Penguin Report

Jack Westmoreland (200394930)

1. Exploritory Data Analysis

Before conducting any statistical analysis we must first conduct some exploratory data analysis on our data set. Firstly, i will take a look at the fields the data set contains.

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

The data set consists of measurements taken from penguins from the Anvers Islands. eight measurements of each penguin were taken:

- **species**: The species the penguin belongs to
- **island**: The island the penguin was measured on
- **bill_length_mm**: The length of the penguin's bill (millimeters)
- **bill_depth_mm**: The depth of the penguin's bill (millimeters)
- **flipper_length_mm**: The length of the penguin's flipper (millimeters)
- **body_mass_mm**: The body mass penguin (grams)
- **sex**: The sex of the penguin (male or female)
- **year**: The year the penguin was added to the data set

1.1 Summary Statistics

Using R we can gather some summary statistics on the penguins in the sample data set.

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :76   Biscoe   :101   Min.   :32.10   Min.   :13.10
##  Chinstrap:45   Dream    : 76   1st Qu.:40.10   1st Qu.:15.47
##  Gentoo   :79   Torgersen: 23   Median :45.20   Median :17.15
##                                 Mean   :44.41   Mean   :17.07
##                                 3rd Qu.:48.62   3rd Qu.:18.62
##                                 Max.   :59.60   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :174.0     Min.   :2850   female: 96   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3569   male  :104   1st Qu.:2007
##  Median :197.5     Median :4100                Median :2008
##  Mean   :201.7     Mean   :4281                Mean   :2008
##  3rd Qu.:215.0     3rd Qu.:4950                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
```

From this we can see a few things. Firstly, most penguins in the sample P set are from the species Adele and Gentoo, with fewer Chinstrap penguins being found. Most of these penguins were found on the islands Biscoe and Dream. Perhaps there is some correlation between species and island. The sample set contains a roughly even split of male and female penguins (being only slightly male skewed) and the penguins were recorded in the data set between 2007-2009.

We can also see the sample mean $\bar{x}$ of the quantitative attributes taken from the penguins. At a cursory glance we can also see that for these values the maximum and minimum values in the sample set do not lie massively far away. This hints at the data set not containing any outliers, however to further assess this more accurate more accurate measures must be taken. Which will be done next by plotting our sample set in histograms.

1.2 Penguin Measurment Plots

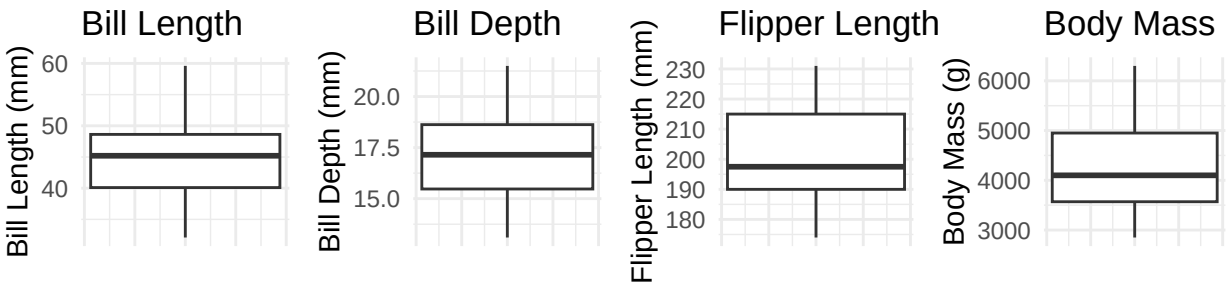


Figure 1: All Penguin Measurements

From plotting the four quantitative measurements taken from P we can see that our measurements are all fairly symmetrical. This is apparent due to Q_1 and Q_3 of our data being a similar distance from our sample's median. Furthermore, we can see that the sample does not exhibit any outliers, as there are no values outside the ranges:

$$[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$$

We can also see that bill length, flipper length and body mass all have a fairly large interquartile range indicating there is some spread in the data set. This can investigated by checking our sample standard deviation (s) for our variables.

