the next section.

In terms of body mass, Gentoo which predominantly lived or in which the penguin was exclusively observed in Biscoe had the biggest mean of body mass over the other penguins lived in Antarctic which were around 5000g over 3 years of observation.

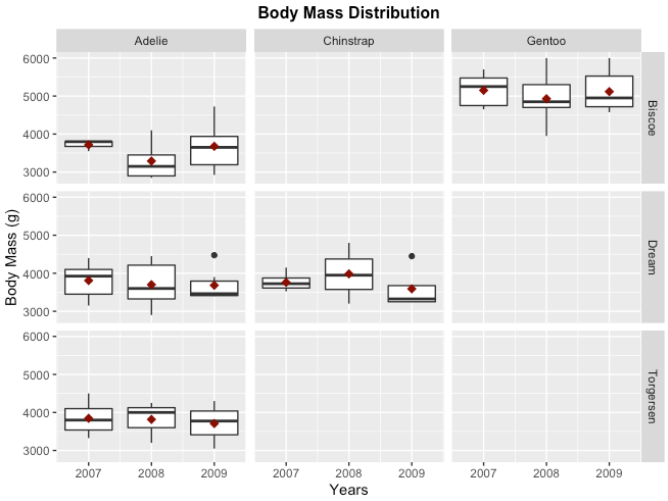

Fig. 6. Body mass distribution in 3 different islands

On the other hand, Chinstrap has a decline in body mass within 2 years in Dream Island and Adelie is relatively stable in their body mass in 3 of those islands over 3-year time (even though it experienced a body mass average decrease in Dream island since 2008. Adelie and Gentoo penguin which were from Biscoe experienced a decline in body mass average and interquartile in 2008 which might indicate there was food scarcity in the region.

On the contrary to body mass, Chinstrap which lived in Dream island has the highest average of bill length among other kinds of Palmer penguins (around 45 – 50 mm). Gentoo comes the second with little variation in the average and quartile of the length of their bills.

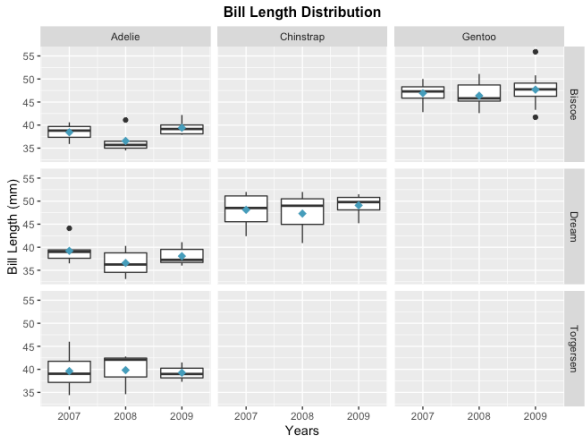

Fig. 7. Bill length distribution in 3 different islands over 3-year period

However, Adelie penguins have some variation depending on the islands they lived. In 2008, the average of the bill length for Adelies experienced a significant decline in Biscoe and Dream, even dthough they're almost at the same length in 2009.

Within 3 years and 3 different places, there's only few difference in the size of the Palmer penguins' bill. However, we can see the bill size difference between Gentoo, Chinstrap and Adelie. Gentoo which lived in Biscoe island had longer but thinner bills while Adelies had thicker but shorter bills. Meanwhile, Chinstrap had quite long and medium size bill ranging from 40 – 55mm long and 16 – 25 mm wide.

