MAS8403: Palmer Achipelago Penguins

210431461 | 21/10/22

Introduction

To the West of the Antarctic Peninsula, extending North and South of the Palmer Basin, is the Palmer Long Term Ecological Research (LTER) (Rutgers and LTER (2022)) study area. Midway down the the Antarctic Peninsula, on Anvers Island, is Palmer Station. Researchers are staffed there to monitor the polar marine biome, including the local penguin population (Foundation (n.d.)).

The LTER researchers are using penguin-borne sensors to inform long-term studies on penguin population dynamics and improve our understanding of how Antarctic penguins are adjusting to rapid climate changes (LTER (n.d.)). The most dramatic effects of climate change are being observed in our polar regions (LTER (n.d.)).

This report is informed by a dataset called `penguins` from the `pamlerpenguins` R package; it is one of two packages provided by Palmer LTER researchers (Hill, Horst and Gorman and Gorman (2020)). The dataset is pre-processed, so accuracy and quality are assumed. It is assumed that this is a sample of the data collected by Palmer Station as the dataset provides 333 observations of penguins between 2007 and 2009. The `set.seed()` and `sample` functions in `R` were to generate a random, representative sample to inform this report.

Objectives

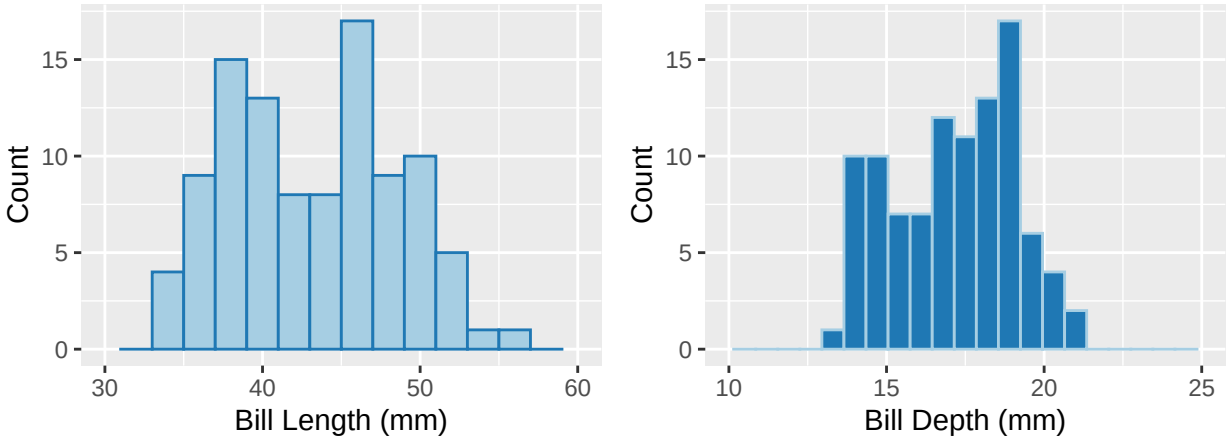
The exploratory data analysis outlined in this report explores a sample of 100 penguins from the `penguins` dataset. There are 4 objectives for this analysis:

1. identify an appropriate probability distribution to represent at least one measurement variable (bill length, bill depth, flipper length and body mass);
2. find estimates for the parameters of the distribution of your data;
3. identify which variables are likely to reliably estimate the sex of a penguin; and
4. identify if the penguins' location (island) appears to have a significant impact on any of its physical characteristics.

Data Exploration

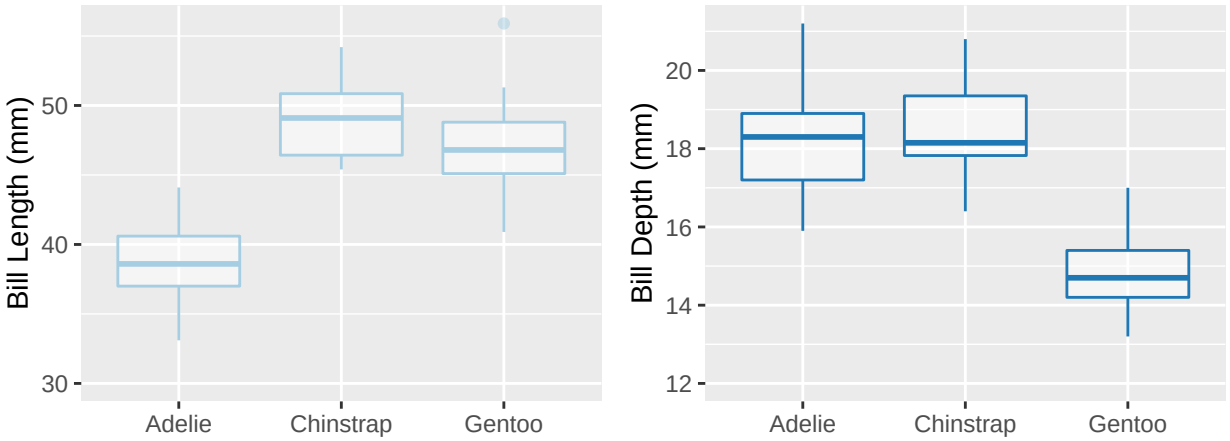
Objective 1 (Distribution)

The data sample includes 8 variables that provide information relating to the 100 penguins. The species (Adelie, Chinstrap or Gentoo), island (Biscoe, Dream or Torgerson), and sex (male or female) are nominal, qualitative values. The year is discrete quantitative data that identifies when the variables were recorded (2007, 2008 or 2009). The variables bill length (mm), bill depth (mm), flipper length (mm) and body mass (g) are quantitative, numerical measures of each penguin. These 4 variables of measurement data are continuous and random; the variables adopt a smooth range of values (Newcastle University (2022)). The plots below show this (all histograms are provided in `Appendix A`).



These histograms visualise the distribution of penguins' bill length (BL), bill depth (BD), flipper length (FL), and body mass (BM) data. The measurement data appears to be multimodal with some variance. However, there may be elements, such as the penguins' species, which are influencing the distribution of this data.