```
## Standard Deviation of Bill Length (mm): 5.444084
```

```
## Standard Deviation of Bill Depth (mm): 1.985753
```

```
## Standard Deviation of Flipper Length (mm): 14.62589
```

```
## Standard Deviation of Body Mass (g): 825.7953
```

As we can see there is some sample standard deviation in our bill/flipper length and body mass variables. This is most noticeable for body mass where our samples are on average 825 grams from the body mass mean ($\bar{x}$). s is also fairly large for bill length and flipper length, whereas bill depth is concentrated at it's $\bar{x}$

1.3 Difference in Species

We could expect that different species of penguins could have different mean $\bar{x}$ measurements for our sample data set attributes. To check this we can plot box plots comparing between the species.

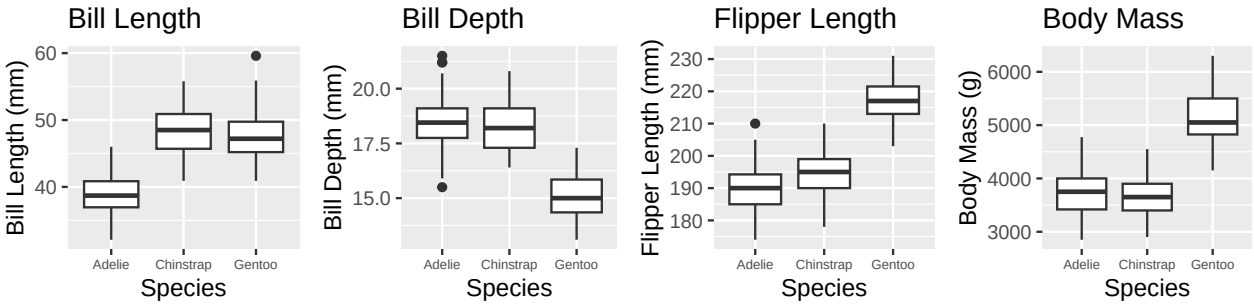


Figure 2: Penguin Measurements by Species

As we can see there is large variation of these measurements in our sample set P depending on the species. For example Gentoo penguins seems to have much longer flippers, weigh more, but have a smaller bill depth than the other penguins. This makes penguin species an important measurement when attempting to calculate our expected values of these measurements for these penguins.

2. Fitting A Distribution

From sample set P it would be helpful to fit a probability distribution for some of the penguins' measurements. This will allow for us to estimate the probability of penguins outside the sample holding specific measurements. To do this I will first plot a histogram of body mass for our sample set s, which should give a rough idea of what distribution describes the sample data.

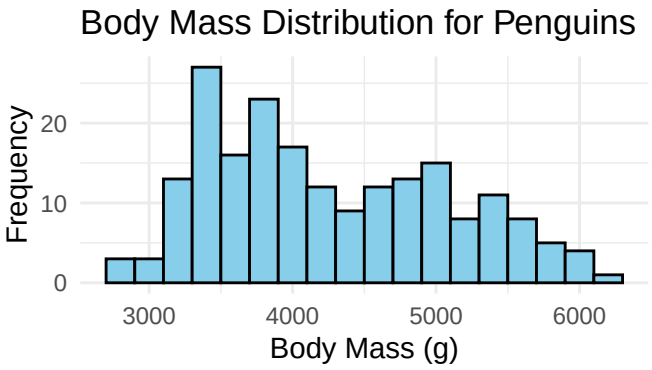


Figure 3: Body Mass Histogram for all Species

From plotting body mass we can see that the distribution of the sample data does not seem to resemble any continuous distributions. This is likely due to the large variation of body mass between species we observed in Figure 2. Therefore splitting our penguin species up and plotting the distribution by species may get us better results.

From the sample set, only 45/200 of our penguins are Chinstrap species. Therefore to give a more robust statistic for this species it would be beneficial to combine them with the Adelie species. This would increase the sample size for statistical analysis. To check if this is viable, a t-test can be performed to check for statistical differences between these species' body mass. First we must conduct a Barlett test to check if the variance for these sets are equal.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  body_mass_g by species
## Bartlett's K-squared = 2.0204, df = 1, p-value = 0.1552
```

