Proiect

Popescu Cristina Alexandra, grupa 322

Introducere

Scopul proiectului este de a realiza o analiza de varianță (ANOVA), care este o metoda statistică utilizata pentru a studia diferențele între medii.

Setul de date folosit pentru analiza este `penguins`, din (Horst, Hill, and Gorman 2020), care provine dintr-un studiu ce a investigat diverse caracteristici ale speciilor de pinguini: Adélie, Gentoo si Chinstrap. Scopul este sa gasim daca exista diferente in medie intre inaltimile ciocurilor, raportandu-ne la specie.

```
head(penguins)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           NA            NA                  NA          NA
5 Adelie  Torgersen           36.7          19.3               193        3450
6 Adelie  Torgersen           39.3          20.6               190        3650
# i 2 more variables: sex <fct>, year <int>
```

```
pg <- na.omit(penguins)
head(pg)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
```

```
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           36.7          19.3               193        3450
5 Adelie  Torgersen           39.3          20.6               190        3650
6 Adelie  Torgersen           38.9          17.8               181        3625
# i 2 more variables: sex <fct>, year <int>
```

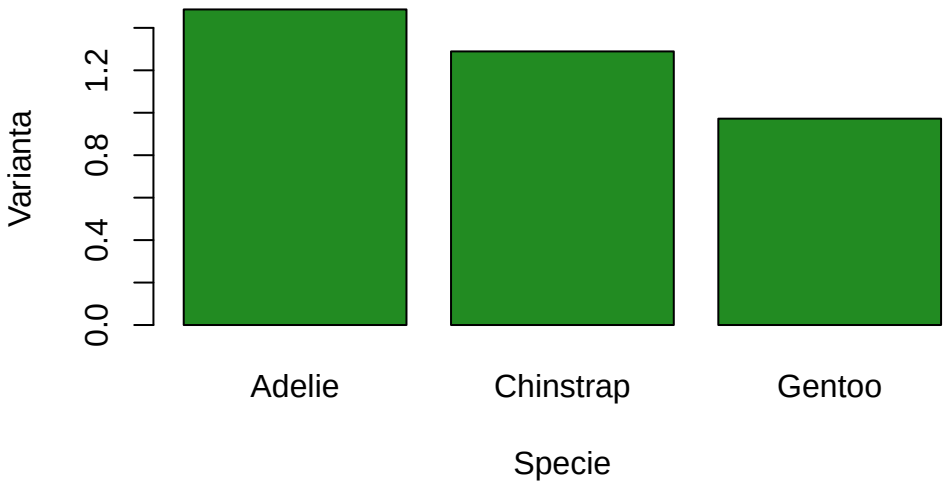
```r
length(pg$species)
```

```
[1] 333
```

Testul Levene este o statistica utilizata pentru a vedea daca variantele dintre mai multe grupuri de date sunt egale sau nu. Acest test este folosit pentru a verifica ipoteza nula conform careia toate grupurile au aceeasi varianta. Daca valoarea $Pr(> F)$ este mai mica decat 0.05, aceasta indica faptul ca exista diferente intre variante, caz in care se respinge ipoteza nula. Daca $Pr(> F)$ este mai mare decat 0.05, nu avem suficiente dovezi pentru a respinge ipoteza nula, deci variantele grupurilor nu difera foarte mult. Vom face o reprezentare grafica a variantelor inaltimii in functie de specie, dupa care vom aplica testul Levene.

```r
varianta <- aggregate(bill_depth_mm ~ species, data = pg, FUN = var)


barplot(height = varianta$bill_depth_mm,
        names.arg = varianta$species,
        xlab = "Specie",
        ylab = "Varianta",
        col = "forestgreen",
        cex.main = 0.7)
```