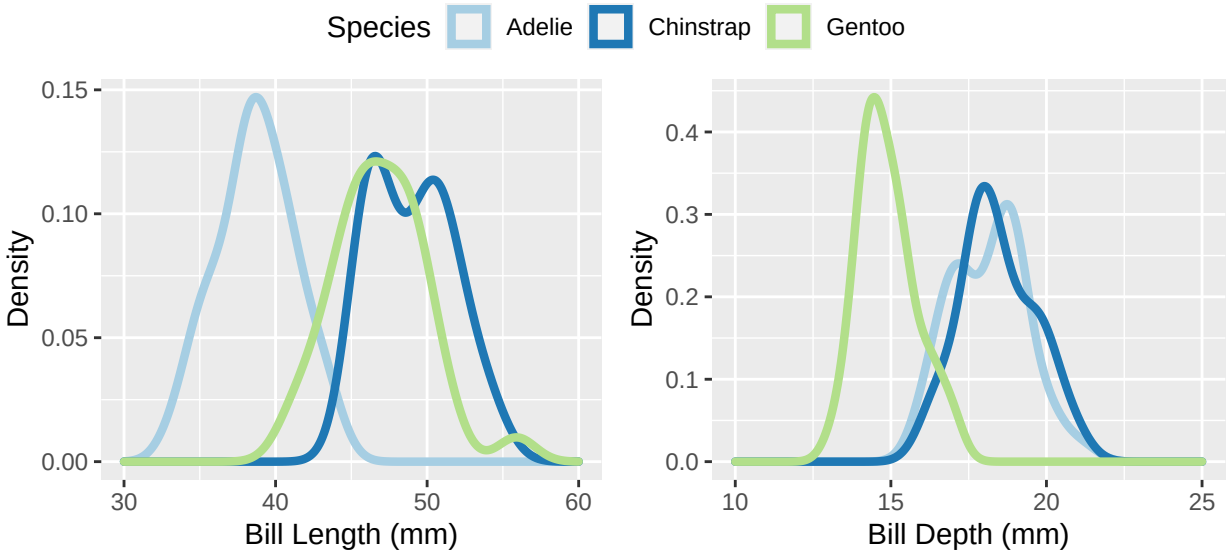
Filtering the data demonstrates that the BL, BD, FL, and BM changes by penguin species. The boxplots for BL and DP are provided below (all boxplots are provided in `Appendix B`).



In these plots, we observe that Adelie penguins have shorter bills and Gentoo penguins have

thinner bills. It is likely to be more significant, therefore, to observe the distribution, and estimate the parameters of the population, separately for each penguin species.

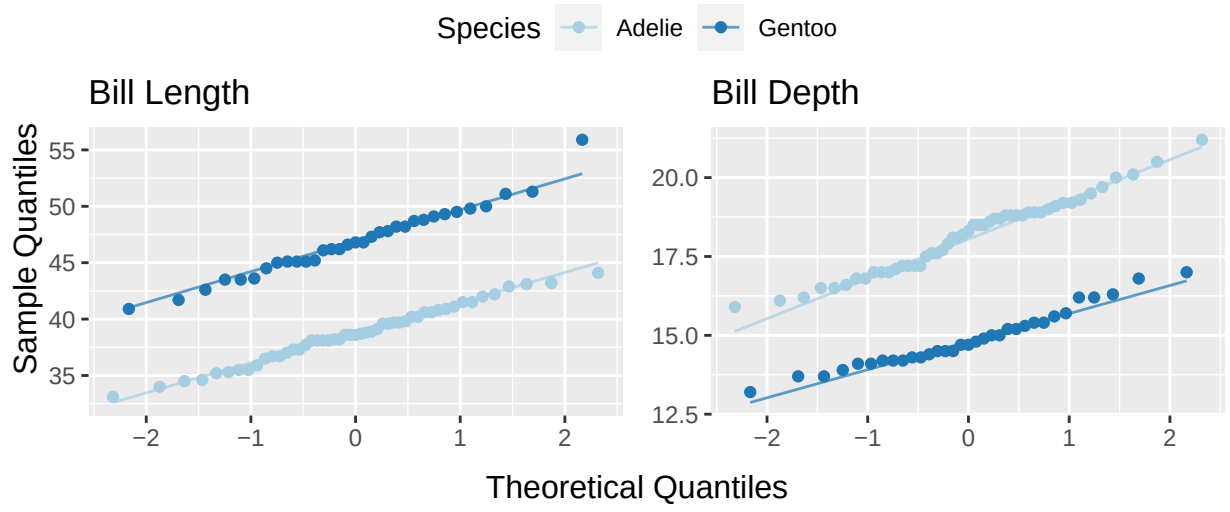
For each species, the distribution of BL and BD are visualised below (all density distributions are provided in `Appendix C`).



Measurement data (e.g. weight or height) of a population frequently follows a normal probability distribution. Initially, however, the density distributions of BL, BD, FL, and BM of the penguin species do not appear to. On closer inspection, some variables do appear to approximate a normal distribution, especially the BL, FL and BM of Adelie penguins.

It is important to remember that our data sample is small. In the sample of 100 penguins, there are 49 Adelie, 33 Gentoo, and 18 Chinstrap. It is likely that there is too little measurement data for Chinstrap penguins to reliably identify its appropriate probability distribution.

The Q-Q plots below, therefore, test whether or not the BL and BD data of Adelie and Gentoo penguins are normally distributed (all Q-Q plots are provided in `Appendix D`).



These Q-Q plots demonstrate that the BL, BD, FL and BM of Adelie and Gentoo penguins does approximate a normal probability distribution. It is important to note that the previous density distributions plots suggest there is another element influencing the distribution of our data. `Appendix E` presents Q-Q plots that clearly demonstrate the data distributions for male and female penguins are different for all species, and this will be explored further in `Objective 2`.

Objective 2 (Distribution Parameters)

A normal distribution is characterised by two paramenters; these are the 'mean' and 'standard deviation' (Bhandari (2022)). The 'mean,' its 95% 'confidence interval' (95% CI), and 'standard deviation' (SD) of BL and BD from the sample data for each penguin species are presented in the tables below (all parameters are provided in `Appendix F`).

Table 1: Summary Statistics

	Bill Length			Bill Depth		
Species	Mean (mm)	95% CI (mm)	SD	Mean (mm)	95% CI (mm)	SD
Adelie	38.70	37.98-39.42	2.58	18.15	17.81-18.49	1.21
Gentoo	46.88	45.83-47.93	3.09	14.88	14.57-15.19	0.91
Chinstrap	48.99	48.08-49.90	2.68	18.44	18.03-18.85	1.19

The 'mean' and 'standard deviation' as parameters of our BL, BD, FL and BM distributions will be limited as estimators for the population parameters. There is only one sample to inform the estimators and, when filtered by species, the sample set is small. The 95% 'confidence interval,' therefore, provides us with the interval that indicates how close the estimated 'mean' is likely to be to the true value; we can be 95% confident that the population mean will be between this interval.

However, as our data approximates a normal distribution, the central limit theorem applies and as we increase the sample size the sample 'mean' will approach the population 'mean.' Similarly, if there were more sample sets, the 'mean' and 'standard deviation' of the samples' parameters would better estimate the population parameters.

Objective 3 (Estimate Penguin Sex)

LTER researchers at Palmer station would like to estimate the sec of a penguin from measurement data to avoid the need for invasive procedures that cause penguin distress.