Our Bartlett gives us a p-value of 0.15, which is significantly large. Therefore we can assume that the null hypothesis H_0, that these is no significant variance σ^2 differences between the species body masses, holds and we can proceed with a t-test for body mass between adelie and chinstrap species.

```
## P-value: 0.3376541
```

```
## 95% Confidence Interval: -79.20087 229.0839
```

From this two sample t-test we can see that the p value is significantly above the 0.05 threshold for saying there is significant differences in these species body size. Additionally, the 95% confidence interval of this test contains 0. Further indicating there are no significant differences. This should allow for the combination of these species when estimating their body mass probability distributions.

2.1 Normal Distribution.

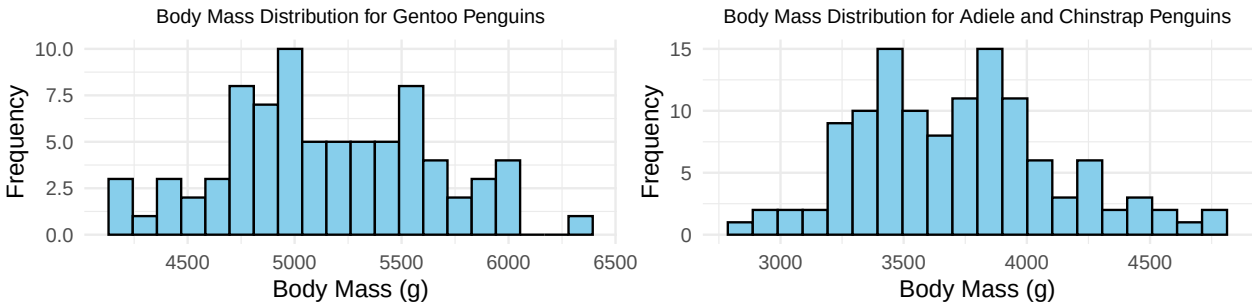


Figure 4: Body Mass Histogram for Gentoo and combined Adelie,Chinstrap Penguins

These histograms seem to be resemble a normal distribution. Therefore it could prove useful to fit to our data. This can be further investigated by using quantile-quantile plots to visually see if the values from G and $(C \cup A)$ resemble the normal distributions. If the plotted quantile closely align with the normal distribution quantile line it is likely that this is the case.

