Palmer Penguins: Statystyczna analiza danych

408453, Łukasz Wala, poniedziałek 14^{40}
AGH, Wydział Informatyki, Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2021/2022

Kraków, 26 stycznia 2022

Ja, niżej podpisany własnoręcznym podpisem deklaruję, że przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....

1 Streszczenie raportu

Raport powstał w oparciu o analizę cech 344 pingwinów zamieszkujących archipelag Palmera zebranych w latach 2007-2009.

2 Opis danych

Dane do projektu pochodzą z strony projektu `palmerpenguins`. Jest to projekt open-source mający dostarczyć dataset do wizualizacji i eksploracji danych będący alternatywą dla *Iris Dataset*, klasyka pośród materiałów do nauki statystycznej analizy danych i uczenia maszynowego.

W skład projektu wchodzą dwa zestawy danych: *penguins_raw* oraz jego uproszczona wersja *penguins*. W tym projekcie użyta zostanie wersja *penguins*. Dane zawierają 344 rekordy, gdzie każdy odpowiada innemu pingwinowi opisanemu ośmioma atrybutami:

- **species** - gatunek pingwina (zmienna jakościowa spośród *Adelie*, *Chinstrap*, *Gentoo*),
- **island** - wyspa w archipelagu Palmera (zmienna jakościowa spośród *Biscoe*, *Dream*, *Torgersen*),
- **bill_length_mm** - długość dzioba (liczba zmiennoprzecinkowa, w milimetrach),
- **bill_depth_mm** - głębokość dzioba (liczba zmiennoprzecinkowa, w milimetrach),

- **flipper_length_mm** - długość płetwy/skrzydła (liczba całkowita, w milimetrach),
- **body_mass_g** - masa ciała (liczba całkowita, w gramach),
- **sex** - płeć (zmienna jakościowa spośród *male*, *female*),
- **year** - rok obserwacji (liczba całkowita, z zamiarem zmiany na zmienną jakościową)

Do załadowania ich do środowiska R można podejść na dwa sposoby. Z racji tego, że jest to zestaw danych przygotowanych z myślą o edukacji, został spakowany do pakietu R, który można zainstalować poprzez CRAN, a następnie załadować:

```
> install.packages("palmerpenguins",
+                  repos = "https://cloud.r-project.org/")
> library(palmerpenguins)
> data(package = 'palmerpenguins')
> penguins_data <- penguins
```

Alternatywnym podejściem będzie pobranie pliku *.csv* z `repozytorium projektu` i załadowanie go:

```
> penguins_data <- read.csv(file = "penguins.csv",
+                           header = TRUE,
+                           stringsAsFactors = TRUE)
```

Czyszczenie danych sprowadzi się tylko do usunięcia rekordów z wartościami nieznanymi, które są numeryczne (są tylko dwa takie wiersze i nie zawierają one żadnego z pomiarów, więc nie poskutkuje to dużą utratą informacji). W przypadku nieznanych zmiennych jakościowych (tylko płci pingwinów w tym przypadku), jedną z metod byłoby zastąpienie ich losowymi wartościami, jednak ze względu na potencjalnie dużą korelację płci z pozostałymi cechami, nie jest to najlepsze rozwiązanie. Najlepszą metodą byłoby zbadanie zależności cech i płci, i na tej podstawie uzupełnienie danych, jednak, dla uproszczenia oraz ze względu na niewielką ilość takich wierszy, tutaj zostaną usunięte. Inną niewielką zmianą będzie będzie konwersja typu roku obserwacji z liczby na zmienną jakościową, w tym przypadku ma to więcej sensu, bo bardziej interesującymi informacjami są cechy pingwinów w danym roku niż średnia lat wykonywania pomiarów:

```
> penguins_data <- na.omit(penguins_data)
> penguins_data$year <- as.factor(penguins_data$year)
```

