Data Analysis of Palmer Penguins using Statistical Methods

2024-10-18

Amey Madane
Student No.: 240698836 Newcastle University, Newcastle upon Tyne
Project Supervisor: Dr. Clement Lee

Introduction

The current research project would be applying different statistical methods to analyze the given Palmer Archipelago penguin dataset of Antarctica, focusing more on the investigation of the association between penguin populations on different islands and their gender in connection with various features of the dataset. These skills include exploratory data analyses, estimation of population proportion, and testing of hypotheses, which provide the necessary tools for scientists to make meaningful inferences and predictive assessments.

Methodology

```
library(ggplot2)
library("GGally")
library(vegan)
library(gridExtra)
library(palmerpenguins)
data("penguins")
penguins = na.omit(penguins) # Removes missing rows

n <- 200
my.student.number <- 240698836 #student number
set.seed(my.student.number)
my.penguins <- penguins[sample(nrow(penguins), n), ]
```

Summary of my.penguins

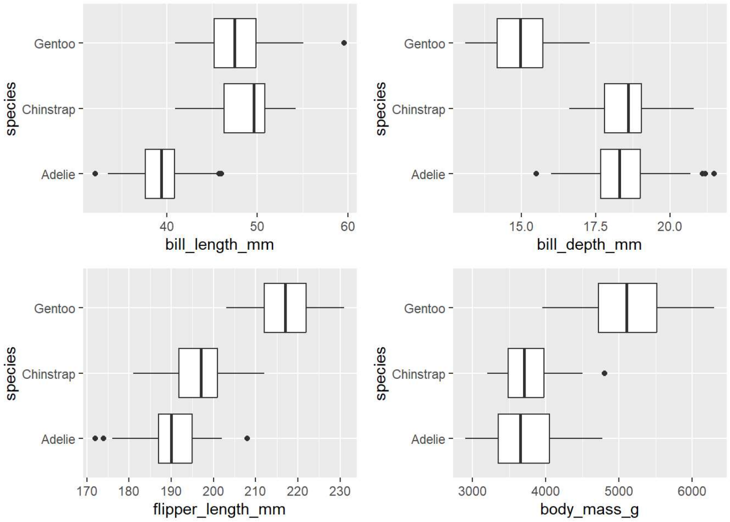
```
summary(my.penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm  
##  Adelie   :84   Biscoe   :101   Min.   :32.10   Min.   :13.10  
##  Chinstrap:40   Dream    : 69   1st Qu.:39.70   1st Qu.:15.47  
##  Gentoo   :76   Torgersen: 30   Median :45.05   Median :17.30  
##                                 Mean   :44.26   Mean   :17.09  
##                                 3rd Qu.:49.00   3rd Qu.:18.62  
##                                 Max.   :59.60   Max.   :21.50  
##  flipper_length_mm  body_mass_g       sex          year     
##  Min.   :172.0     Min.   :2900   female: 95   Min.   :2007  
##  1st Qu.:190.0     1st Qu.:3550   male  :105   1st Qu.:2007  
##  Median :198.0     Median :4150                Median :2008  
##  Mean   :201.9     Mean   :4262                Mean   :2008  
##  3rd Qu.:214.0     3rd Qu.:4881                3rd Qu.:2009  
##  Max.   :231.0     Max.   :6300                Max.   :2009  
```

The penguin dataset consists of a diverse sample of 200 penguins across three species and three islands. Adelie penguins and individuals from Biscoe Island are slight predominations. The sex ratio is roughly balanced with a small majority of males. Physical measurements vary: bill lengths range from 32.1 mm to 59.6 mm, bill depth varies from 13.1 mm to 21.5 mm, flipper length varies between 172 mm and 231 mm, while body mass varies between 2900 g and 6300 g. The mean of 4262 g being higher than the median of 4150 g indicates right-skewness, hence the presence of some very heavy birds. Of all the measures, flipper length is the closest approximation to normality. This dataset ranges over three years, 2007-2009, and gives an overview of the penguin population within these three years. The previous summary statistics expose the variability of the sample and give a background on what further analysis would disclose with respect to the species, island location, and sex as influencing variables on penguin morphology.

