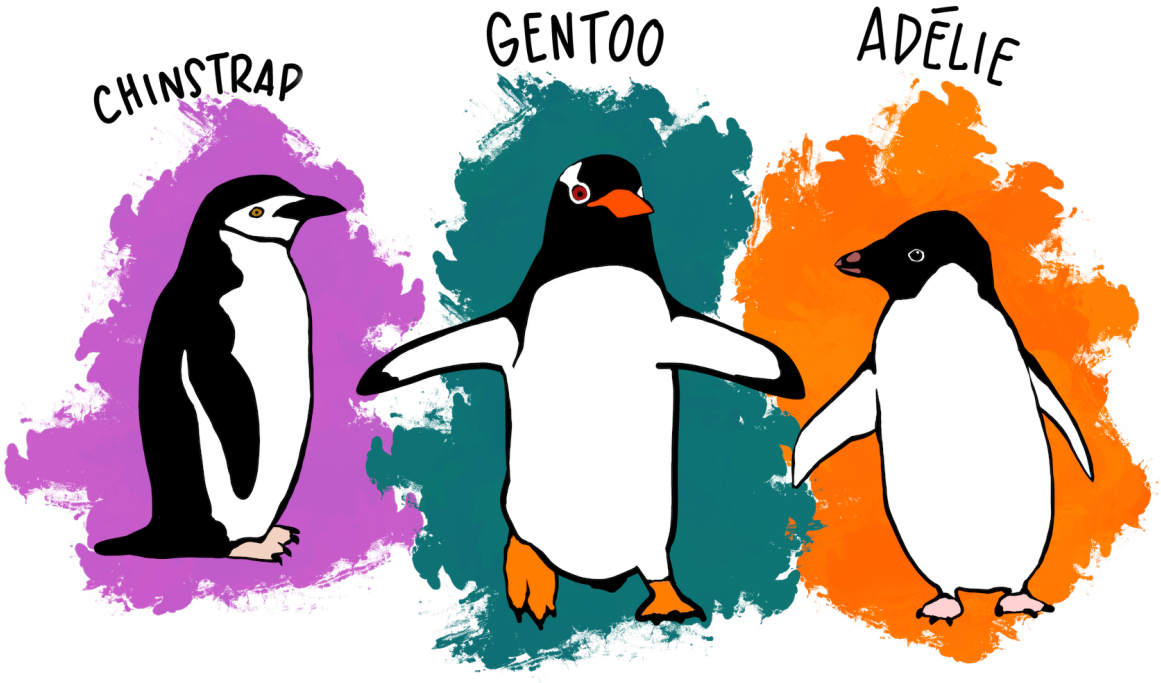
Análisis descriptivo de los datos

Contents

1 Introducción

El archivo pinguinos.csv contiene los datos de 333 pingüinos del archipiélago Palmer, en la Antártida. En concreto, el archivo consta de 7 columnas:

- especie: especie del pingüino (Chinstrap, Adelie, Gentoo)
- long_pico: longitud del pico (mm)
- prof_pico: medida de la parte más ancha del pico (mm)
- long_ala: longitud del ala (mm)
- peso: peso del pinguino (g)
- isla: nombre de la isla donde se recogió al pingüino (Dream, Torgersen, Biscoe)
- genero: genero del pingüino (hembra, macho)



Cada columna del archivo corresponde a una **variable**, y cada fila corresponde a una **observación** de dichas variables. Por tanto, tenemos 7 variables y 333 observaciones.

Las variables son de dos tipos:

- cualitativas: especie, isla y género.
- cuantitativas: long_pico, prof_pico, long_ala y peso.

2 Análisis de variables cualitativas

Las variables cualitativas se denominan **factores** en R. Y los posibles valores que puede tomar cada variable cualitativa se llaman **niveles del factor**. Por ejemplo, los niveles de especie son: Chinstrap, Adelie, Gentoo.