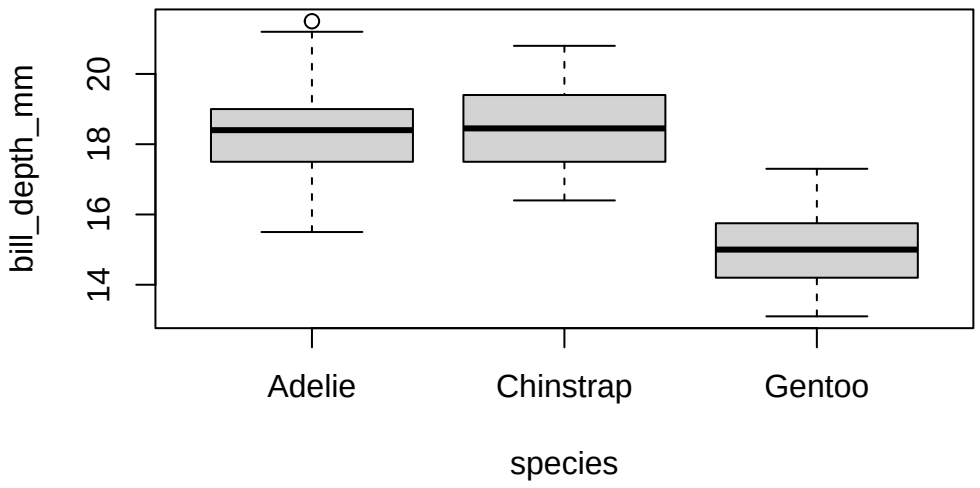

```
levene_r <- leveneTest(bill_depth_mm ~ species, data = pg)
print(levene_r)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  1.9124 0.1494
      330
```

Se observa ca sunt aproximativ egale, fapt sustinut si de valoarea obtinuta in urma testului.

`0.1494 > 0.05.`

```
plot(bill_depth_mm ~ species, data=pg)
```



Este doar o observatie in afara boxplot-ului in cazul primei specii, insa nu va influenta modelul in mod semnificativ.

Analiza de Varianta (ANOVA) Unifactoriala

Aceasta sectiune se bazeaza pe informatiile din (DeGroot and Schervish 2010).

Fie datele y_{ij} impartite in $i = 1, \dots, p$ grupuri si avand $j = 1, \dots, n_i$ elemente pe fiecare grup. Modelul general este:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

unde μ media, α_i efectul nivelului i, fixat si necunoscut si ε_{ij} valorile reziduale. Consideram ca $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Definim

$$\beta_i = \mu + \alpha_i$$

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{bmatrix}$$

si matricea de design X

$$X = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 1 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 1 & 0 & 0 & ... & 0 \\ -- & -- & -- & -- & -- \\ 0 & 1 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 1 & 0 & ... & 0 \\ -- & -- & -- & -- & -- \\ ... & ... & ... & ... & ... \\ -- & -- & -- & -- & -- \\ 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & ... & 1 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

Matricea are dimensiunea $n \times p$, unde n numarul de observatii si fiecare coloana corespunde unui grup. Coloana pentru primul grup are n_1 de 1 si $n_2 + \cdots + n_p$ de 0. Coloana pentru al doilea grup are n_1 de 0 si n_2 de 1, urmate de $n_3 + \cdots + n_p$ de 0. Obtinem modelul liniar:

$$Y = X\beta + \varepsilon$$

In cazul setului nostru de date, matricea X este:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \dots & \dots & \dots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \dots & \dots & \dots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \dots & \dots & \dots \\ 0 & 0 & 1 \end{bmatrix}$$

deoarece avem 3 specii si este o matrice de 333 linii si 3 coloane.

```
l1 <- length(pg$species[pg$species == "Adelie"])
l2 <- length(pg$species[pg$species == "Chinstrap"])
l3 <- length(pg$species[pg$species == "Gentoo"])
cat("Numarul de pinguini din specia Adelie:",l1, "\n")
```

```
Numarul de pinguini din specia Adelie: 146
```

```
cat("Numarul de pinguini din specia Chinstrap:",l2, "\n")
```

```
Numarul de pinguini din specia Chinstrap: 68
```

```
cat("Numarul de pinguini din specia Gentoo:",l3, "\n")
```

```
Numarul de pinguini din specia Gentoo: 119
```

Deci, $n_1 = 146, n_2 = 68, n_3 = 119$.