The Q-Q plots in `Appendix E` suggest that the data distribution of male and female penguins differ, regardless of their species. The difference between the BL and BM appears most visually obvious (boxplots are also provided in `Appendix G`. Two-sample t-tests are used to test whether this difference is significant enough to reliably estimate the sex of a penguin.

For example, the BM two-sample t-test for Adelie penguins is set out below.

Adelie two-sample t-test for BM: Of the 49 Adelie penguins, there are 26 females and 23 males. We would like to compare BM observations from Adelie females (F) to determine that their 'mean' is different to observations from Adelie males (M).

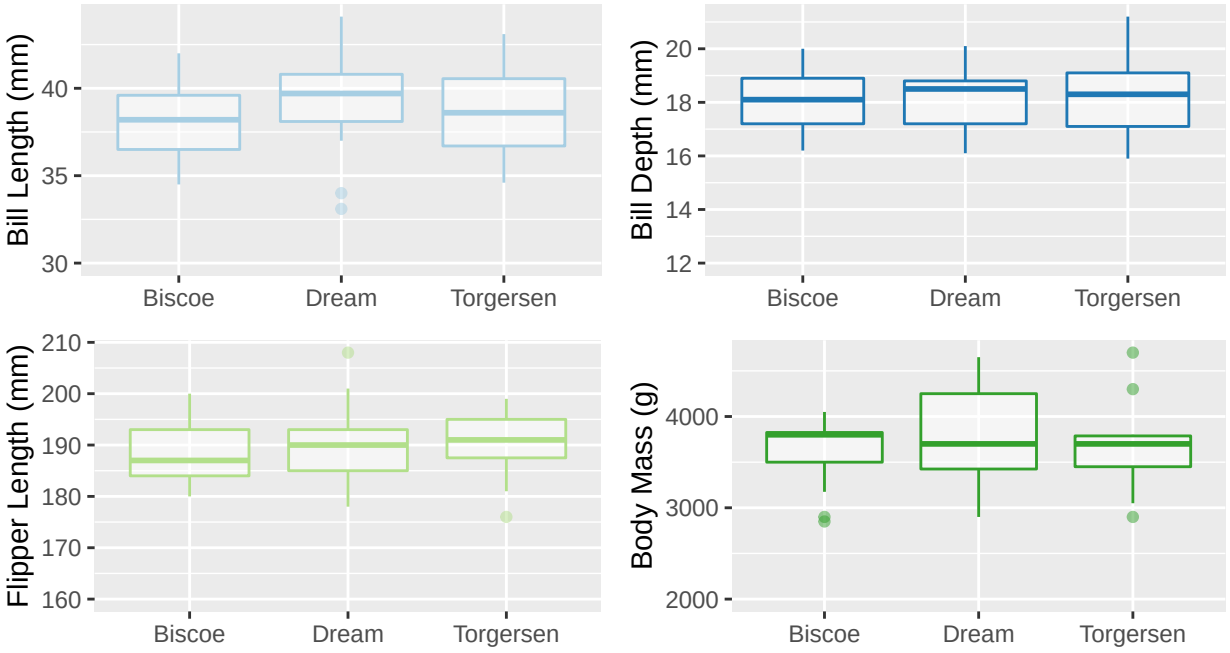
$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.59). The `t.test` function provides us with a p-value of 1.21e-7. This suggests that there is very strong evidence against H_0 and to reject it in favour of H_1.

All two-sample t-tests are provided in `Appendix H`. These tests suggest that in our sample of Adelie and Gentoo penguins, the BM 'mean' is more significant than the BL 'mean' and, therefore, more likely to reliably estimate the penguin's sex. In our sample of Chinstrap penguins, the BL 'mean' appears to be more significant than the BM 'mean'; however, this finding is limited by the Chinstrap sample size.

Objective 4 (Island Impact)

Only one species of penguin is found on all three islands; this is the Adelie penguin. The relevant bar chat is in `Appendix I`. The boxplots below show the BL, BD, FL and BM for the Adelie penguins on each island.



The boxplots suggest that location may impact the phsyical characteristics of the Adelie penguins. for example, the range of Adelie penguins' BM is larger on Dream than other islands. To confirm the impact of location, two-sample t-tests were used (and are provided in `Appendix J`). An example hypothesis test is set out below.

Dream and Biscoe two-sample t-test for BM: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BM observations from Dream (D) to determine that body mass 'mean' on the island is significantly different to observations from Biscoe (B).

$$\begin{aligned} &\mathrm{H}_0 : \mu_D = \mu_B \\ &\mathrm{H}_1 : \mu_D \neq \mu_B \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.15). The `t.test` function provides us with a p-value of 0.26. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

All two-sample t-tests are provided in `Appendix G`. These tests suggest that location does not appear to have have a significant impact on any of the Adelie penguins' physical characteristics.

Evalation

This report set out the exploratory data analysis of a sample of 100 penguins from the LTER study area surrounding Palmer station.

The penguin measurement variables BL, BD, FL and BM all approximate a normal probability distribution. The accuracy of these distributions is limited as the report has confirmed its findings with Q-Q plots. This report has not confirmed whether the distribution is symmetric about the 'mean' or that the 'mean' equals the 'median' for the sample set.