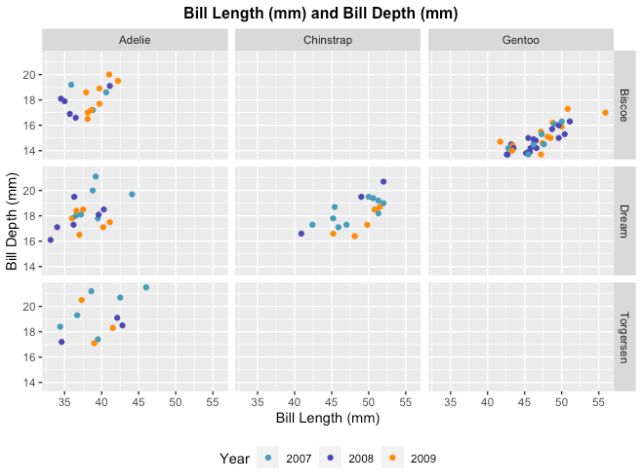

Fig. 8. Bill size difference in 3 different islands over 3-year period

There's a significant increase in the average of flipper length of Adelie penguins which lived in Biscoe over the 3-year time compared to the other islands. They had almost the same size within 3 years in Dream and Torgersen. Even though Adelies flipper length grew longer but it didn't have any effect on Gentoo's. There was an incline in Chinstrap flipper's length in 2008, but it experienced a slight decrease the year after. Overall, Gentoo had longer flippers than the other penguins while Adelie remained to have the shortest flippers of all penguins observed in the dataset.

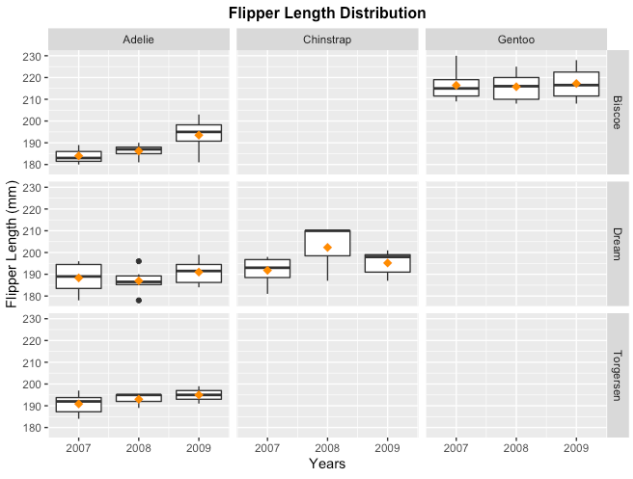

Fig. 9. Flipper Length Distribution in 3 different islands over 3-year period

To sum up, Gentoos which lived in Biscoe island had bigger body, longer bill and longer flipper than the rest of the penguins. Meanwhile, Adelies experienced ups and downs in their body size average and bill size in Biscoe island and was relatively stable in the other islands. They had shorter but thicker bills

with no significant bill growth in three islands. However, their flipper size average increased quite significantly within 3 years in Biscoe.

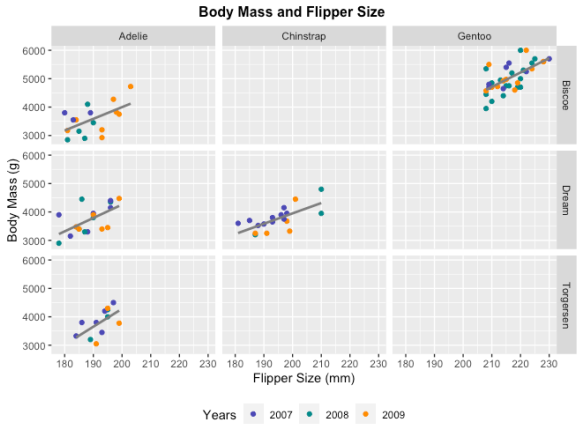

Fig. 10. Body Mass and Flipper Size Increase in 3 different islands over 3-year period

As stated above, Adelies had thicker bill and Gentoo had longer bill in general. Chinstrap had the medium size compared to the three of the penguin kinds.

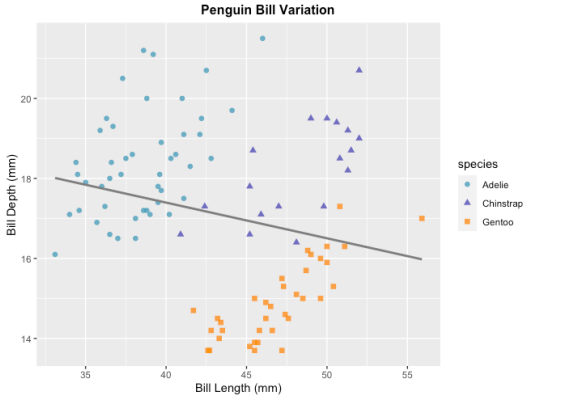

Fig. 11. Bill Size Variation in 3 different islands over 3-year period

IV. Sexing

We'll see the bigger picture of the comparison between male and female penguins by excluding the species.