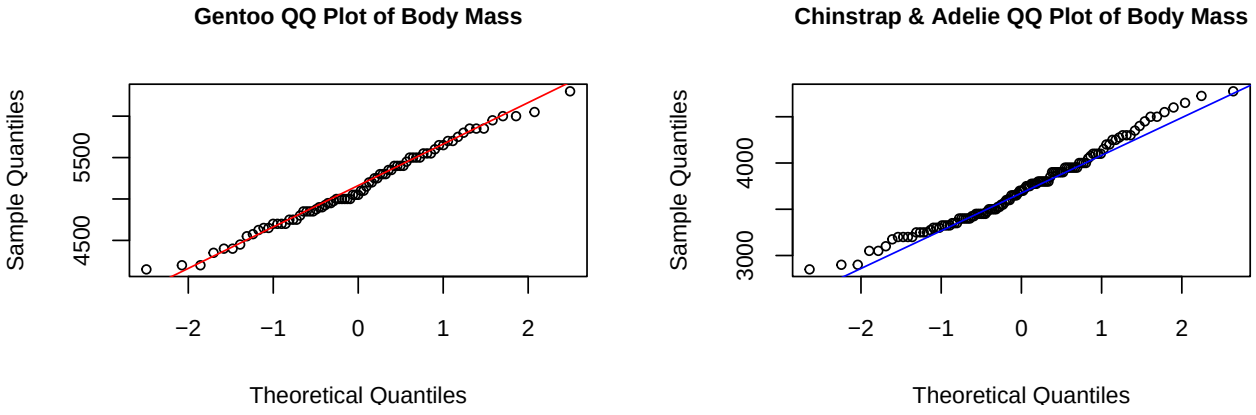


Figure 5: QQPlots for body mass normal distribution

From our QQplots we can further verify that both our Gentoo set G and the combined Adelie Chinstrap set $C \cup A$ seem to resemble a normal distribution. To fit this distribution we must first calculate the mean μ and variance that σ^2 best describe the normal distribution of G and $C \cup A$. To do this the log likelihood function or normal distribution can be used,

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

The minimized negative log likelihood function for G and $C \cup A$ will be calculated using R.

G $\hat{\mu}$**:** 5143.35. G $\hat{\sigma}^2$**:** 2.3188636×10^5

G $\bar{x}$**:** 5143.35. G s^2**:** 2.3188636×10^5

$(C \cup A)$ $\hat{\mu}$**:** 3718.18. $(C \cup A)$ $\hat{\sigma}^2$**:** 1.7117708×10^5

$(C \cup A)$ $\bar{x}$**:** 3718.18. $(C \cup A)$ s^2**:** 1.7117708×10^5

From maximizing the log normal distribution likelihood function we get the variables for both these sample sets' normal distribution variables. We can check these against the sample mean $\bar{x}$ and sample variance s^2 of these sets to see that they are identical. This is expected as the recommended estimators for normal distribution are actually these values.

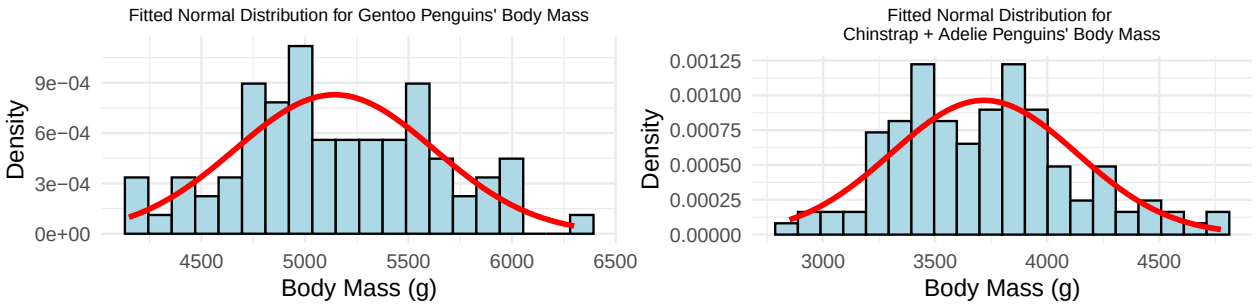


Figure 6: Fitted Normal Distributions

G **Skewness:** 0.12. $(C \cup A)$ **Skewness:** 0.38

G **Kurtosis:** 2.44. $(C \cup A)$ **Kurtosis:** 2.7

From the fitted plots (fig 6) we can visually see that the normal distribution does a relatively good job fitting to our data. The histogram destiny looks to fairly resemble the red normal distribution line fitted to the sets. To further test these distributions the skewness and kurtosis of the sets have been measured. From these measures we can see that G is slightly right skewed compared to the normal distribution skew of 0. However this is still quite close to 0 thus reinforcing this models validity. The skewdness of $(C \cup A)$ is a little more, implying a right skew to the data set however this is still fairly close to 0.

By measuring the Kurtosis of G and $(C \cup A)$ we can see that the tails of these sets are slightly heavier than that of the normal distribution. Meaning that there are more extreme outlines in the sample sets than expected in the normal distribution, which has a kurtosis of 3. Overall from visual and quantitative assessment, sets G and $(C \cup A)$ appear to be reasonably close to a normal distribution. With the small size of our sets ≈ 200, likely effecting G and $(C \cup A)$s accuracy when approximating the penguin population. The normal distribution likely is the best fit for our data. However, due to the slight differences in the normal distribution and the sets the sample measurements may vary from the true population to some degree.