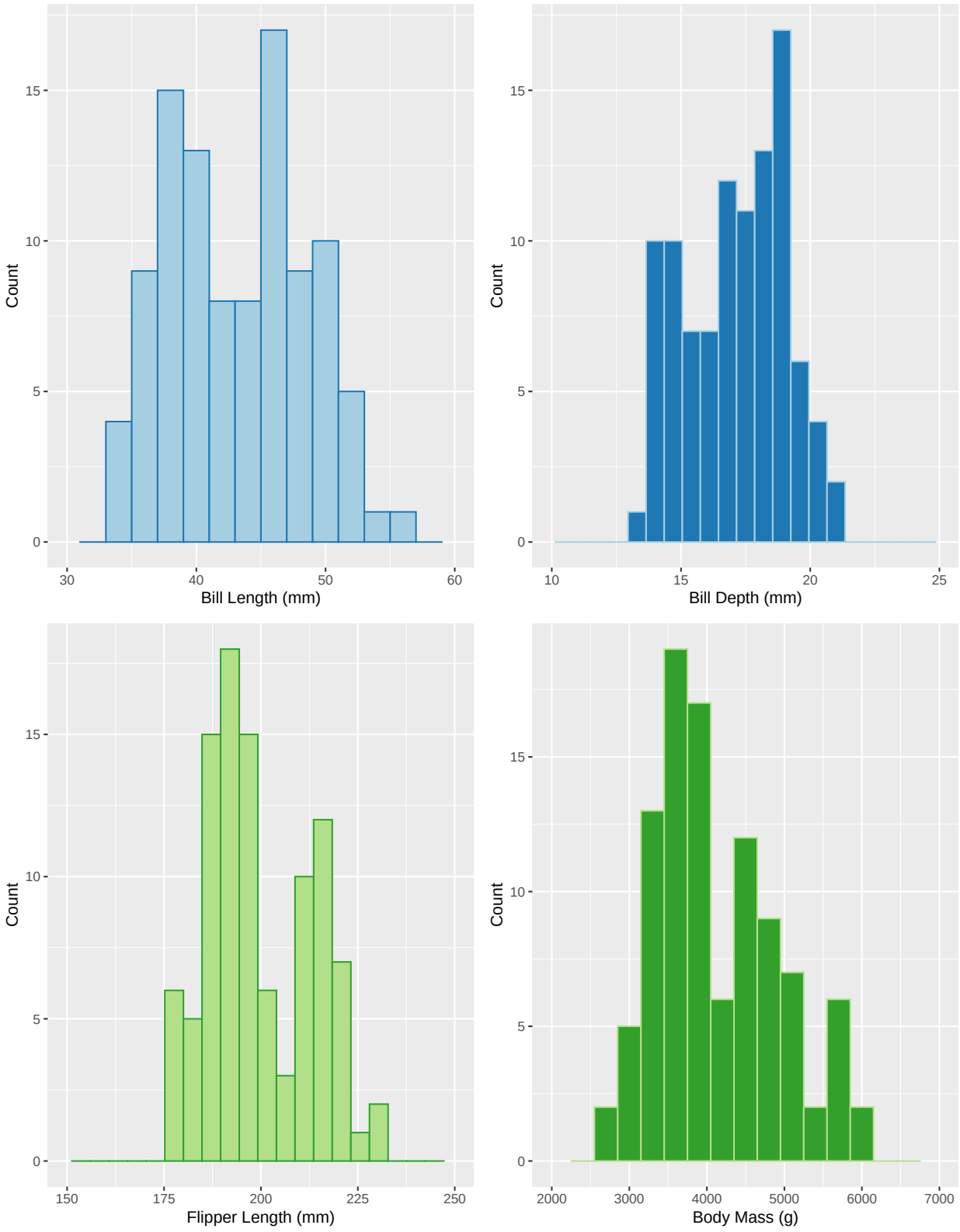
`Appendix F` sets out the estimates for the parameters of the BL, BD, FL and BM distributions with a 95% confidence interval. The accuracy of these estimates as estimators for the population is likely to be limited, especially for the chinstrap penguins.

The analysis highlights that BM is likely to best estimate the sex of Adelie and Gentoo penguins. For Chinstrap penguins, the BL appear to be best. Further investigation is required to identify if BM is likely to reliably estimate the sex of a penguin as the range of BM values for male and females overlap. For example, using both BM and BL measurement data may be more reliable.

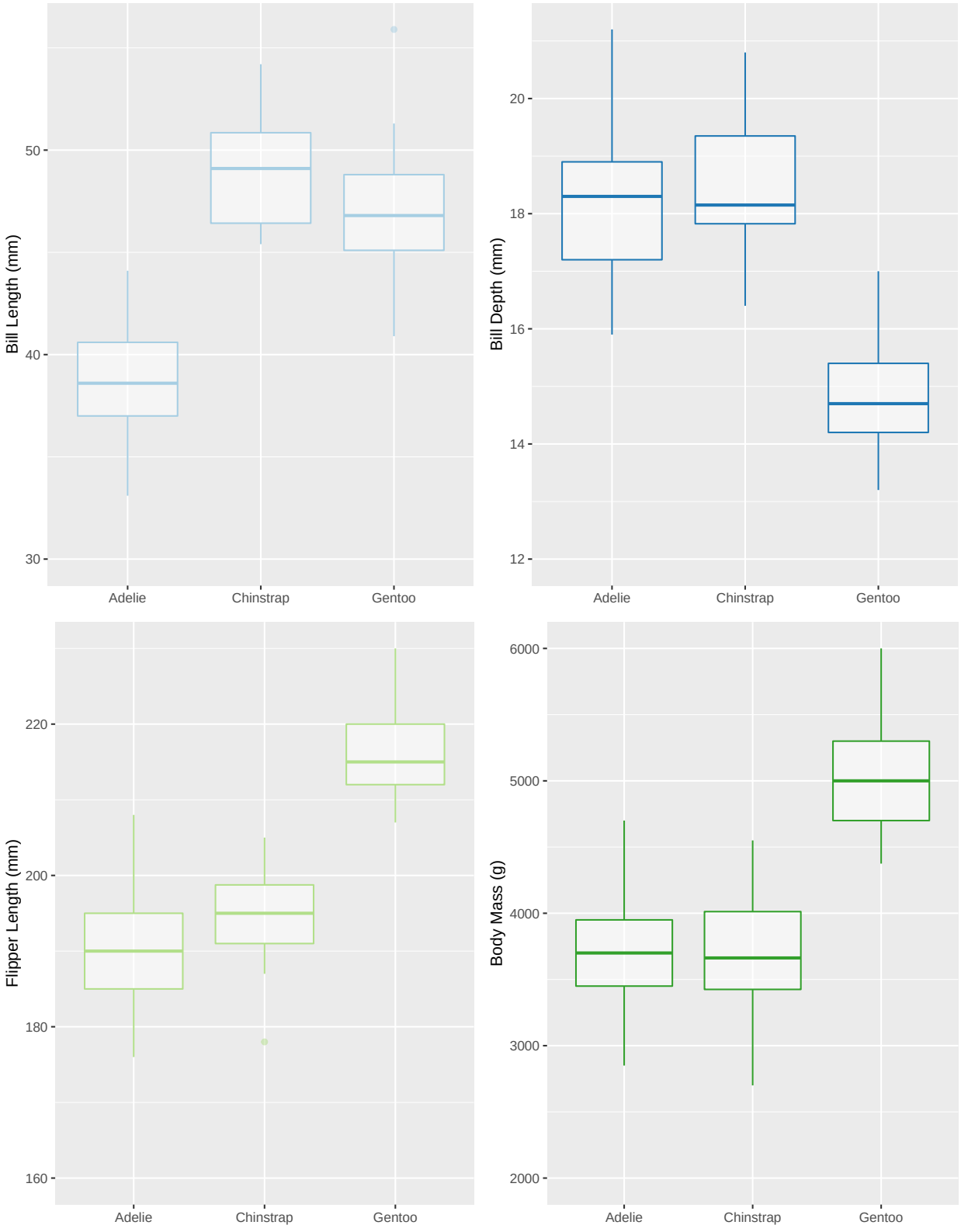
Adelie penguins are the only species found on all 3 islands, and the location does not appear to have a statistically significiant impact on its phsyical characteristics. Further investigation is needed to conclusively determine this; for example, exploring a larger data set filtered by year.

