MAS8403

Harvey Yuan 0077439

2022-10-31

Aim and Objective

The aim of this analysis is to explore and see if there is any correlation between the which sex the penguins are from and their body mass index. This will hopefully allow researchers to understand what causes the difference in weight for the penguins such as environmental and genetic factors.

From this data, we hopefully will be able to use body mass to distinguish between male and female penguins without using invasive procedures. By using analysis methods which are non-invasive for sexing, less stress will be inflicted onto the penguins which will reduce the weight fluctuations of the penguins due to stress.

Technologies and Libraries Used

R version 4.1.2

R is a language used for statistical computing and graphics.

RStudio 2021.09.01 Build 372

RStudio is an IDE that is used for R. RStudio includes an in-built console and terminal that allows for direct code executing and also a pane which allows for files, plots and packages management. Different libraries can be installed through this pane.

dplyr version 1.0.7

The dplyr library allows for the user to use a set of verbs to write code for common data manipulation steps. It allows the user to use familiar words when scripting analysis and pre-processing scripts (such as filter), both for easy understanding of the code and proof reading (Hadley Wickham, n.d.).

readr version 2.0.1

The readr library provides for an easy way to read the csv files that are being used as part of this analysis. As well as being easier to for users to understand and read the data, it also makes analysis more reproducible as base R functions inherit behaviours from the operating system and environment variables, as such, importing code from one environment to another using readr will work without issue (Grolemund, 2016).

ggplot2 version 3.3.5

ggplot is a system for creating graphics (where the data is a selected data frame). The data is either provided by the user or created in script, which allows for said user to iteratively add new layers, components and functionality. ggplot will be used to demonstrate the analysis as part of this investigation.

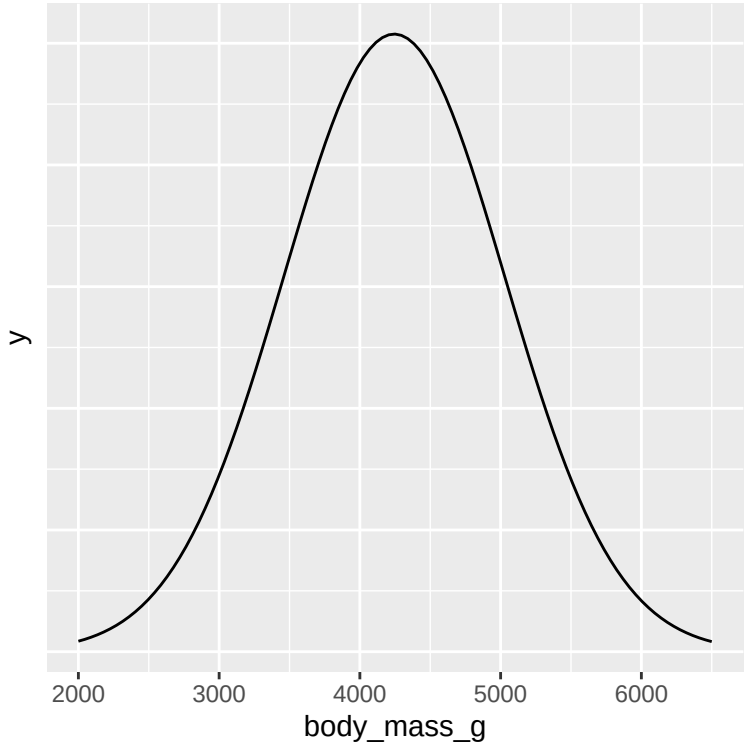
Data Preparation and Analysis

As part of the project, the prepared dataset which provided the data on the penguins was taken from the "palmerpenguins" package in R. The dataset contains data on 333 penguins with a large number of different variables but for the investigation, we're going to wrangle and analyse the following data:

- species - species of the penguin (Adelie, Chinstrap or Gentoo)
- island - island the penguin lives (Biscoe, Dream or Torgerson)
- body mass g - penguin's body mass (in grams)
- sex - sex of the penguin (male or female)

For the investigation, a sample of 100 penguins were used and the collected data of these penguins were analysed.

With the understanding of this data and the aim of this investigation, in the first instance, it was decided that the first thing to look at is the sample mean of the body masses of the penguins aggregated by species, island and sex. The aggregation of the data provides the sample mean which is a good estimator of the population mean.



By plotting the body mass, we can tell that the data is normally distributed. This means that the mass of the penguins distribute symmetrically around the mean mass of the penguins and that a majority of the penguins fall within the ± 1.96 standard deviations of the sample mean.

Table 1: Aggregated Mean Body Mass

	species	island	sex	body_mass_g
1	Adelie	Biscoe	female	3509.375
3	Adelie	Dream	female	3355.000
5	Adelie	Torgersen	female	3360.714
6	Adelie	Biscoe	male	4110.714
8	Adelie	Dream	male	4129.545
10	Adelie	Torgersen	male	4037.500
4	Chinstrap	Dream	female	3635.000
9	Chinstrap	Dream	male	3859.615
2	Gentoo	Biscoe	female	4766.667
7	Gentoo	Biscoe	male	5423.810

The above table shows the sample mean of the 100 penguins sorted by species.