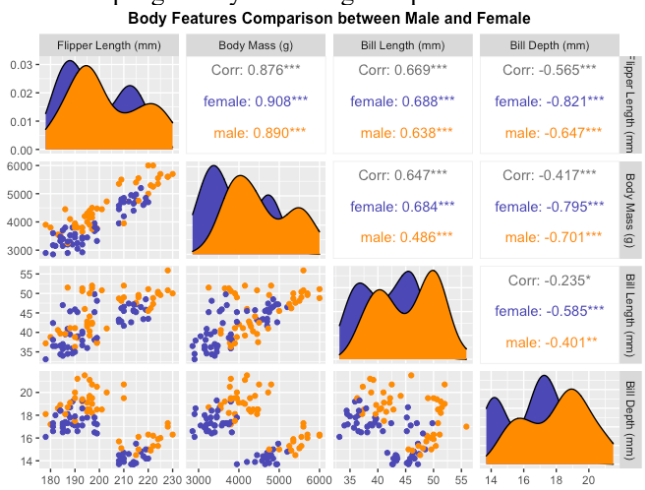

Fig. 12. Body Features Comparison between Male and Female Penguins (Exclude Species)

In general, body mass might be able to be used to distinguish between male and female but they're still a bit mixed in the body mass – bill length plotting since the average of body mass of Gentoo penguin was the biggest of all the three species put together. Nonetheless, we can say that male penguins have longer and bigger bills than the females from figure 12.

We'll try to look at the distribution between species to get more detailed information before taking a final summary about which feature might be the best to distinguish between male and female. In terms of body mass, Adelie penguins normally had bigger size than the female ones, the flippers can't be used to determine the gender. Nonetheless, bill size between male and female is seen clearly distinguished on the bottom right-most scatterplot within the 3 ggpair panels.

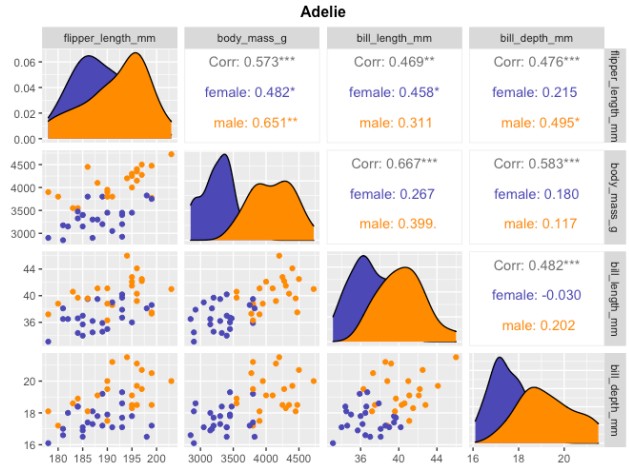

Fig. 13. Body Features Comparison between Male and Female Adelie Penguins

The same goes with Gentoo, most of females had smaller body mass and shorter flippers. However, The body mass between male and female are more premoninent among Gentoos. Few male Gentoos also have longer flipper than the female ones. The most prominent body feature that can tell the male and female Gentoos is the bill size as the one we have to distinguish the gender of Adelie penguins.