Wówczas dane wyglądać będą następująco:

```
> str(penguins_data)
```

```
'data.frame':        333 obs. of  8 variables:
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num  39.1 39.5 40.3 36.7 39.3 38.9 39.2 41.1 38.6 34.6 ...
 $ bill_depth_mm    : num  18.7 17.4 18 19.3 20.6 17.8 19.6 17.6 21.2 21.1 ...
 $ flipper_length_mm: int  181 186 195 193 190 181 195 182 191 198 ...
 $ body_mass_g      : int  3750 3800 3250 3450 3650 3625 4675 3200 3800 4400 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 2 1 2 2 ...
 $ year             : Factor w/ 3 levels "2007","2008",..: 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:11] 4 9 10 11 12 48 179 219 257 269 ...
  ..- attr(*, "names")= chr [1:11] "4" "9" "10" "11" ...
```

3 Analiza danych

3.1 Wydobywanie podstawowych informacji z danych

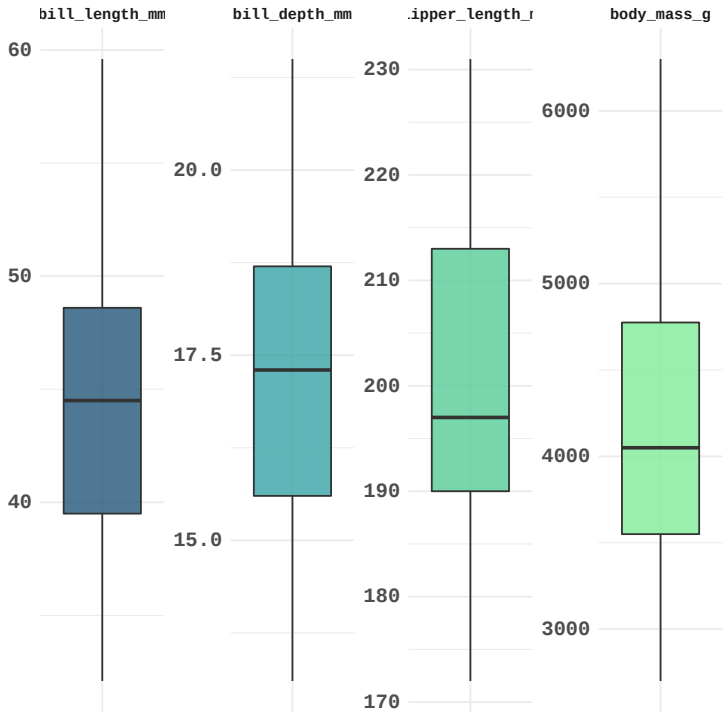
Na początku zostaną obliczone wartości estymatorów punktowych podstawowych wielkości statystyki opisowej:

```
> summary(penguins_data)
```

```
      species          island    bill_length_mm  bill_depth_mm
 Adelie   :146   Biscoe   :163   Min.   :32.10   Min.   :13.10
 Chinstrap: 68   Dream    :123   1st Qu.:39.50   1st Qu.:15.60
 Gentoo   :119   Torgersen: 47   Median :44.50   Median :17.30
                                 Mean   :43.99   Mean   :17.16
                                 3rd Qu.:48.60   3rd Qu.:18.70
                                 Max.   :59.60   Max.   :21.50
 flipper_length_mm  body_mass_g       sex         year
 Min.   :172       Min.   :2700   female:165   2007:103
 1st Qu.:190       1st Qu.:3550   male  :168   2008:113
 Median :197       Median :4050                2009:117
 Mean   :201       Mean   :4207
 3rd Qu.:213       3rd Qu.:4775
 Max.   :231       Max.   :6300
```