1. Exploratory Data Analysis

```
gg1 = ggplot(my.penguins, mapping = aes(bill_length_mm, species)) + geom_boxplot()
gg2 = ggplot(my.penguins, mapping = aes(bill_depth_mm, species)) + geom_boxplot()
gg3 = ggplot(my.penguins, mapping = aes(flipper_length_mm, species)) + geom_boxplot()
gg4 = ggplot(my.penguins, mapping = aes(body_mass_g, species)) + geom_boxplot()
grid.arrange(gg1, gg2, gg3, gg4)
```



The box plots illustrate that each of the three penguin species has entirely different morphologies. Generally, Gentoo penguins possess much longer bills and larger flippers; their body mass is higher compared to Adelie and Chinstrap penguins. In terms of bill depth, it can be observed that Chinstrap penguins have the biggest depth, whereas generally, Adelie has the shortest and shallowest bill. These species have the lowest body mass and flipper length among the three species. Interestingly, while the Gentoo penguins exhibit the greatest overall size, they also show more variability in bill length and body mass, as evidenced by the wider boxes and longer whiskers in these plots. This plot really shows that each species has a unique combination of physical characteristics, which may be related to their particular adaptations to different ecological niches in the Antarctic environment.

```
table = xtabs(~ year + species + island, data =my.penguins)
table
```

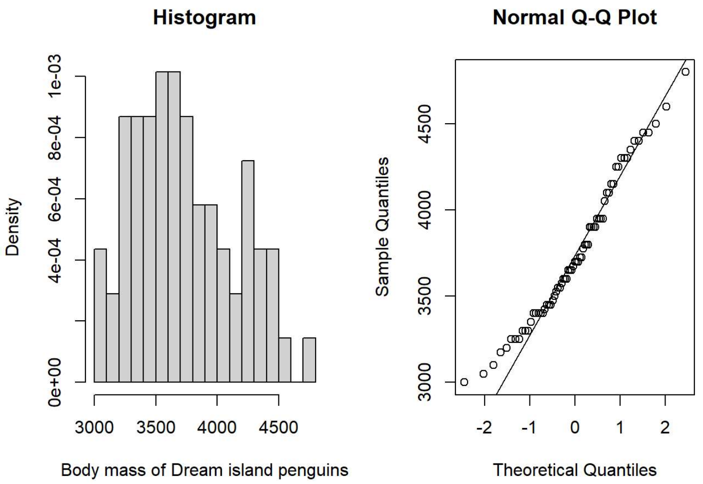
```
## , , island = Biscoe
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007      4         0     23
##   2008      8         0     28
##   2009     13         0     25
## 
## , , island = Dream
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007     12        15      0
##   2008      6        11      0
##   2009     11        14      0
## 
## , , island = Torgersen
## 
##       species
## year   Adelie Chinstrap Gentoo
##   2007     10         0      0
##   2008      9         0      0
##   2009     11         0      0
```

It is observed that Adelie species can be found on every island, Gentoo species are only found on Biscoe island and Chinstrap could be found on Dream island.
The number of penguins on Biscoe island have increased over the years.

2. Estimating probability/proportions for Penguin population.

The data sample is used to estimate the parameters of body mass for the population of penguins inhabiting Dream Island.

```
par(mfrow = c(1, 2))
hist = hist(my.penguins$body_mass_g[my.penguins$island=='Dream'],
            breaks = 20, freq = FALSE, xlab = "Body mass of Dream island penguins", main = "Histogram")
qq = qqnorm(my.penguins$body_mass_g[my.penguins$island == "Dream"])
qqline(my.penguins$body_mass_g[my.penguins$island == "Dream"])
```



The histogram reveals a near-normal distribution of body mass for Dream Island’s penguins. To delve deeper, a normal Q-Q plot was generated, displaying slight deviations at the tails, yet the majority of data points closely align with the expected line. To find the mean and standard deviation of the sample we use most likelihood estimation.

Equation for normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Calculating log likelihood function

$$\log(f(x)) = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$\log(f(x)) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

Differentiate the log-likelihood function with respect to μ

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

Differentiate the log-likelihood function with respect to σ

$$\frac{\partial}{\partial\sigma}\log(\mathcal{L}(\mu,\sigma)) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

Setting the result to 0 and solving for MLE of μ and σ Step 1: Set the derivative with respect to μ to zero:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

Step 2: Solve for μ:

$$\sum_{i=1}^{n}(x_i-\mu)=0$$

Step 3: Rearrange the equation and solve for μ:

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Step 4: Now, set the derivative with respect to σ to zero:

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

Step 5: Solve for σ:

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=\frac{n}{\sigma}$$

Step 6: Rearrange the equation and solve for σ:

$$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

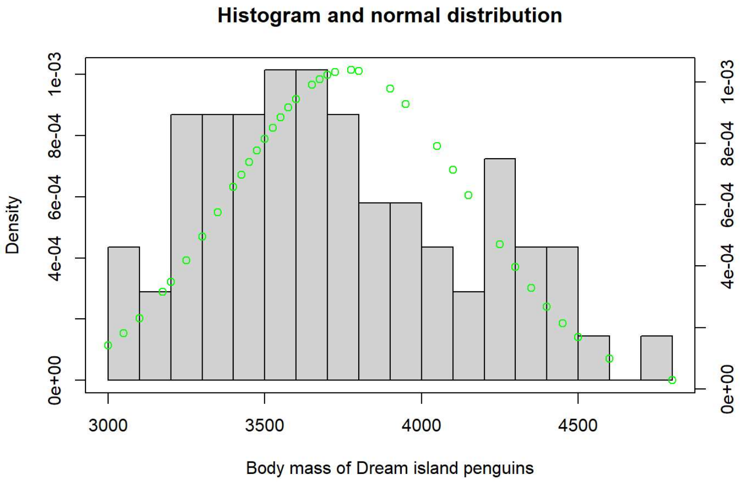
further we can compute the parameters in R using dnorm function

```r
l0 = function(param, x) {mu = param[1]; sigma = param[2]; -sum(dnorm(x, mu, sigma, log = TRUE))}

x = optim(c(0, 200), l0, x = my.penguins$body_mass_g[my.penguins$island == "Dream"])
x[1]
```

```
## $par
## [1] 3752.8363  415.6296
```

```r
myhist = hist(my.penguins$body_mass_g[my.penguins$island =='Dream'], breaks = 20,
              freq = FALSE, xlab = "Body mass of Dream island penguins",
              main = "Histogram and normal distribution")
par(new = TRUE)
plot(x = my.penguins$body_mass_g[my.penguins$island == "Dream"],
     dnorm(x = my.penguins$body_mass_g[my.penguins$island == "Dream"],
     mean = 3767.1958, sd=383.8225, log = FALSE), col = 'green',
     xlab = "", ylab = "", xaxt = 'n', yaxt = 'n')
axis(4)
```



The computed mean is 3767.1958 and standard deviation is 383.8225.

Following the construction of a normal distribution graph using the Maximum Likelihood Estimates for the mean and standard deviation over the histogram, it becomes evident that the body mass of penguins residing on Dream Island conforms to a normal distribution. This observation is further affirmed by the Q-Q plot.

When the model is accurately assumed, the Maximum Likelihood Estimator (MLE) stands out as the most efficient estimator, as it leverages the available data optimally to estimate the model parameters. This makes it a preferred choice for a wide range of applications, even in situations where assumptions of other models are not met. Moreover, in larger samples, the MLE tends to yield unbiased estimates, further enhancing its reliability and utility in statistical analysis.

A major disadvantage of MLE is that it tends to be biased for small samples of data.

To find the parameters of other measurement variables we need to first convert the data to a normal distribution by applying appropriate transformations.

Calculating the confidence interval of mean

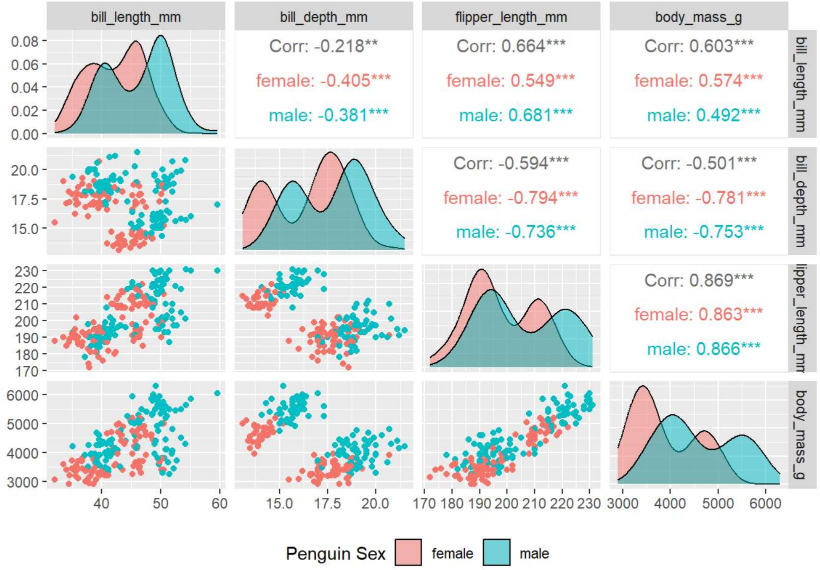
```
# 95% confidence interval for mean
#input sample size, mean and SD
n = 69
xbar = 3767.1958
sd = 383.8225
#calculating margin of error
margin <- qt(0.975,df=n-1)*sd/sqrt(n)
lower_bound <- xbar - margin
upper_bound <- xbar + margin
print(paste("mean values of body mass for penguins from Dream island lie in the range of",
            round(upper_bound, digits = 1), "and", round(lower_bound,digits = 1)))
```

```
## [1] "mean values of body mass for penguins from Dream island lie in the range of 3859.4 and 3
675"
```