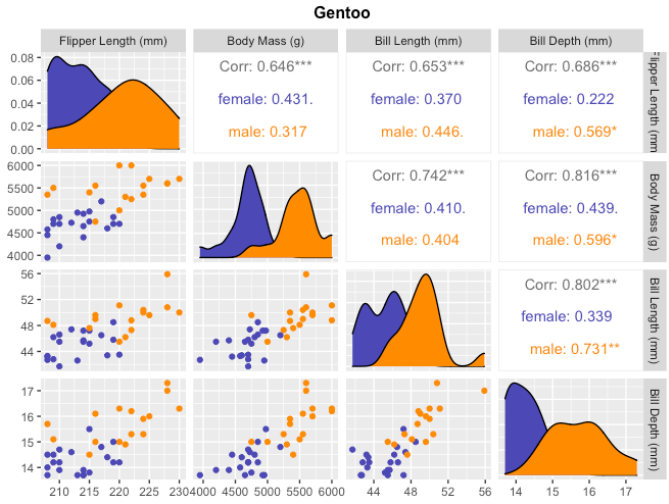

Fig. 14. Body Features Comparison between Male and Female Gentoo Penguins

On the other hand, we can't tell the Chinstrap gender apart by seeing the body mass. Only few male Chinstrap might be a little bit bigger than the other male Chinstraps. The best body

feature we can use to tell them apart is the bill size as the other species.

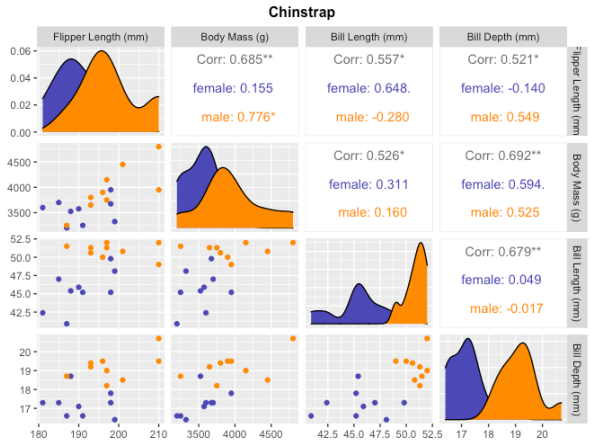

Fig. 15. Body Features Comparison between Male and Female Chinstrap Penguins

Based on the 4 scatterplots, the most promising pair of body features to tell the Palmer penguins gender apart is bill length and bill depth. The clusters are separated visually if we see the difference of the bill size between male and female regardless of their species.

To prove our hypothesis is correct, we'll try to use a machine learning technique called Logistic Regression. Logistic regression is often used for classification and predictive analytics such as suitable parameters we can use to determine a gender of a penguin by estimating the probability of an event happening based on a given dataset of independent variables. We calculate the outcome by dividing the probability of success and the probability of failure which is widely known as the log odds (natural logarithm of odds) [4].

We use a Repeated k-fold Cross validation to see improvement or difference in performance of the model in repeated folds.

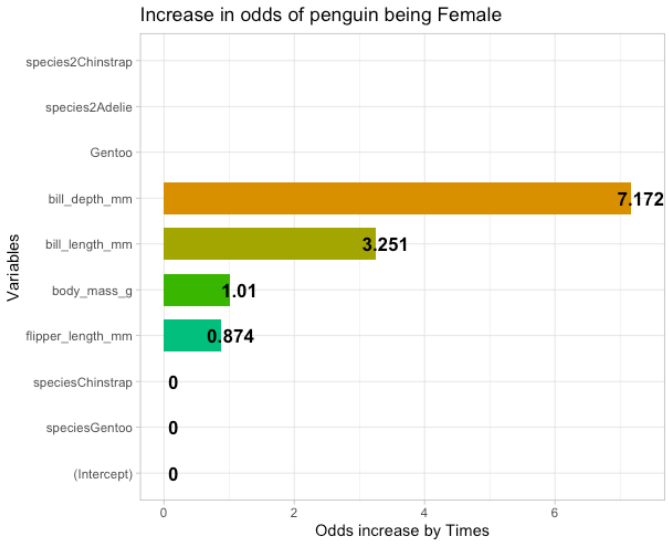

Fig. 16. Logistic regression to determine gender of the Palmer Penguins by comparing 1mm increase in each variable

Through the bar plot in Fig. 16, we found that bill depth and bill length are the two body features we can use to determine the gender of the penguins. 1 mm in bill depth increases the odds of the penguin being female by 7 times and almost 4 times for bill length. Species doesn't matter in terms of determining the sex of the penguins. The next part, I will be talking about the details about the two of the body features to see the spread of the dataset used in this analysis.