En general R no declara automáticamente los factores, por lo que el primer paso es definirlos:

```
# se leen los datos
d = read.csv("datos/pinguinos.csv", sep = ";")
# se muestra la estructura de la variable d: tipo de dato, tamaño, contenido,...
str(d)
```

```
## 'data.frame':    333 obs. of  7 variables:
##  $ especie   : chr  "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ isla      : chr  "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ long_pico : num  39.1 39.5 40.3 36.7 39.3 38.9 39.2 41.1 38.6 34.6 ...
##  $ prof_pico : num  18.7 17.4 18 19.3 20.6 17.8 19.6 17.6 21.2 21.1 ...
##  $ long_aleta: int  181 186 195 193 190 181 195 182 191 198 ...
##  $ peso      : int  3750 3800 3250 3450 3650 3625 4675 3200 3800 4400 ...
##  $ genero    : chr  "macho" "hembra" "hembra" "hembra" ...
```

Como se observa, especie, isla y genero son de tipo carácter (chr) y no son factores. Por tanto se convierten a factor:

```
# se define como factor y lo se guarda en genero1
d$genero1 = factor(d$genero)
head(d)
```

```
##   especie      isla long_pico prof_pico long_aleta peso genero genero1
## 1  Adelie Torgersen      39.1      18.7        181 3750  macho   macho
## 2  Adelie Torgersen      39.5      17.4        186 3800 hembra  hembra
## 3  Adelie Torgersen      40.3      18.0        195 3250 hembra  hembra
## 4  Adelie Torgersen      36.7      19.3        193 3450 hembra  hembra
## 5  Adelie Torgersen      39.3      20.6        190 3650  macho   macho
## 6  Adelie Torgersen      38.9      17.8        181 3625 hembra  hembra
```

La variable no ha cambiado, solo la manera en que se define. Lo habitual es guardarlas en la variable de origen para no duplicar información:

```
d$especie = factor(d$especie)
d$isla = factor(d$isla)
```

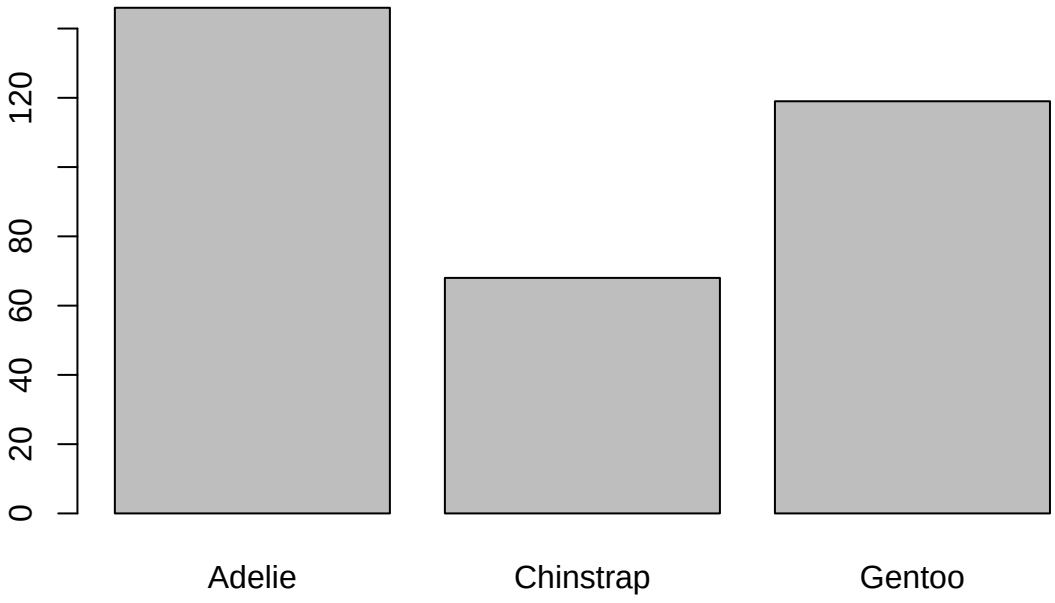
Ya se puede realizar un análisis descriptivo de las variables. En primer lugar se van a contar cuantos pingüinos hay de cada especie:

```
(t = table(d$especie))
```

```
##
##    Adelie Chinstrap    Gentoo
##       146        68       119
```