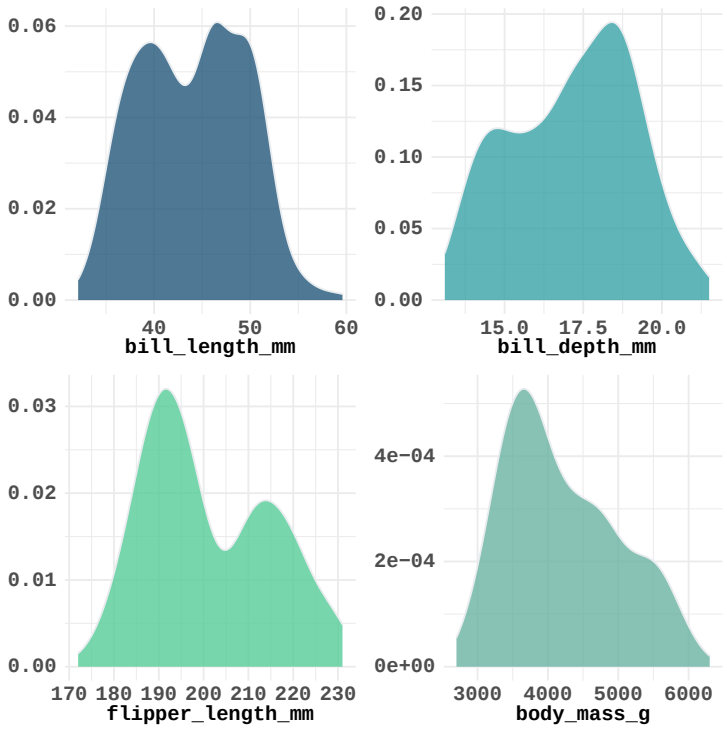
```
> # zmienne jakościowe ignorowane
> describeBy(penguins_data[c(3:6)])

                  vars   n    mean     sd median trimmed    mad    min    max
bill_length_mm       1 333   43.99   5.47   44.5   43.98   6.97   32.1   59.6
bill_depth_mm        2 333   17.16   1.97   17.3   17.19   2.22   13.1   21.5
flipper_length_mm    3 333  200.97  14.02  197.0  200.36  16.31  172.0  231.0
body_mass_g          4 333 4207.06 805.22 4050.0 4159.46 889.56 2700.0 6300.0
                   range  skew kurtosis    se
bill_length_mm      27.5  0.04    -0.90  0.30
bill_depth_mm        8.4 -0.15    -0.91  0.11
flipper_length_mm   59.0  0.36    -0.98  0.77
body_mass_g       3600.0  0.47    -0.75 44.13
```

Poniżej kilka wykresów przedstawiających cechy numeryczne:



Jak widać, dla wszystkich cech, z wyjątkiem *bill_depth_mm*, *skew*, czyli współczynnik skośności przyjmuje wartości dodatnie, oznacza to, że rozkłady są (w niewielkim stopniu, bo wartości *skew* są nadal bliskie zera) prawostronnie asymetryczne (mają wydłużone prawe ramię rozkładu). Rozkład *bill_depth_mm* jest w podobnym stopniu lewostronnie asymetryczny. Na podstawie wartości kurtozy, która dla każdej cechy jest ujemna, można stwierdzić, że intensywność

wartości skrajnych jest mniejsza niż w przypadku rozkładu normalnego. Dobrze obrazują to poniższe wykresy, tzw. *density plots*:



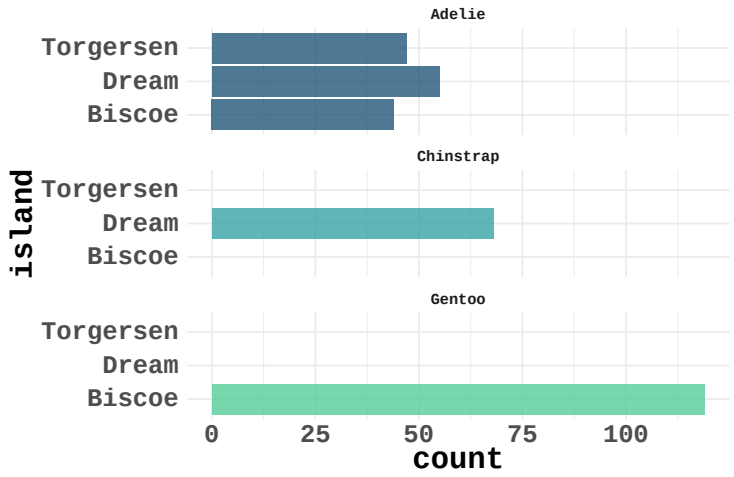
Badanie, czy dane pochodzą z jakiegoś rozkładu, zostanie przeprowadzone w dalszej części dokumentu, ale już teraz, na podstawie wykresów i ich podobieństwa do wykresu gęstości rozkładu normalnego, można snuć pewne przypuszczenia na ten temat.