V. Bill Size Probability Distribution and Estimation

We select one of the best features which determines the gender of the penguins to project our probability distribution and talk a bit further about the estimation. We plot the density and the distribution using normal distribution as the data we use is continuous.

We do a plot to see the bigger picture of our sample data. The red line shows the normal distribution of our sample data while the purple one shows the true density of our sample data. Based on our data, we have a very long tail of our normal distribution line which indicates that our data is not normally distributed. After doing the Shapiro-Wilko test, our p-value is 0.01682. The test rejects the hypothesis of normality when p-value is less than or equal to 0.05 [3].

I also did the hypothesis testing for bill depth. However, the p-value for all species put together is 0.03192 which is below 0.05. It rejects the null hypothesis.

Therefore, we will try to divide them based on species to get the normal distribution of the population in our sample data. Nonetheless, we did the bill depth normal distribution for 3 penguins and only Chinstrap distribution rejects the null hypothesis which is 0.03192.

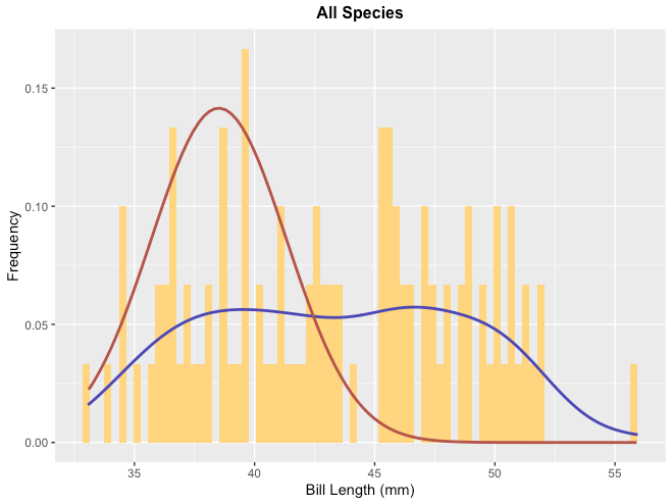

Fig. 17. Population estimation of all penguin species

Here is the density and the distribution of Adelie penguin from our dataset. From the plot, we can say that it seems to be normally distributed.

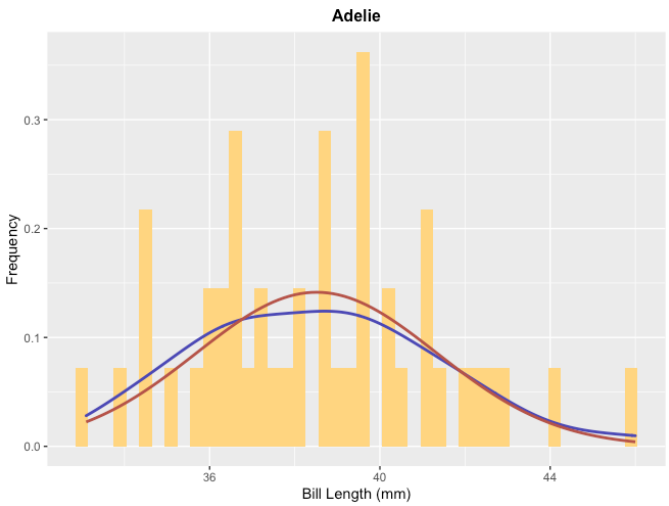

Fig. 18. Normal Distribution of Adelie's Bill Length

We can estimate the population mean (μ) by the sample mean and the population SD (σ) by the sample SD $\boldsymbol{S}$. The sample mean (μ) of Adelie penguin's bill length is 38.51304, the sample variance is 7.947826 and the sample standard deviation (σ) is 2.819189. To see the accuracy, we calculate the standard error of the mean (SEM). It reveals how far the sample deviates from the true mean. Our SEM for Adelie is 0.4156667.