Overall, the insights and findings outlined would be improved with access to more sample sets of the penguins' data or a larger sample size.

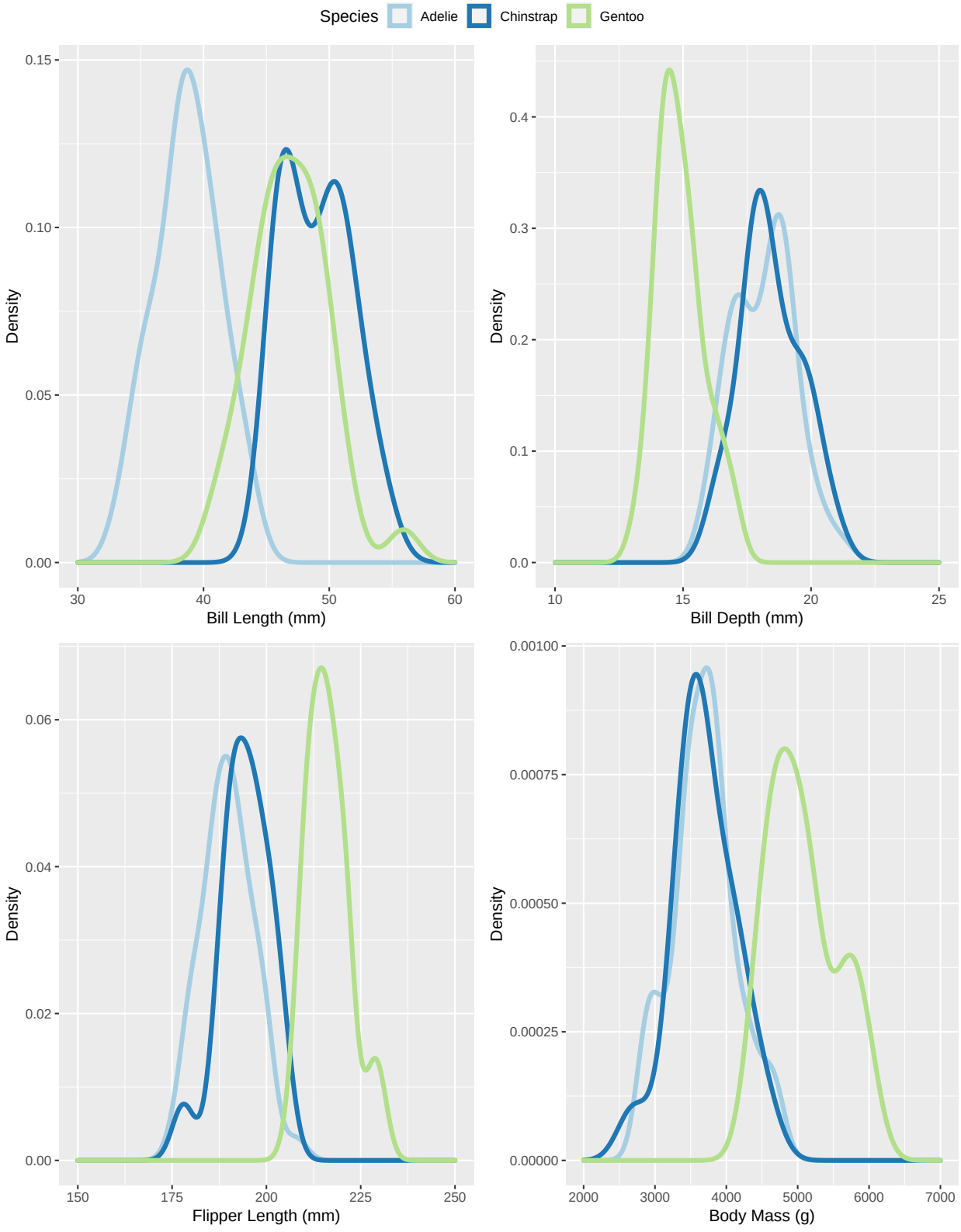
Appendix A: Histograms of Measurement Variables



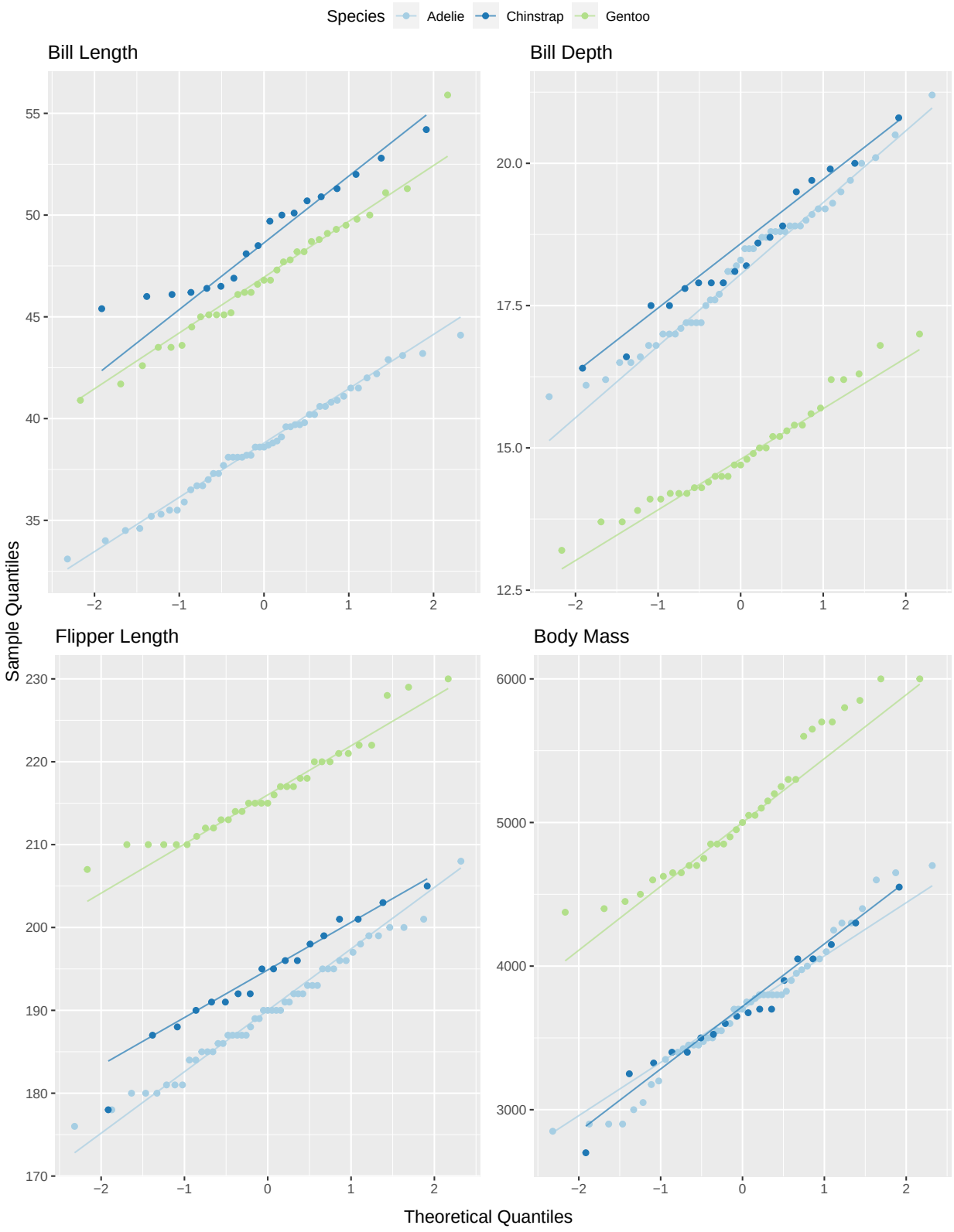
Appendix B: Boxplots of Measurement Variables by Species