3.2 Badanie zależności pomiędzy danymi

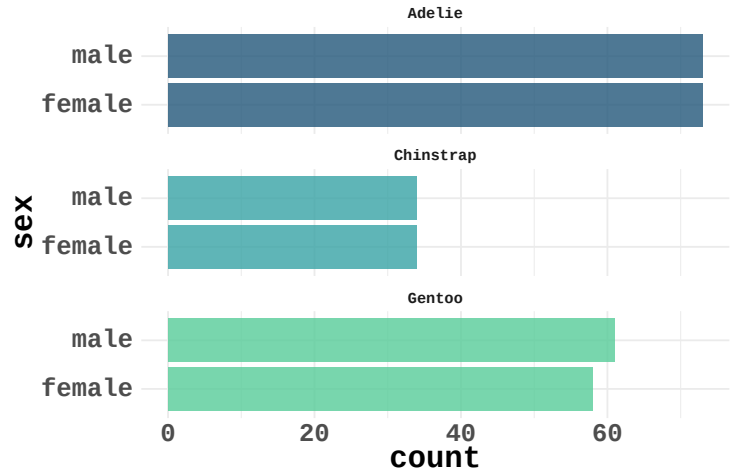
Na początku pod lupę zostanie wziętych kilka zmiennych jakościowych i zależności pomiędzy nimi.

```
> # gatunek, a wyspa na ktorej zostal zaobserwowany
> penguins_data %>% count(species, island)
```

```
    species    island   n
1    Adelie    Biscoe  44
2    Adelie     Dream  55
3    Adelie Torgersen  47
4 Chinstrap     Dream  68
5    Gentoo    Biscoe 119
```




```
> # gatunek, a plec
> penguins_data %>% count(sex, species)

     sex   species  n
1 female    Adelie 73
2 female Chinstrap 34
3 female    Gentoo 58
4   male    Adelie 73
5   male Chinstrap 34
6   male    Gentoo 61
```

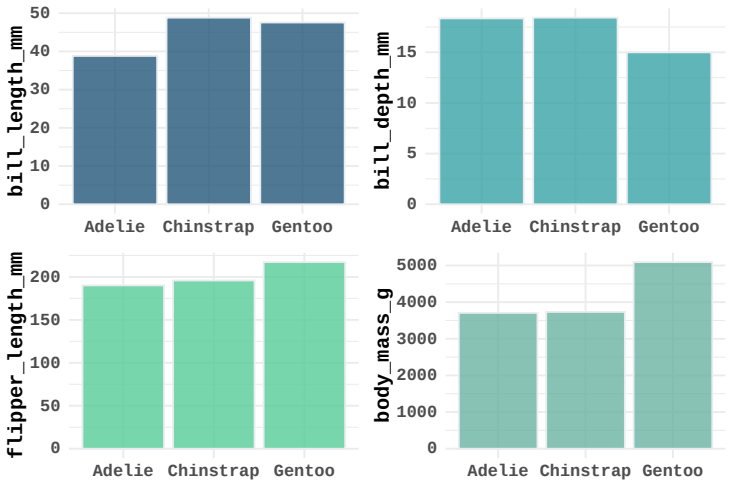


Ciekawą cechą, którą można tutaj zauważyć, jest to, że tylko gatunek *Adelie* żyje na wszystkich trzech wyspach, reszta gatunków zamieszkuje jedną wyspę.

Rozkład płci względem gatunku jest praktycznie równomierny, trudno zaobserwować inne zależności pomiędzy cechami jakościowymi.