The mean body mass of the female penguins is 4,219g whereas the mean body mass of the male penguins come to 3,818g. Comparing the body mass of the penguins by species reveals that the female Adelie penguins, on average, have 168g less body mass whereas for the other two species, the females have a higher body mass of 224g and 657g respectively for the Chinstrap and Gentoo species.

It is worth noting that for both the Chinstrap and Gentoo species of penguins, they only appear on one island each. The Chinstrap species only appear on the Dream island whereas the Gentoo species only appear on the Biscoe island. This is true for the population mean and by extension, also true for the sample that we have.

Table 2: Mean Body Mass by Species

species	sex	body_mass_g
Adelie	female	3408.363
Chinstrap	female	3635.000
Gentoo	female	4766.667
Adelie	male	4092.587
Chinstrap	male	3859.615
Gentoo	male	5423.810

In table 2, the data shows that for the Adelie species of penguin, the male penguins have a larger body mass regardless of which island they are from by approximately 685g.As for the Chinstrap and Gentoo penguins, they data also shows that the male penguins have a larger mean body mass than their female equivalents by approximately 224g for the Chinstrap and 657g for the Gentoo.

Table 3: Mean Body Mass by Island

island	sex	body_mass_g
Biscoe	female	4263.750
Dream	female	3541.667
Torgersen	female	3360.714
Biscoe	male	5095.536
Dream	male	3983.333
Torgersen	male	4037.500

This realisation is confirmed by table 3, as shown above. Table 3 shows the mean sample mass for the penguins aggregated by the islands they are from. This suggests that, on a whole, that the male population of penguins have a higher body mass than the female penguins.

By using the values from tables 1,2 and 3, we can calculate the confidence interval, by the means of the t distribution, of the population mean with 95% confidence that it would fall within the ranges of:

$3842.278477 < \mu$ (female) < 3868.435809

$4510.617446 < \mu$ (male) < 4541.106691

By calculating the 95% confidence interval of our estimator, the sample mean, we can say that the population mean for body mass will fall within the given ranges for both sexes 95% of the time.

From the data and analysis, it seems as though using the body mass of the penguins is a good way to distinguish the sexes of the penguins. This is assuming that the penguins are not abnormally large or abnormally small for the sex of the penguin. This means that for any one individual penguin, if their body mass lies within ± 1.96 standard deviations of the sample mean range as shown above, they are likely to fall within they sex of penguin.

To analyse if the island which the penguin is on has any significant impact on it's physical characteristic, we would need to discount the Chinstrap and Gentoo species because the Adelie species is the only observed species which are present across all three islands.

As part of the analysis to identify if the islands made any significant impact on any physical characteristic, the confidence internals were worked out of the sample body mass for the male and female penguins of the Adelie penguins. By calculating this, we would be 95% confident that the population body mass for the male and female penguins lie within the range. The ranges of the upper and lower mean of the Adelie penguins are as follows:

$3305.007179 < \mu$ (female) < 3532.492821

$3949.159351 < \mu$ (male) < 4252.923982

Comparing the mean body mass of the Adelie species across the islands, it doesn't look like the island affects the body mass as the means of the body mass across the species varies from the range of the confidence interval.

Table 4: Adelie by Sex and Island

species	island	sex	body_mass_g
Adelie	Biscoe	female	3509.375
Adelie	Dream	female	3355.000
Adelie	Torgersen	female	3360.714
Adelie	Biscoe	male	4110.714
Adelie	Dream	male	4129.545
Adelie	Torgersen	male	4037.500

From table 4, all body mass sample means for the Adelie penguins fall within the population mean calculated using the confidence intervals. This suggests that which island the Adelie penguins are on doesn't affect the body mass.

Critical Evaluation

The aim of this investigation was to determine if there is any correlation between which sex the penguins are and their body mass index and hopefully be able to determine the sex of the penguins using non-invasive procedure.

It was chosen that the body mass would be used as part of the physical trait to evaluate as generally in nature, male versions of the same species are larger than their female counterparts. The data was plotted to show that the body mass of the sample was normally distributed, which allowed the for the population mean to be approximated to fall within the ranges 95% of the time.

The body mass of the penguins where aggregated by species, island and sex in the first instance which showed that only the Adelie penguin were present on all three islands. The body mass of the penguins were also evaluated against their species and sex. This showed that the across all three species, the male penguins were on average approximately 520g larger than the female penguins.

Lastly, as the Adelie penguins were the only penguins to span across all three islands, they were the only penguins used in the investigation for if which island penguins come from have any significant impact on any characteristic. This meant that the Chinstrap and Gentoo penguins did not affect the sample mean of the body mass and skew it to negatively impact the Dream and Biscoe islands.

To improve this investigation in the future, the body mass of an individual penguin should be evaluated against the calculated figures above and the probability of the sex of the penguin could be determined.

Another way to improve this investigation would be to evaluate the body mass in conjunction with another physical characteristic such as bill depth to allow for more accurate sexing of the penguins. By doing so, the probablity fitting both criteria of male and female characteristics are smaller if the more traits that are compared.

Bibliography

Hadley Wickham, R. F. (n.d.). dplyr. Retrieved from dplyr part of the tidyverse 1.0.7: https://dplyr.tidyverse.org/ (Last accessed 1st of November 2022)

Grolemund, H. W. (2016, December). Data Import. Retrieved from R for Data Science: https://r4ds.had.co.nz/data-import.html (Last accessed 1st of November 2022)