De manera gráfica:

```
barplot(t)
```



También se pueden calcular proporciones:

```
prop.table(t)
```

```
##
##    Adelie Chinstrap    Gentoo
## 0.4384384 0.2042042 0.3573574
```

También es posible trabajar con dos variables:

```
(t1 = table(d$especie, d$isla))
```

```
##
##             Biscoe Dream Torgersen
##   Adelie        44    55        47
##   Chinstrap      0    68         0
##   Gentoo       119     0         0
```

3 Análisis de variables cuantitativas

El resto de variables son cuantitativas. Por ejemplo, si se analiza la variable peso.

```
# minimo
min(d$peso)
```

```
## [1] 2700
```

```
# maximo
max(d$peso)
```

```
## [1] 6300
```

```
# media
mean(d$peso)
```

```
## [1] 4207.057
```

```
# mediana
median(d$peso)
```

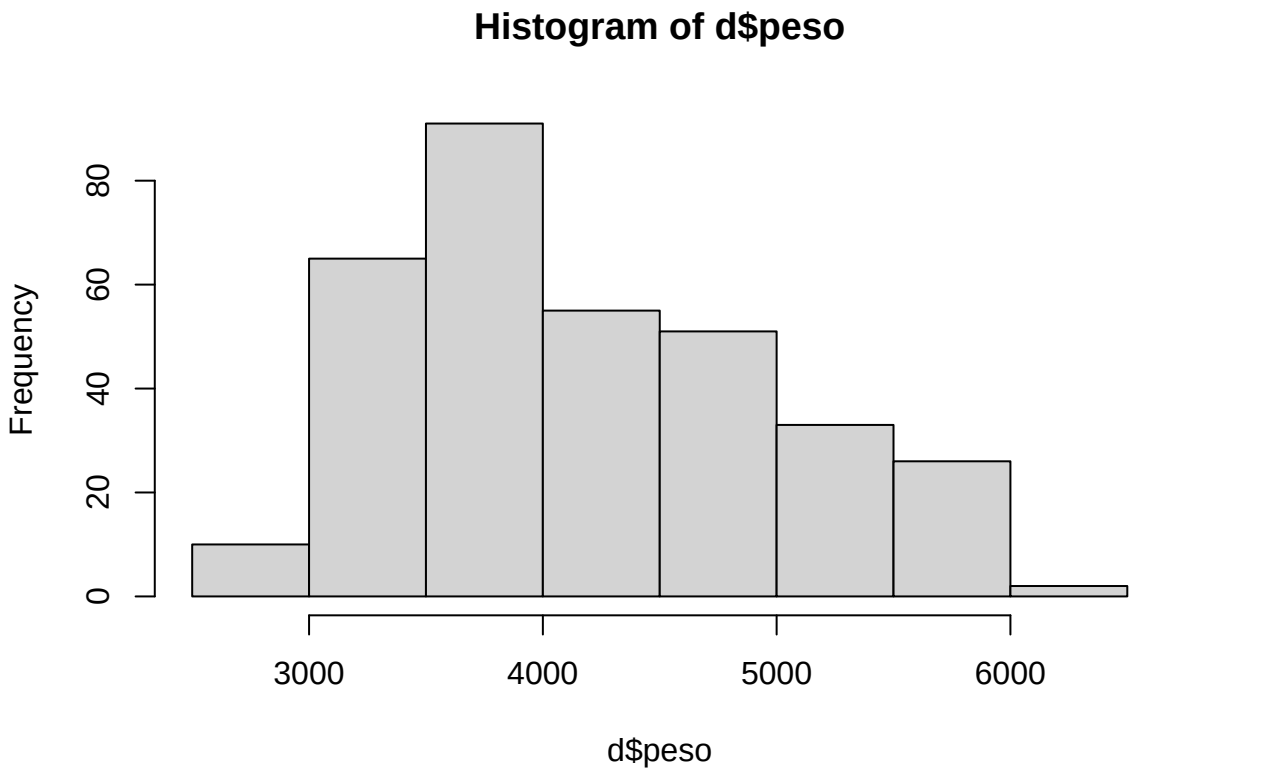
```
## [1] 4050
```

```
# desviacion típica
sd(d$peso)
```

```
## [1] 805.2158
```