3. Sexing

To find out which characteristics of penguins would be best at estimating their sex we can first plot a box plot of each penguins' measurements by their sex.

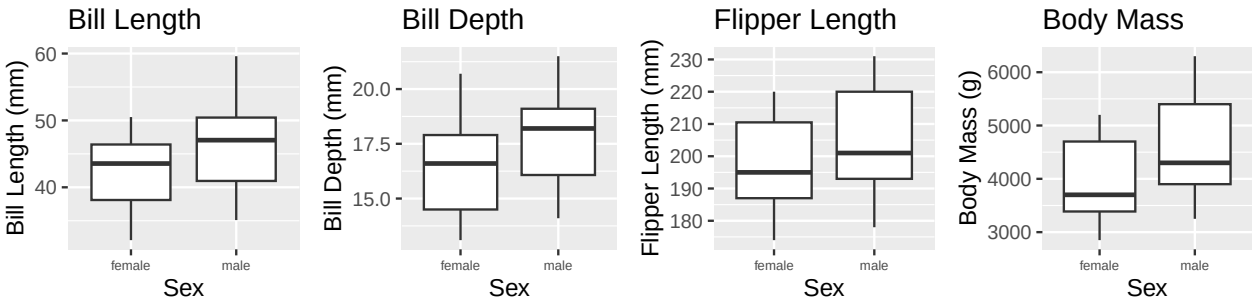


Figure 7: Penguin Measurements by Sex

From these plots we can observe some difference in the median X for these measurements depending on their sex. Specifically the male penguins seem to have a much larger body mass median vs female. This hints at there being some statistical difference in penguins from our male subset and female subset.

We can represent the hypotheses for further testing as:

- H_0: No difference in measurements between males and females, $\mu_{\text{male}} = \mu_{\text{female}}$.
- H_1: There is a difference, $\mu_{\text{male}} \neq \mu_{\text{female}}$.

3.1 Hypothesis Testing

To confirm if hypothesis h_1 holds statistical significance a two-sample t-test can be performed on $M \subset P$ $F \subset P$ where:

$M = \{p \in P \mid p.sex = "male"\}$

$F = \{p \in P \mid p.sex = "female"\}$

For running t-tests the sets F & M must be **IID**, independent and identically distributed. It is safe to say that the penguins measurements are not effected by each other. One penguin p will not have its flipper length effected by another and therefore are independent. F & M also have measurements coming from the same distributions within their sets and are therefore identically distributed.

First me must check if the sample variance between male and female penguins us roughly equal for both sets so that $\sigma_M^2 = \sigma_F^2$. To check this a a Barlett's test can be performed with R.

```
## P-value: 0.09668118
```

With this result we can accept the null hypothesis that there is no difference in body mass variation for F and M as $pvalue < 0.05$ therefore a two sample t-test should be conducted.

```
## Sex Body Mass Welch Test Result
```

```
## P-value: 0.00000002141344
```

```
## 95% Confidence Interval: 418.339 845.2027
```

From this result we can see there is a large difference in body size for our sets F and M, with a p value far blow 0.05. The 95% confidence interval puts this difference around 420-845g. Making this a potentially great marker for distinguishing between male and female penguins. From performing the same tests on each of the penguins measurements we get the following.

Measurement	Bartlett's P-value	T-test's P-value	95% Confidence Interval (Lower)	95% Confidence Interval (Upper)
Bill Length	0.0903	0.00000396	2.019150	4.888382
Bill Depth	0.9202	0.00000003	0.997780	2.025137
Flipper Length	0.116	0.0007661	2.915752	10.869504
Body Mass	0.0967	0.00000002	418.338992	845.202675

This table shows that there are several strong differences in male and female penguin measurements. Body mass and Bill Depth seem to be the two best measurements for this. However, the other measurements are all strong candidates, with p-values far below the 0.05 threshold needed to reject H_0. Furthermore, for all measurements the 95% confidence intervals places the differences fairly far from 0 meaning these measurements quite reliable for sexing male and female pigeons. Therefore with a high degree of certainty we can say for all measurements H_1 holds.