Pentru aceasta parte am folosit informatiile din (Faraway 2004). Pentru realizarea analizei datelor vom folosi functia `aov`. Variabila raspuns este inaltimea ciocului (bill_depth_mm), deoarece este cea care este influentata de specie. Variabila explicativa este specia (species). Denumirea de unifactoriala vine de la faptul ca fiecare variabila raspuns este clasificata intr-un singur mod. ANOVA nu testeaza daca o anumita medie este mai mica decat alta, ci doar daca sunt egale sau nu. Ipoteza nula este ca toate mediile sunt egale.

```r
a <- aov(bill_depth_mm ~ species, data = pg)
summary(a)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
species       2  870.8   435.4   344.8 <2e-16 ***
Residuals   330  416.7     1.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vom presupune ca reziduurile sunt independente si normal repartizate.

Din cauza valorii mici $Pr(> F)$, exista diferente intre medii. Tabelul ofera urmatoarele informatii:

- *Df (Degrees of Freedom):* numarul de grade de libertate pentru "species" si pentru "Residuals". Cum sunt 333 de observatii si 3 specii, numarul de grade de libertate pentru specii va fi 2 si pentru valorile reziduale 330. Adunate ne dau numarul de grade de libertate total al modelului.
- *Sum Sq (Sum of Squares):* suma patratelor diferentelor intre valorile observate si cele estimate, i.e. pentru "species" reprezinta variatia totala explicata de efectele speciei asupra inaltimii, iar pentru "Residuals", variatia care nu poate sa fie explicata de model. Valoarea 870.8 este mare, ceea ce inseamna ca efectul speciei explica o mare parte din variatia datelor. Are o influenta semnificativa asupra inaltimii ciocurilor.
- *Mean Sq (Mean Square):* media sumelor patratelor (Sum Sq) impartita la gradele de libertate corespunzatoare.
- *F value:* raportul dintre Mean Sq al speciilor si Mean Sq al reziduurilor.
- *Pr(>F)/ valoarea p:* ne indica faptul ca exista o diferenta mare intre medii.

Concluzii ANOVA unifactoriala

Vom efectua o analiza post-hoc folosind functia `pairwise.t.test` si metoda Bonferroni pentru ajustarea valorii p. Nivelul ales initial era de 0.05. Se imparte acest numar la numarul de comparatii, in cazul acesta fiind 3 si se obtine aproximativ 0.17, noul nivel cu care trebuie comparat fiecare rezultat. In acest fel, se reduce probabilitatea de a avea erori in interpretare.

```r
#Testam daca sunt egale mediile pentru fiecare pereche
perechi <- pairwise.t.test(pg$bill_depth_mm, pg$species, p.adjust= "bonf")
perechi
```

```
    Pairwise comparisons using t tests with pooled SD

data:  pg$bill_depth_mm and pg$species

          Adelie Chinstrap
Chinstrap 1      -
Gentoo    <2e-16 <2e-16

P value adjustment method: bonferroni
```

Nu sunt diferente mari intre inaltimea ciocurilor pinguinilor din specia Chinstrap si celor din specia Adelie. Intre a celor din specia Gentoo si Adelie, precum si Gentoo si Chinstrap exista diferente semnificative.

DeGroot, Morris H., and Mark J. Schervish. 2010. *Probability and Statistics*. 4th ed. Addison-Wesley.

Faraway, Julian J. 2004. *Linear Models with r*. Chapman; Hall/CRC. https://doi.org/10.4324/9780203507278.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. "Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data." https://doi.org/10.5281/zenodo.3960218.