Teraz zależności pomiędzy cechami numerycznymi i jakościowymi. Dla uproszczenia pod uwagę wzięte zostaną tylko wartości średnie grupowane po cechach jakościowych.

```
> # zaleznosc pomiedzy gatunkiem a cechami numerycznymi
> penguins_data %>% group_by(species) %>%
+   summarise(across(-c(island, sex, year), mean))

# A tibble: 3 x 5
  species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>              <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie              38.8          18.3              190.       3706.
2 Chinstrap           48.8          18.4              196.       3733.
3 Gentoo              47.6          15.0              217.       5092.
```



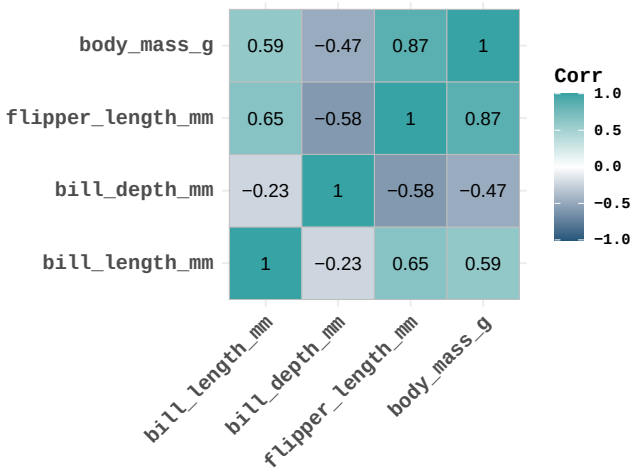
Jak widać na podstawie średnich, cechy numeryczne pingwinów z gatunków *Adelie* oraz *Chinstrap* nie różnią się bardzo z wyjątkiem długości dzioba (średnio nieznacznie dłuższe u *Chinstrap*). Pingwiny gatunku *Gentoo* wyróżniają się natomiast średnio krótszymi głębokościami dzioba, ale większymi masami ciała.

```
> # zaleznosc pomiedzy plcia a cechami numerycznymi
> penguins_data %>% group_by(species, sex) %>%
+   summarise(across(-c(island, year), mean))

# A tibble: 6 x 6
# Groups:   species [3]
  species   sex    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>     <fct>           <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie    female           37.3          17.6              188.       3369.
2 Adelie    male             40.4          19.1              192.       4043.
3 Chinstrap female           46.6          17.6              192.       3527.
4 Chinstrap male             51.1          19.3              200.       3939.
5 Gentoo    female           45.6          14.2              213.       4680.
6 Gentoo    male             49.5          15.7              222.       5485.
```

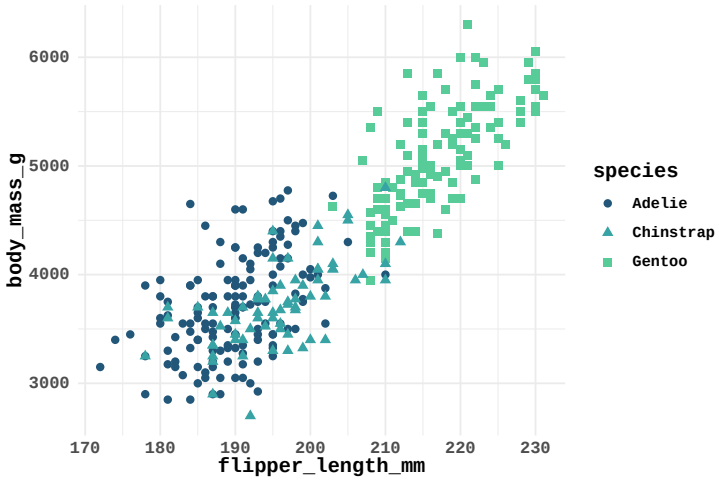
Samce pingwinów, niezależnie od gatunku, przeważają pod względem każdej cechy, szczególnie masy ciała.