3.2 Logistic Regression

To actually make predictions of penguin sex we can fit a logistic regression model to our data using the four measurements. Logistic regression is a statistical model used to predict binary outcomes (in this case male or female) from from our variables. It works by predicting the log odds (the odds of male or female) as a linear function of our penguins variables. These odds are then turned into a probability of the penguin belonging to a sex with use of a logistic function. Essentially finding the boundaries that separate our males

and females for these variables. This differs from the two sample t-tests done in section 3.1 which can only take into account one variable of P to determine differences between F and M

```
## # A tibble: 5 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       -54.0      10.2        -5.31  1.10e- 7
## 2 body_mass_g         0.00527   0.00103     5.10  3.36e- 7
## 3 bill_length_mm      0.0985    0.0676      1.46  1.45e- 1
## 4 flipper_length_mm  -0.0312    0.0447     -0.699 4.84e- 1
## 5 bill_depth_mm       1.98      0.305       6.49  8.56e-11
```

```
## Model Accuracy:  88.5 %
```

From fitting the logistic regression model we can that the most significant measures for determining sex are identified as body mass and bill depth. These values have much lower p values than others for sex estimation. This is consistent with our two sample analysis. Furthermore, the σ error and z vales (statistic) are extremely small for these measures, further proving that there are good variables for predicting sex. The model achieved from these parameters achieved a 88.5% accuracy, meaning the sex of each penguin was correctly predicted 88.5% of the time.

4. Island Characteristics

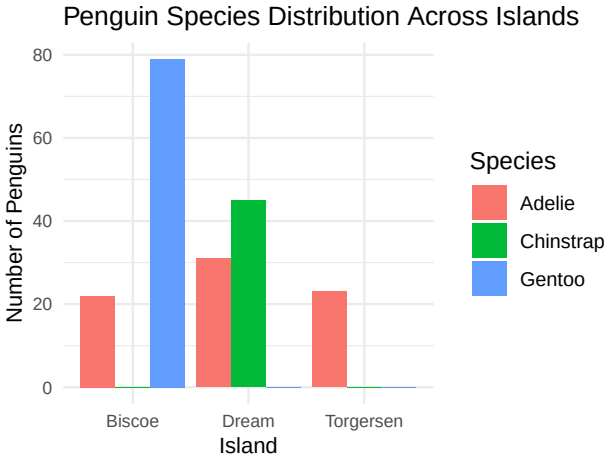
The penguin sample set P has penguins who come from three different islands "Biscoe", "Torgersen" and "Dream" to find out if islands have significant statistical impact on penguin measurements we can split sample set P into three subsets. and create our hypothesis.

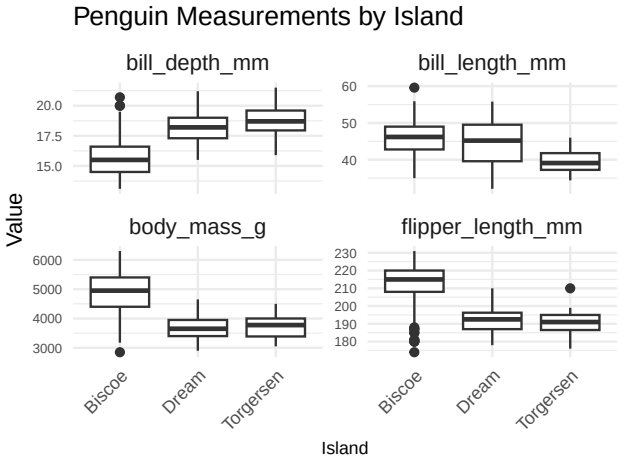
$B = \{p \in P \mid p.island = "Biscoe"\}$

$T = \{p \in P \mid p.island = "Torgersen"\}$

$D = \{p \in P \mid p.island = "Dream"\}$

- H_0: No difference in measurements between B, T and D, $\mu_B = \mu_T = \mu_D$.
- H_1: There is a difference, $\mu_B \neq \mu_T \neq \mu_D$.