3. To identify the variables that are useful for determining sex of

the penguin.

```
ggpairs(my.penguins, columns = c(3:6), aes(color = sex), legend = 1,
        diag = list(continuous = wrap("densityDiag", alpha = 0.5 )), progress = FALSE) +
        theme(legend.position = "bottom") + labs(fill = "Penguin Sex")
```



By showing side-by-side plots on the same scale we can see directly compare the distribution of the measurement variables for both genders. From this above plot, it can be concluded that male penguins have longer flippers and greater body masses. All continuous variables are asymmetrically distributed across the genders and are multi-modal in nature.

Hypothesis test - Two Sample t-test

```
t.test(body_mass_g ~ sex, data = my.penguins)
```

```
##
##  Welch Two Sample t-test
##
## data:  body_mass_g by sex
## t = -6.9952, df = 196.3, p-value = 4.069e-11
## alternative hypothesis: true difference in means between group female and group male is not e
qual to 0
## 95 percent confidence interval:
##  -939.1137 -526.0492
## sample estimates:
## mean in group female   mean in group male
##             3877.895             4610.476
```

Test Statistic: t = -6.9952 The negative value indicates that the first group (females) has a lower mean than the second group (males)

Degrees of Freedom: df = 196.3 This indicates the sample size and variability in the data.

P-value: p-value = 4.069e-11 This is an extremely small number (0.00000000004069). This is much smaller than the typical significance level of 0.05.

Confidence Interval: [-939.1137 and -526.0492] This range represents the 95% confidence interval for the true difference in mean body mass between females and males.

Sample Estimates: Mean for body mass females:3877.895 grams Mean body mass for males: 4610.476 grams

The very low p-value (4.069e-11) indicates that the difference in body mass between male and female penguins is highly statistically significant. Males are consistently heavier than females. The average male penguin (4610.476 g) is about 732.581 grams heavier than the average female penguin (3877.895 g).The 95% confidence interval [-939.1137, -526.0492] tells us that we can be 95% confident that the true population difference in mean body mass (female minus male) falls within this range.The large t-statistic (-6.9952) and the tight confidence interval suggest that this difference is consistent across the sample, not just driven by a few outliers Finally from this test it concludes that MALES ARE MORE HEAVIER THEN FEMALES.

ANOSIM test

ANOSIM test lets us compare a categorical variable with more than two groups to more than one numerical variable at a time.
Null Hypothesis (H0): There is no significant difference between numerical variable and different sexes. Alternative Hypothesis (H1): There is a significant difference between numerical variable and different sexes.