We can also use Kernel Density Estimator (KDE) of f based on our sample data. Here is our KDE value with bandwidth 1.179:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
x	29.56	34.56	39.55	39.55	44.54	49.54
y	8.285e-05	6.714e-03	3.412e-02	5.001e-02	9.723e-02	1.241e-01

Fig. 19. Kernel Density Estimator (KDE) of Adelie Penguins

The population estimation (μ and σ) corresponding to the KDE can be described as follows:

Approximate Population Mean	Approximate Population Variance	Approximate Population Variance
38.51301	155.7423	12.47968

Fig. 20. Population estimation of Adelie penguin

Then, we test the normality of our dataset and we get its p-value equal to 0.8729 which is more than 0.05 (indicates that it's normally distributed).

Next is the density and the distribution of Gentoo penguin. The mean (μ) of Gentoo penguin's bill length is 48.24444, the variance is 11.6132 and the standard deviation (σ) is 3.407815. To see the accuracy, the standard error of the mean (SEM) for Gentoo is 0.8032297.

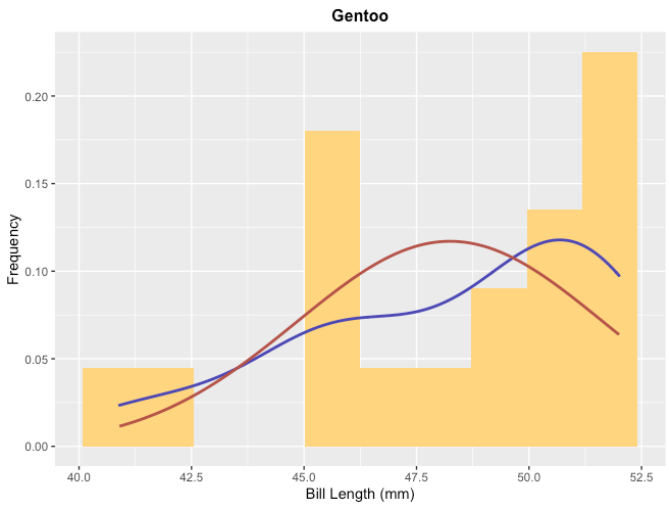

Fig. 21. Normal Distribution of Gentoo's Bill Length

The estimation of the population mean (μ) by the sample mean and the population SD (σ) by the sample SD $\boldsymbol{S}$ using Kernel Density Estimator (KDE) of f based on our sample data with bandwidth 1.721 as follows:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
x	35.74	41.09	46.45	46.45	51.81	57.16
y	0.0001518	0.0094600	0.0387598	0.0466190	0.0758013	0.1178387

Fig. 22. Kernel Density Estimator (KDE) of Gentoo Penguins

The p-value of Gentoo penguin is equal to 0.05561 which is more also than 0.05. It indicates that it's normally distributed. The population estimation (μ and σ) as follows:

Approximate Population Mean	Approximate Population Variance	Approximate Population Variance
48.24337	19.55582	4.422197

Fig. 23. Population estimation of Gentoo penguin

Here is the density and the distribution of Chinstrap penguin. The mean (μ) of Chinstrap penguin's bill length is 46.91944, the variance is 9.123325 and the standard deviation (σ) is 3.020484. The standard error of the mean (SEM) for Gentoo is 0.503414.