Teraz pora na zbadanie zależności pomiędzy cechami numerycznymi. Również można by to podzielić, rozbijając dane na grupy względem cech jakościowych, ale tutaj, dla uproszczenia, użyte zostaną dane z całej próby. Najpierw obliczona zostanie macierz korelacji dla cech numerycznych:

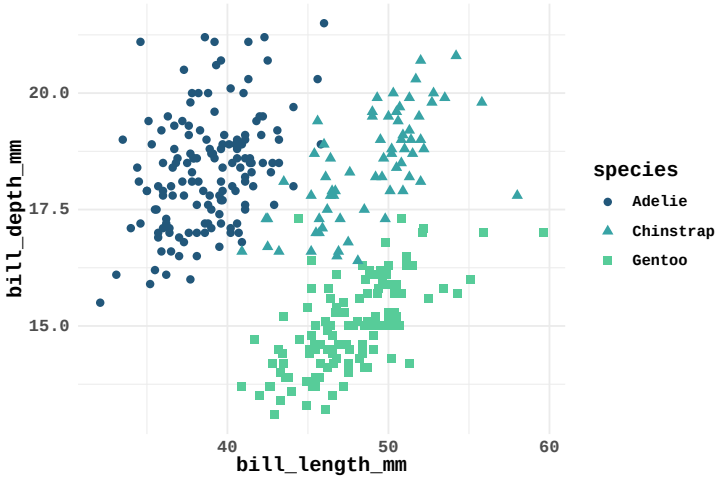
```
> corr <- round(cor(penguins_data[c(3:6)]), 2)
```



Można zaobserwować m.in. wysoki współczynnik korelacji pomiędzy masą ciała pingwina, a długością płetwy (skrzydła). Łatwo to zaobserwować na poniższym wykresie:



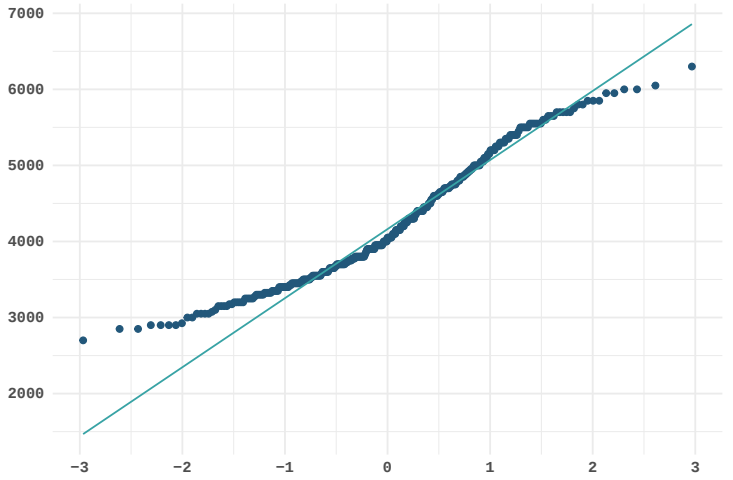
Długość i głębokość dzioba wykazują nieznacznie ujemny poziom korelacji dla całej próby danych, natomiast dla poszczególnych gatunków są to cechy nadal silnie skorelowane, co wynika z poniższego wykresu (tzw. *paradoks Simpsona*):



3.3 Analiza rozkładu badanych cech

Teraz nastąpi test tego, na ile rozkład jednej z cech numerycznych odpowiada rozkładowi normalnemu. Dla uproszczenia zostanie on przeprowadzony tylko dla jednej cechy. Bazując na wykresach z sekcji **3.1**, wybrana została cecha *body_mass_g*.

Histogram raczej przeczy temu, że rozkład ten jest normalny. Dla pewności użyty zostanie *Q-Q plot* (*quantile-quantile plot*, inna wizualna metoda testowania, czy rozkład jest normalny):



W przypadku rozkładu normalnego punkty wykresu tworzyłyby prostą, Tutaj natomiast można zaobserwować stosunkowo duże odchylenie od prostej, co również sugeruje, że rozkład nie jest normalny.