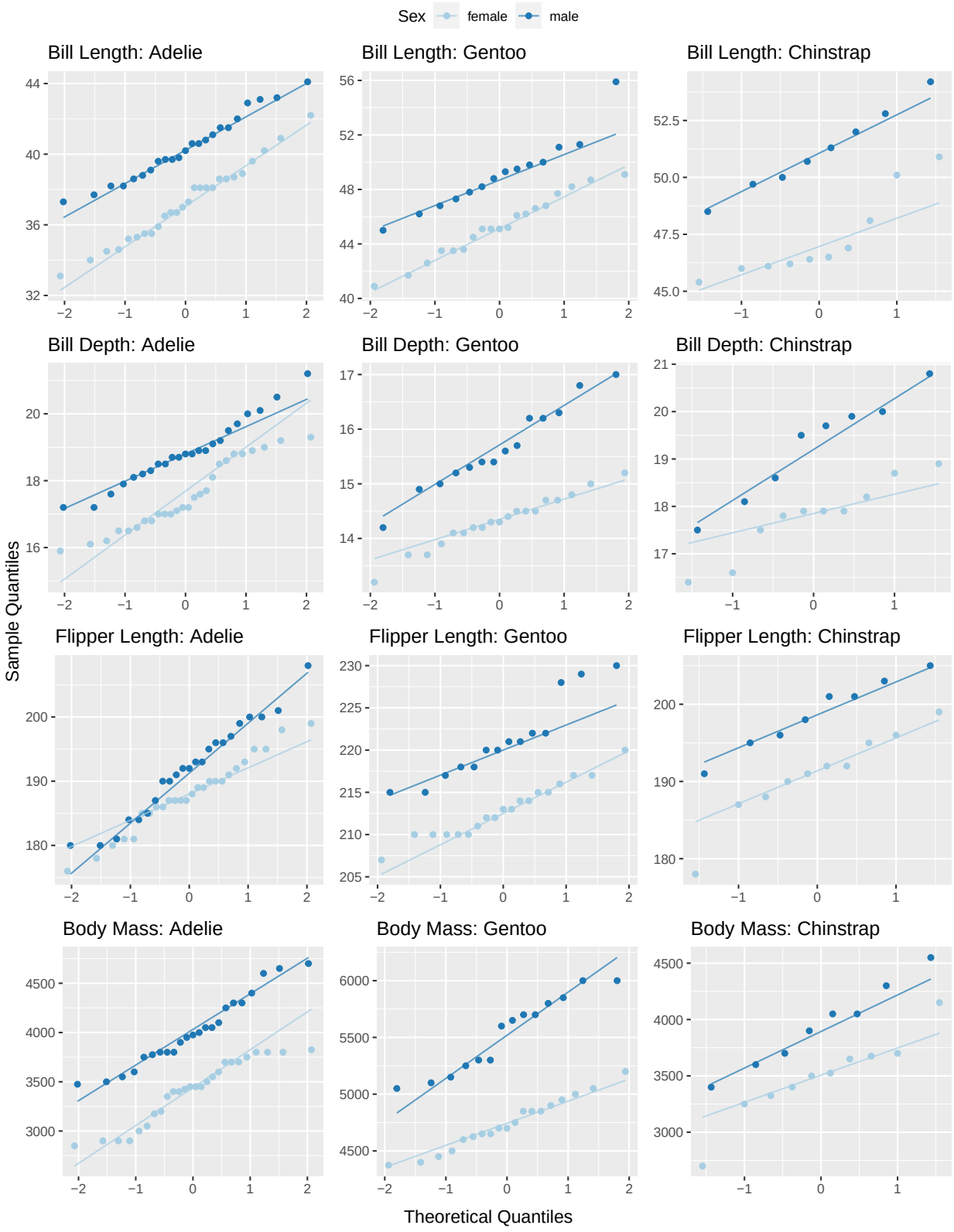
Appendix C: Density Distributions of Measurement Variables by Species



Appendix D: Q-Q Plots of Measurement Variables of Adelies and Gentoos



Appendix E: Q-Q Plots of Measurement Variables by Species



Appendix F: Distribution Parameter Estimates

Table 2: Bill Length Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	38.70	37.98-39.42	2.58
Gentoo	46.88	45.83-47.93	3.09
Chinstrap	48.99	48.08-49.90	2.68

Table 3: Bill Depth Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	18.15	17.81-18.49	1.21
Gentoo	14.88	14.57-15.19	0.91
Chinstrap	18.44	18.03-18.85	1.19