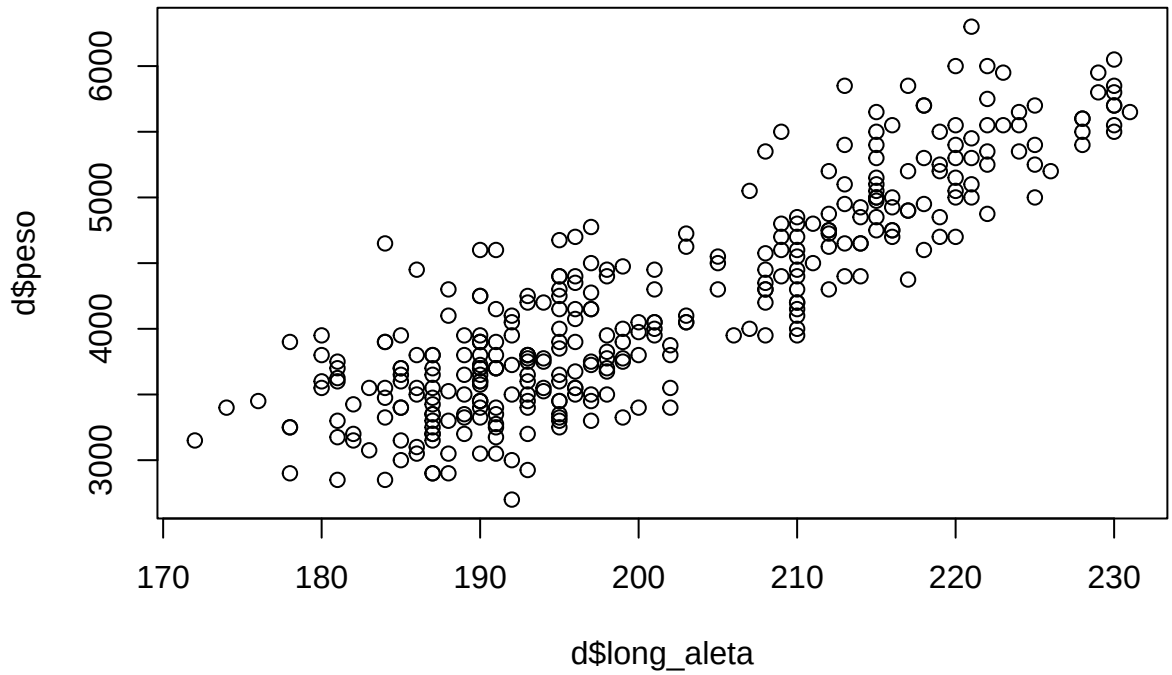
El gráfico más importante es el histograma ya que indica el rango de la variable y como se distribuyen sus valores:

```
hist(d$peso)
```



Cuando se tienen dos variables cuantivativas se puede dibujar el gráfico de dispersión para ver la relación entre ambas:

```
plot(d$long_aleta, d$peso)
```



De manera analítica se puede calcular la covarianza o el coeficiente de correlación lineal:

```
# covarianza
cov(d$long_aleta, d$peso)
```

```
## [1] 9852.192
```

```
# coef. correlación
cor(d$long_aleta, d$peso)
```

```
## [1] 0.8729789
```

El valor de la covarianza depende de las unidades de las variables: un valor alto o bajo es difícil de interpretar. Sin embargo, el coeficiente de correlación lineal toma valores entre -1 y 1. En este caso, 0.873 indica una correlación positiva (a mayor longitud de aleta, mayor peso) y alta (la longitud de la aleta es un buen indicador del peso).