Ostatnim krokiem będzie przeprowadzenie testu Shapiro-Wilk'a, czyli testu sprawdzającego, czy zbiór danych jest dobrze modelowany przez rozkład normalny. Hipotezą zerową w tym teście jest H_0: rozkład jest normalny, natomiast hipotezą alternatywną H_1: rozkład nie jest normalny. Należy wspomnieć, że test został stworzoną z myślą o liczebności próbki poniżej 50 obserwacji.

```
> shapiro.test(sample(penguins_data$body_mass_g, 50))

        Shapiro-Wilk normality test

data:  sample(penguins_data$body_mass_g, 50)
W = 0.9179, p-value = 0.001982
```

Interpretacja testu sprowadza się do porównania wartości p z przyjętym poziomem istotności α. Tutaj za poziom istotności przyjęta zostanie wartość $\alpha = 0.05$. Z uwagi na wykorzystanie tylko losowych 50 wartości z całego zbioru danych, wyniki mogą się różnić w zależności od wywołania, jednak w większości przypadków przyjmują wartości < 0.05, co pozwala na odrzucenie hipotezy zerowej.

Ostatecznie, na podstawie powyższych trzech metod, można stwierdzić, że rozkład cechy *body_mass_g* nie jest normalny.

3.4 Estymatory przedziałowe

W tej sekcji zostaną wyznaczone niektóre wartości estymatorów przedziałowych. Dla uproszczenia, badaną cechą będzie tylko *body_mass_g*. Cecha ta nie należy do rozkładu normalnego, co zostało ustalone w poprzedniej sekcji, więc badanym przedziałem ufności będzie tylko przedział dla średniej. Wartość wariancji czy odchylenia standardowego nie są znane. W związku z tymi ograniczeniami użyta zostanie statystyka

$$T = \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$$

Wiadomo, że jeżeli X ma rozkład normalny to

$$T \sim t_{n-1}$$

gdzie t_n jest rozkładem t o n stopniach swobody, zwanym też **rozkładem Studenta**. W tym przypadku, kiedy rozkład X nie jest normalny, statystyka T ma (dla $n >$ ok. 20) rozkład zbliżony do rozkładu Studenta. Za poziom ufności zostanie przyjęta wartość $\alpha = 1 - 0.95$, wówczas, dla najkrótszego przedziału ufności

$$P(t_{\frac{\alpha}{2},n-1} \leqslant T \leqslant t_{1-\frac{\alpha}{2},n-1}) = 1 - \alpha$$

Czyli przedział ufności dla średniej to

$$\left[\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$$

Funkcja w R obliczające krańce tego przedziału:

```
> meanCI <- function(x, conf.level) {
+   alpha <- 1 - conf.level
+   n <- length(x)
+   return (c(mean(x) - qt(1-alpha/2, n-1)*sd(x)/ sqrt(n),
+             mean(x) + qt(1-alpha/2, n-1)*sd(x)/ sqrt(n))
+   )
+ }

> meanCI(penguins_data$body_mass_g, 0.95)

[1] 4120.256 4293.858
```

Natomiast funkcja z pakietu *stats* zwraca następujący wynik:

```
> t.test(penguins_data$body_mass_g)$"conf.int"

[1] 4120.256 4293.858
attr(,"conf.level")
[1] 0.95
```

tak więc przyjęta metoda jest poprawna. Stąd przedział ufności średniej dla rozkładu cechy *body_mass_g* wynosi $[4120.256, 4293.858]$.

4 Wnioski

Analiza statystyczna zestawu danych użytych w projekcie dostarcza wielu potencjalnie wartościowych informacji dla ornitologii i zoologii:

- pozwala przewidywać podstawowe cechy pingwinów,
- traktuje o zależnościach pomiędzy cechami tych ptaków,
- umożliwia powiązanie miejsc występowania z zamieszkującymi je gatunkami,
- pokazuje zmiany zachodzące w cechach pingwinów na przestrzeni lat.

Oczywiście ów projekt jest jedynie przykładem zastosowania statystycznej analizy danych, wymienione zalety jej wykorzystania można uogólnić, statystyka daje ogromne możliwości w badaniu danych i odkrywaniu ciekawych i użytecznych zależności.

Również należałoby wspomnieć o użytych technologiach: język **R** i jego biblioteki zawierające wiele przydatnych narzędzi do analizy statystycznej, **ggplot** pozwalający na stworzenia atrakcyjnych wizualnie wykresów oraz **Sweave** - narzędzie umożliwiające używania kodu R wewnątrz LaTeX. Użycie tych technologii ułatwiło i uprzyjemniło proces tworzenia owego projektu.