B, T and D all contain a differing amounts of penguins from each species. For example Gentoo penguins are only found on B and Chinstrap on D in the sample set. Therefore, measuring the differences of all penguins on each island (as seen in the above box plot) would more than likely instead measure the differences in species. Which has already been completed in section 2. Therefore, to ensure I am only measuring statistical differences due to the island of the penguins I will be updating the sets to only contain Adelie penguins. Which are present on all 3 islands.

$B = \{p \in P \mid p.island = \text{"}Biscoe\text{"} \text{ and } p.species = \text{"}Adelie\text{"}\}$

$T = \{p \in P \mid p.island = \text{"}Torgersen\text{"} \text{ and } p.species = \text{"}Adelie\text{"}\}$

$D = \{p \in P \mid p.island = \text{"}Dream\text{"} \text{ and } p.species = \text{"}Adelie\text{"}\}$

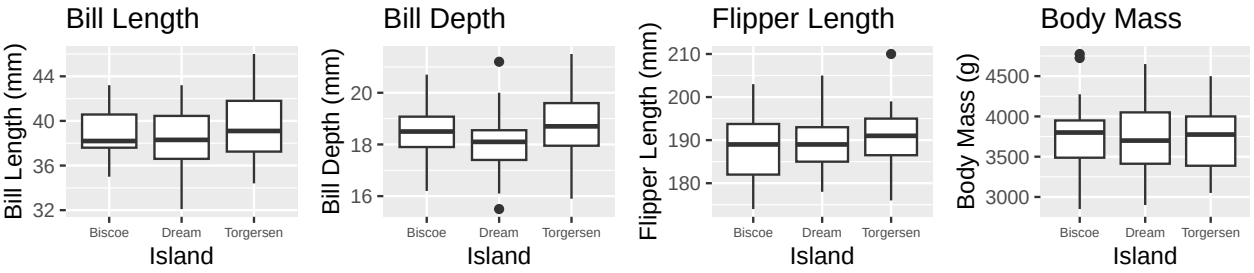


Figure 8: Adelie Measurements by Island

From the box plots of sets B, D and T we can see little variation between the medians and IQR of Adelie penguin measurements between the three islands.

4.1 ANOVA Tests

For statistical analysis between the three islands an ANOVA test will be conducted on each penguin measurement. ANOVA tests work by calculating the between group variability SSB and within group variability SSW for each group (island).

$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$

$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$

k: number of groups

n_i: number of observations in group i

k: number of groups

$\bar{X}_i$: mean of group i

$\bar{X}$: overall mean

X_{ij}: individual data point j in group i

SSB measures the variation in between groups, by calculating how much the groups mean differs from the overall mean and multiplying this by the number of observations in the group. This is summed for each group (island). SSW measures the in group variability by summing the squared difference from the mean for each data point in the group. This is done and summed for all groups. We then calculate F ratio $f = \frac{SSB}{SSW}$ which is used in a look up table to find it associated p value.

ANOVA tests assume that the data is normally distributed, and of equal variance between groups. From barlett's testing we can see that all data has equal variation (p-vale > 0.5 for all measurements). Additionally, by plotting Q-Q plots like as done in section 2.1 confirms the data is normally distributed. Making ANOVA tests a good choice for statistical analysis across islands.

```
## body mass: 0.9167 bill length: 0.4562 bill depth: 0.096 flipper length: 0.3732
```

From the p values, derived from ANOVA tests on each measurement. there is not enough statistical difference between adelie penguins on different island to reject null hypothesis H_0. The p-values from these tests are too large to suggest the data shows significant differences in measurements between islands. The p-value for bill depth could indicate that there is variation in measurements between islands, but this is still too large to reject H_0. Therefore it is unlikely that there is differences in penguin measurements between islands when adjusted for the differences for species.