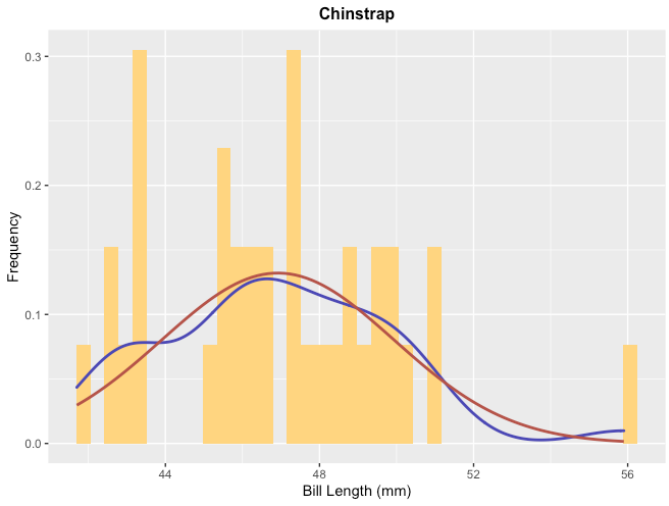

Fig. 24. Normal Distribution of Chinstrap's Bill Length

The estimation of the population mean (μ) by the sample mean and the population SD (σ) by the sample SD $\boldsymbol{S}$ using Kernel Density Estimator (KDE) of f based on our sample data with bandwidth 1.123 as follows:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
x	38.33	43.56	48.80	48.80	54.04	59.27
y	0.000111	0.005033	0.029858	0.047704	0.088232	0.127586

Fig. 25. Kernel Density Estimator (KDE) of Chinstrap Penguins

The population estimation (μ and σ) as follows:

Approximate Population Mean	Approximate Population Variance	Approximate Population Variance
46.91935	23.82208	4.880787

Fig. 26. Population estimation of Chinstrap penguin

The p-value of Chinstrap penguin is equal to 0.2451 which is more also than 0.05. It also indicates that it's normally distributed.

Normal distribution is a really good method to know the distribution of the population in our sample data in which the data is continuous such as penguin bill length. It depends on two parameters such mean and standard deviation. We can see the entire distribution only through those two parameters. The sample mean itself is a good estimator for μ, while the sample standard deviation is a good estimator for σ. Any variable that displays normal distribution is feasible to be forecasted with higher accuracy and it can help simplifying the model we have as shown in the plot of normal distribution with all species (not normally distributes as it's visually clear), Adelie, Gentoo and Chinstrap. We chose the normal distribution because of its central limit theorem in which independent random variables are summed up.

VI. Summary

The sample data we have is not normally distributed when we don't classify the penguins by its species. It gives the p-value of 0.01682 which rejects the null hypothesis to become a normal distribution. However, when we classify them into different species and chose one of the body features such as Bill Length, we get the p-value of each species 0.8729, 0.05561, 0.2451 for Adelie, Gentoo and Chinstrap. I chose Bill Length because it's one of the body features which can help us to determine the gender of the penguins based on the scatterplots we discussed in chapter IV. However, it rejects the p-value if we apply the normal distribution for bill depth with Chinstrap dataset. We might need to consider different kinds of distribution for bill depth such as uniform or gamma distribution for future work.

In terms of islands where the penguins live, it seems there is no significant difference of penguin body mass which lived on Biscoe island and other islands except for Gentoo which lived in Biscoe island had a bigger body than others. Gentoo which lived in Biscoe Island had longer but thinner bills while Adelies had thicker but shorter bills. Meanwhile, Chinstrap had quite long and medium size bill ranging from 40 – 55mm long and 16 – 25 mm wide. Gentoos which lived on Biscoe island had bigger body, longer bill and longer flipper than the other penguins. Meanwhile, Adelies experienced ups and downs in their body size average and bill size on Biscoe island and was relatively stable on the other islands. They had shorter but thicker bills with no significant bill growth in three islands. However, their flipper size average increased quite significantly within 3 years in Biscoe.

References

[1] Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus *Pygoscelis*). PLOS ONE 9(3): e90081. https://doi.org/10.1371/journal.pone.0090081

[2] Wells J 2019, *Evolution and Science Today*, Discovery Institute, accessed 9 October 2022, https://evolutionnews.org/2019/09/is-gender-in-penguins-a-human-construct/

[3] Bevans R 2020, *Understanding P-values | Definition and Examples*, Scribbr, accessed 14 October 2022, https://www.scribbr.com/statistics/p-value/

[4] IBM, 2022, *What is Logistic Regression*, accessed 20 October 2022, https://www.ibm.com/uk-en/topics/logistic-regression