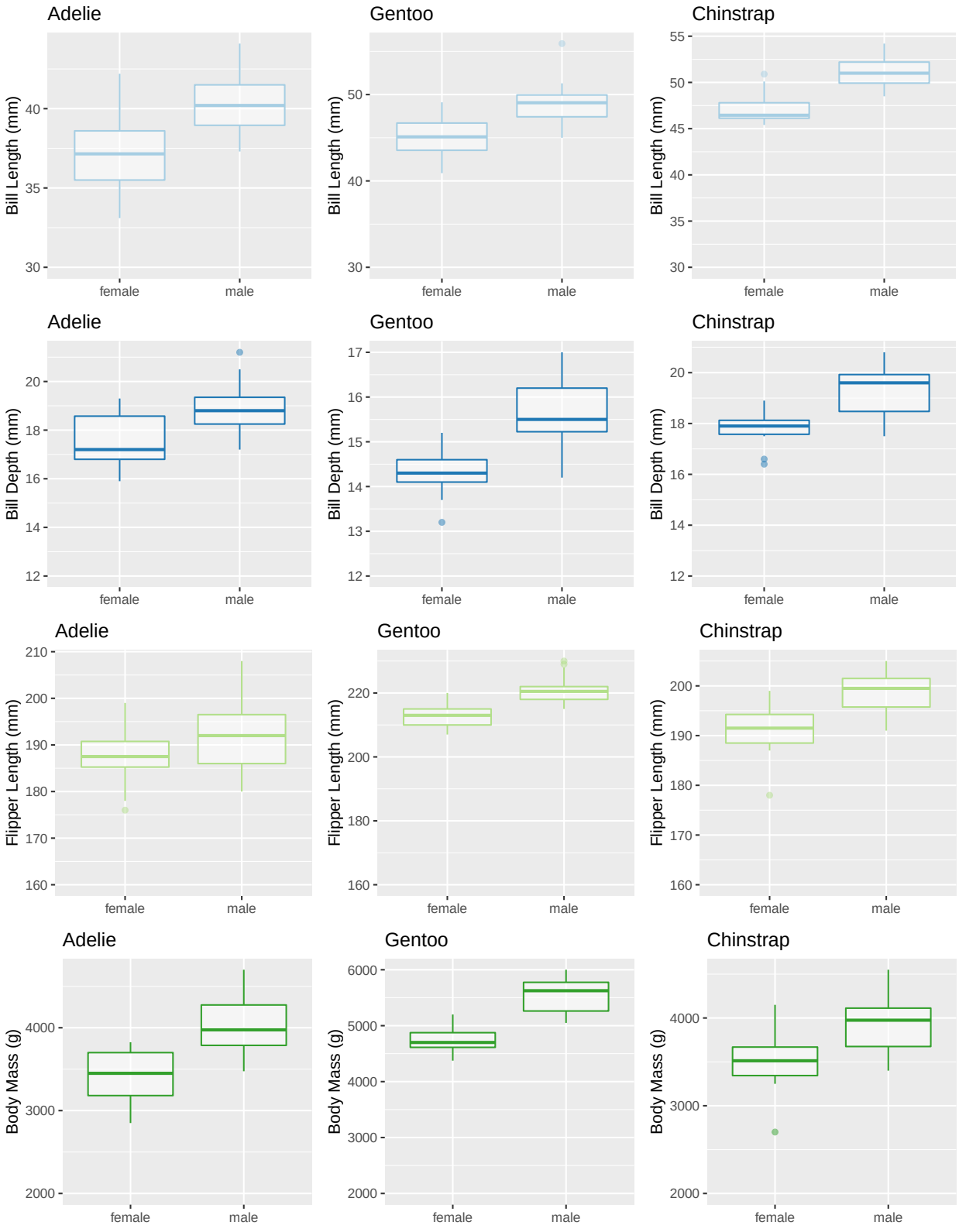
Table 4: Flipper Length Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	189.78	187.88-191.68	6.80
Gentoo	216.42	214.48-218.36	5.68
Chinstrap	194.33	192.06-196.60	6.65

Table 5: Body Mass Summary Statistics

Species	Mean (g)	95% CI (g)	SD
Adelie	3691.84	3565.18-3818.50	452.38
Gentoo	5075.76	4911.39-5240.13	481.76
Chinstrap	3690.28	3542.83-3837.73	432.17

Appendix G: Boxplots of Measurement Data by Penguin Sex



Appendix H: Two-sample t-tests for Penguin Sex

Adelie two-sample t-test for BL: Of the 49 Adelie penguins, there are 26 females and 23 males. We would like to compare BL observations from Adelie females (F) to determine that their 'mean' is different to observations from Adelie males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.42). The `t.test` function provides us with a p-value of 2.93e-6. This suggests that there is very strong evidence against H_0 and to reject it in favour of H_1.

Adelie two-sample t-test for BM: Of the 49 Adelie penguins, there are 26 females and 23 males. We would like to compare BL observations from Adelie females (F) to determine that their 'mean' is different to observations from Adelie males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.59). The `t.test` function provides us with a p-value of 1.21e-7. This suggests that there is very strong evidence against H_0 and reject it in favour of H_1.

Gentoo two-sample t-test for BL: Of the 33 Gentoo penguins, there are 19 females and 14 males. We would like to compare BL observations from Gentoo females (F) to determine that their 'mean' is different to observations from Gentoo males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is not valid (p-value = 0.56). The `t.test` function provides us with a p-value of 1.27e-4. This suggests that there is very strong evidence against H_0 and to reject it in favour of H_1.
Gentoo two-sample t-test for BM: Of the 49 Adelie penguins, there are 26 females and

23 males. We would like to compare BL observations from Adelie females (F) to determine that their 'mean' is different to observations from Adelie males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.14). The `t.test` function provides us with a p-value of 3.30e-9. This suggests that there is very strong evidence against H_0 and reject it in favour of H_1.

Chinstrap two-sample t-test for BL: Of the 18 Chinstrap penguins, there are 10 females and 8 males. We would like to compare BL observations from Chinstrap females (F) to determine that their 'mean' is different to observations from Chinstrap males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