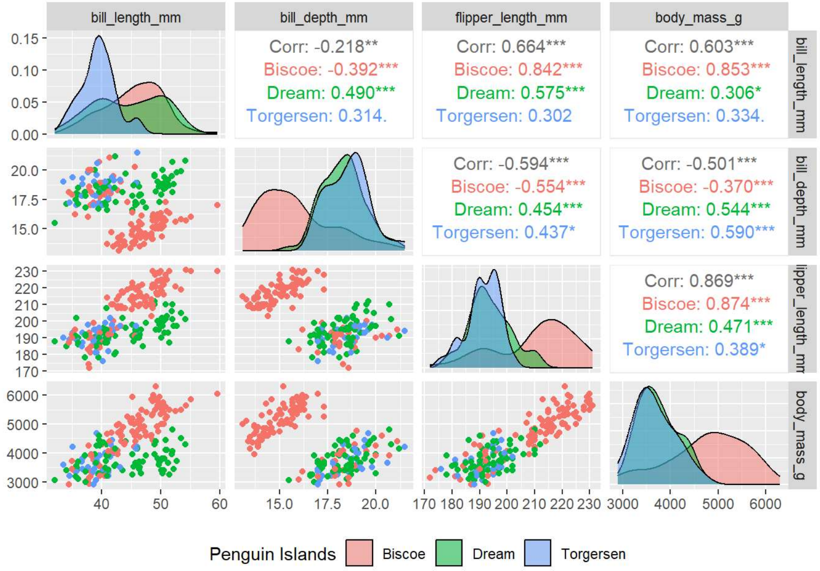
```
a = anosim(my.penguins[3:6], my.penguins$sex, permutations = 999, distance = "bray", strata = NU
LL,
    parallel = getOption("mc.cores"))
print(paste("significance value is",a[2]))
```

```
## [1] "significance value is 0.001"
```

Since the significance value(p-value) is less than 0.05 we can reject the Null hypothesis. a small p-value suggests that there is a significant dissimilarity between groups, indicating that the groups are more different from each other than would be expected by chance.

4. Significant difference in the physical characteristics of penguins living on different islands

```
ggpairs(my.penguins, columns = c(3:6), aes(color = island), legend = 1,
        diag = list(continuous = wrap("densityDiag", alpha = 0.5 )),
        progress = FALSE) + theme(legend.position = "bottom") +
        labs(fill = "Penguin Islands")
```



Using this plot we can visually explore the correlations and distributions of all the continuous variables for penguins living on different islands.
By analyzing the scatter plots we can see three distinctive groups.
We can observe that body mass of penguins from dream island seem to have a normally distributed data. This will be analyzed further.
Penguins living on Biscoe island have longer flipper length and high body mass but lowest bill depth.
Bill length and depth have gotten a weak to moderate correlation.
Body mass has a strong positive correlation with flipper length but weak correlation with bill depth.

Hypothesis test - ANOVA (Analysis of Variance) test

```
aov(body_mass_g ~ island, data = my.penguins)
```

```
## Call:
##    aov(formula = body_mass_g ~ island, data = my.penguins)
##
## Terms:
##                   island Residuals
## Sum of Squares  55637561  81547439
## Deg. of Freedom        2       197
##
## Residual standard error: 643.3867
## Estimated effects may be unbalanced
```

Sum of Squares: Island: 55,637,561 Residuals: 81,547,439 The Sum of Squares for ‘island’ represents the variation in body mass that can be explained by differences between islands.

Degrees of Freedom: Island: 2 (This is because there are three islands, so 3 - 1 = 2) Residuals: 197

Residual standard error: 643.3867 grams The residual standard error of about 643.39 grams indicates the typical amount of variation in body mass that isn’t explained by the island factor.

The Sum of Squares for ‘island’ (55,637,561) is substantial compared to the Residuals (81,547,439). This suggests that there are notable differences in body mass between islands, although there’s also considerable variation within each island.With 2 degrees of freedom for the island factor, we know we’re comparing three different islands.The total sample size is 200 (197 + 3), which matches your earlier mention of a 200-penguin subset. This ANOVA result suggests that there are likely meaningful differences in penguin body mass between the three islands this finding could have important implications for understanding how local environments shape penguin physiology and for conservation efforts that might need to consider island-specific factors affecting penguin health and survival.

Conclusion

Successfully applied various statistical methods to analyze the dataset. The exploratory data analysis provided valuable insights into the relationships between different variables. Maximum likelihood estimation allowed us to determine the mean and standard deviation of specific variables. Several non-parametric tests were conducted to investigate how different variables relate to the sex and island of the penguins.

The results indicate that certain physical characteristics, including bill length, bill depth, flipper length, and body mass, exhibit a higher correlation with the sex of the penguins. This suggests that these variables can be utilized by scientists for sex determination in penguins. However, our chi-squared test suggests that there is no statistically significant relationship between island/species and penguin sex. Additionally, the year of measurement does not appear to be related to the sex of the penguin. Hence island, species and year would not be useful for sex determination.

Based on the results of our exploratory data analysis and the ANOVA (Analysis of Variance) test, we may conclude that males are more heavier then females and we understand the influence of local environments on penguin physiology and informing conservation efforts by considering island-specific factors that affect penguin health and survival also there are probable significant differences in penguin body mass across the three islands studied.