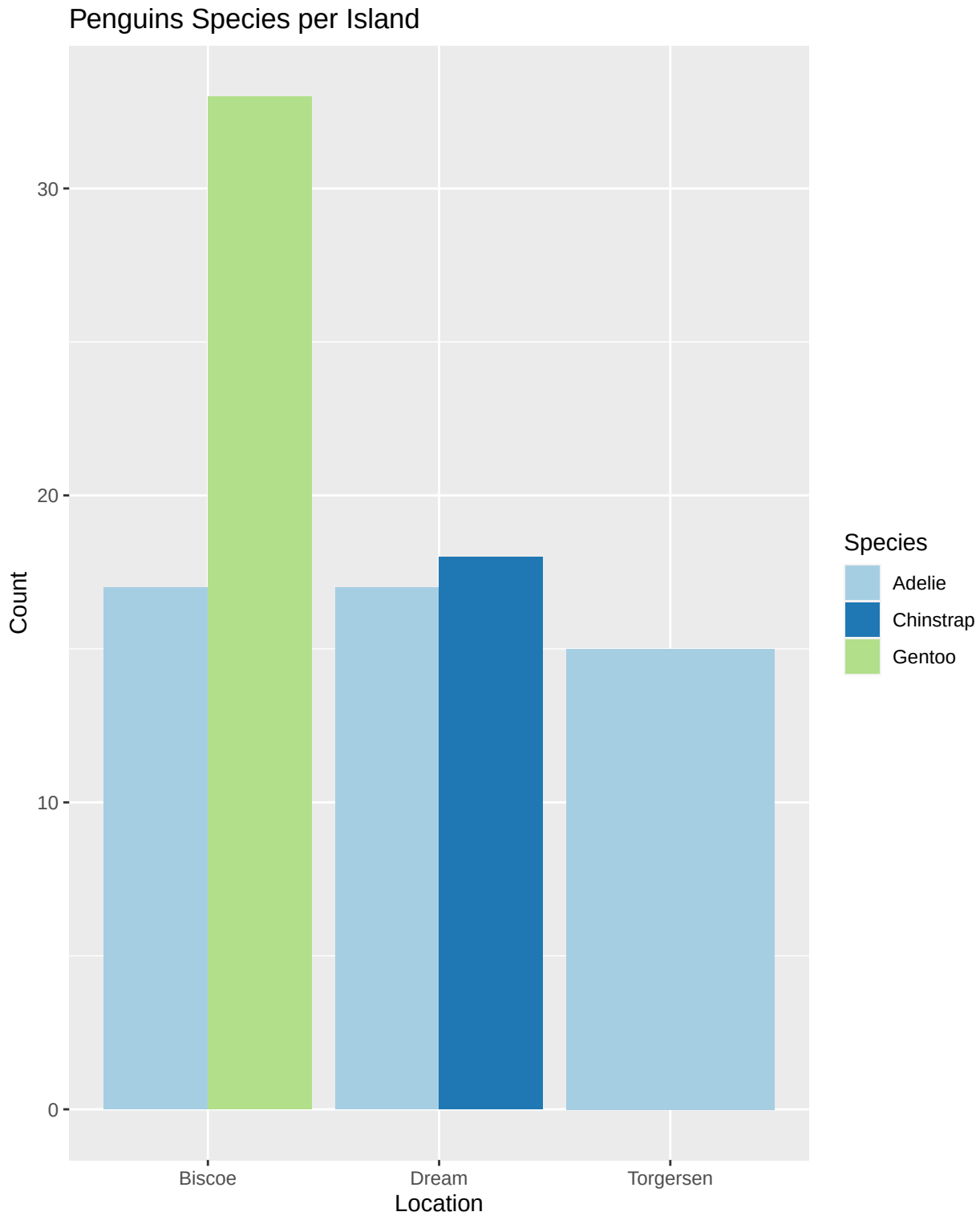
The `barlett.test` function for equal variances confirms that the assumption of equal variance is not valid (p-value = 0.97). The `t.test` function provides us with a p-value of 4.05e-4. This suggests that there is very strong evidence against H_0 and to reject it in favour of H_1.

Chinstrap two-sample t-test for BM: Of the 18 Adelie penguins, there are 10 females and 8 males. We would like to compare BL observations from Chinstrap females (F) to determine that their 'mean' is different to observations from Chinstrap males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.99). The `t.test` function provides us with a p-value of 2.08e-2. This suggests that there is very strong evidence against H_0 and reject it in favour of H_1.

Appendix I: Bar chart of Penguin Species Location



Appendix J: Two-sample t-tests for Adelie location

Dream and Biscoe two-sample t-test for BL: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BL observations from Dream (D) to determine that bill length 'mean' on the island is significantly different to observations from Biscoe (B).

$$\begin{aligned} &\mathrm{H}_0 : \mu_D = \mu_B \\ &\mathrm{H}_1 : \mu_D \neq \mu_B \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.21). The `t.test` function provides us with a p-value of 0.19. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Dream and Torgersen two-sample t-test for BL: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BL observations from Dream (D) to determine that bill length 'mean' on the island is significantly different to observations from Torgersen (T).

$$\begin{aligned} &\mathrm{H}_0 : \mu_D = \mu_T \\ &\mathrm{H}_1 : \mu_D \neq \mu_T \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.77). The `t.test` function provides us with a p-value of 0.51. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Torgersen and Biscoe two-sample t-test for BD: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BD observations from Torgersen (T) to determine that bill depth 'mean' on the island is significantly different to observations from Biscoe (B).

$$\begin{aligned} &\mathrm{H}_0 : \mu_T = \mu_B \\ &\mathrm{H}_1 : \mu_T \neq \mu_B \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.33). The `t.test` function provides us with a p-value of 0.63. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Torgersen and Dream two-sample t-test for BD: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BD observations from Torgersen (T) to determine that bill depth 'mean' on the island is significantly different to observations from Dream (D).

$H_0 : \mu_T = \mu_D$
$H_1 : \mu_T \neq \mu_D$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.33). The `t.test` function provides us with a p-value of 0.82. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Torgersen and Biscoe two-sample t-test for FL: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare FL observations from Torgersen (T) to determine that flipper length 'mean' on the island is significantly different to observations from Biscoe (B).

$H_0 : \mu_T = \mu_B$
$H_1 : \mu_T \neq \mu_B$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.77). The `t.test` function provides us with a p-value of 0.48. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Torgersen and Dream two-sample t-test for FL: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare FL observations from Torgersen (T) to determine that flipper length 'mean' on the island is significantly different to observations from Dream (D).

$H_0 : \mu_T = \mu_D$
$H_1 : \mu_T \neq \mu_D$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.41). The `t.test` function provides us with a p-value of 0.94. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Dream and Biscoe two-sample t-test for BM: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BM observations from Dream (D) to determine that body mass 'mean' on the island is significantly different to observations from Biscoe (B).

$H_0 : \mu_D = \mu_B$
$H_1 : \mu_D \neq \mu_B$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.15). The `t.test` function provides us with a p-value of 0.26. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Dream and Torgersen two-sample t-test for BM: Of the 49 Adelie penguins, 17 are on Biscoe, 17 are on Dream, and 15 are on Torgersen. We would like to compare BM observations from Dream (D) to determine that body mass 'mean' on the island is significantly different to observations from Torgersen (T).

$$\begin{aligned} &\mathrm{H}_0 : \mu_D = \mu_T \\ &\mathrm{H}_1 : \mu_D \neq \mu_T \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.56). The `t.test` function provides us with a p-value of 0.33. This suggests that there is no evidence against H_0 and it is not rejected in favour of H_1.

Bibliography [CHECK AS NOT ALL ARE APPEARING]

Bhandari, Pritha. 2022. "Normal Distribution | Examples, Formulas and Uses." https://www.scribbr.com/statistics/normal-distribution/.

Foundation, Nationa Science. n.d. "Palmer Station Webcams." https://www.usap.gov/videoclipsandmaps/palwebcam.cfm?t=1.

Hill, Horst and Gorman, Allison Horst, Alison Hill, and Krissten Gorman. 2020. "Release the Penguins." https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/.

LTER, Long Term Ecological Research Program. n.d. "Penguin Science." https://penguinscience.com/.

Newcastle University, Joe Matthews on behalf of. 2022. "Mas8403 1: Introduction."

Rutgers, The State University of New Jersey, and Palmer Station Antarctica LTER. 2022. "Palmer Station Antarctica LTER." https://pallter.marine.rutgers.edu/.