Différencier le *Trichosurus cunninghami* du *Trichosurus caninus* d'après ses caractéristiques morphologiques

SCI 1402 – Projet en science des données

TÉLUQ

Automne 2024

Dupuis, Geneviève

No d'étudiante : 22307619

Introduction

En 1995, le scientifique David Lindenmayer entreprend une étude sur les variations morphométriques entre des populations de *Trichosurus caninus Ogilby*, des opossums connus sous le nom vernaculaire anglais de *mountain brushtail possums*. Pour réaliser l'étude, des opossums sont capturés sur sept sites répartis entre le sud de Victoria et le centre de Queensland, dans l'est de l'Australie, et neuf mesures morphologiques sont consignées pour chaque animal. Lindenmayer et son équipe observent des variations significatives entre les populations du sud et du nord et publient leurs résultats dans l'*Australian Journal of Zoology* (https://www.publish.csiro.au/zo/ZO9950449). Une seconde étude, dont les résultats sont aussi publiés dans l'*Australian Journal of Zoology* (https://www.publish.csiro.au/zo/ZO01047) en 2002, permet au chercheur de démontrer qu'en plus des variations morphométriques, il existe des différences constantes entre les caractéristiques génétiques des deux populations. À la suite de ces conclusions, les opossums étudiés sont classés en deux espèces distinctes de *Trichosurus*, soit *T. caninus* pour la population du nord (du centre de New South Wales au centre de Queensland), et *T. cunninghami* pour la population du sud (Victoria).

L'objectif de ce projet est de déterminer s'il est possible d'observer un dimorphisme entre les opossums étudiés à partir des mesures consignées par Lindenmayer, d'une part en fonction du sexe, et d'autre part en fonction de la région. Les relations entre les caractéristiques morphologiques et le sexe seront d'abord explorées, et l'hypothèse selon laquelle la circonférence du ventre des femelles serait plus grande que celle du ventre des mâles en raison de leur marsupium sera vérifiée. Les relations entre les caractéristiques morphologiques et la région seront ensuite explorées dans le but d'identifier quelles caractéristiques sont les plus corrélées à l'habitat. À partir des résultats obtenus, un modèle sera développé pour prédire l'espèce à laquelle appartient un opossum en fonction des ses mesures morphologiques.

Les données utilisées sont disponibles dans le paquetage DAAG de R. Elles incluent les mesures morphologiques, *Possum Measurements* (https://search.r-project.org/CRAN/refmans/DAAG/html/possum.html), et la localisation des sites de capture des opossums, *Possum Sites* (https://search.r-

project.org/CRAN/refmans/DAAG/html/possumsites.html). De plus, des données spatiales de l'Australie ont été récupérées sur le site opendatasoft.com (https://data.opendatasoft.com/explore/dataset/georef-australia-state%40public/export/?disjunctive.ste_code&disjunctive.ste_name).

1. Rencontre avec les données

```
# Importation des données
library(DAAG)
data("possum")
data("possumsites")
```

1.1 Exploration du jeu de données `possum`

```
dim(possum)
```

```
## [1] 104  14
```

```
names(possum)
```

```
##  [1] "case"     "site"     "Pop"      "sex"      "age"      "hdlngth"
##  [7] "skullw"   "totlngth" "taill"    "footlgth" "earconch" "eye"
## [13] "chest"    "belly"
```

```
str(possum)
```

```
## 'data.frame':    104 obs. of  14 variables:
##  $ case    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ site    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Pop     : Factor w/ 2 levels "Vic","other": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex     : Factor w/ 2 levels "f","m": 2 1 1 1 1 1 2 1 1 1 ...
##  $ age     : num  8 6 6 6 2 1 2 6 9 6 ...
##  $ hdlngth : num  94.1 92.5 94 93.2 91.5 93.1 95.3 94.8 93.4 91.8 ...
##  $ skullw  : num  60.4 57.6 60 57.1 56.3 54.8 58.2 57.6 56.3 58 ...
##  $ totlngth: num  89 91.5 95.5 92 85.5 90.5 89.5 91 91.5 89.5 ...
##  $ taill   : num  36 36.5 39 38 36 35.5 36 37 37 37.5 ...
##  $ footlgth: num  74.5 72.5 75.4 76.1 71 73.2 71.5 72.7 72.4 70.9 ...
##  $ earconch: num  54.5 51.2 51.9 52.2 53.2 53.6 52 53.9 52.9 53.4 ...
##  $ eye     : num  15.2 16 15.5 15.2 15.1 14.2 14.2 14.5 15.5 14.4 ...
##  $ chest   : num  28 28.5 30 28 28.5 30 30 29 28 27.5 ...
##  $ belly   : num  36 33 34 34 33 32 34.5 34 33 32 ...
```

```
head(possum)
```

```
##     case site Pop sex age hdlngth skullw totlngth taill footlgth earconch  eye
## C3     1    1 Vic   m   8    94.1   60.4     89.0  36.0     74.5     54.5 15.2
## C5     2    1 Vic   f   6    92.5   57.6     91.5  36.5     72.5     51.2 16.0
## C10    3    1 Vic   f   6    94.0   60.0     95.5  39.0     75.4     51.9 15.5
## C15    4    1 Vic   f   6    93.2   57.1     92.0  38.0     76.1     52.2 15.2
## C23    5    1 Vic   f   2    91.5   56.3     85.5  36.0     71.0     53.2 15.1
## C24    6    1 Vic   f   1    93.1   54.8     90.5  35.5     73.2     53.6 14.2
##     chest belly
## C3   28.0    36
## C5   28.5    33
## C10  30.0    34
## C15  28.0    34
## C23  28.5    33
## C24  30.0    32
```

Le jeu de données comprend 104 observations d'opossums, et 14 variables les décrivant, soit :

- un numéro d'identification,
- le site sur lequel l'animal a été capturé,
- la région dans laquelle se trouve le site,
- le sexe,
- l'âge,
- la longueur de la tête,
- la largeur du crâne,
- la longueur totale,
- la longueur de la queue,
- la longueur du pied,
- la longueur de la conque de l'oreille,
- la distance entre le canthus médial et le canthus latéral de l'œil droit,
- la circonférence de la poitrine,
- la circonférence du ventre.

Toutes les variables sont numériques, à l'exception de `Pop` et `sex`, qui sont des facteurs de deux niveaux.

Vérification de la présence de données manquantes

```
library(naniar)
summary(possum)
```

```
##       case             site          Pop     sex         age
##  Min.   :  1.00   Min.   :1.000   Vic  :46   f:43   Min.   :1.000
##  1st Qu.: 26.75   1st Qu.:1.000   other:58   m:61   1st Qu.:2.250
##  Median : 52.50   Median :3.000                     Median :3.000
##  Mean   : 52.50   Mean   :3.625                     Mean   :3.833
##  3rd Qu.: 78.25   3rd Qu.:6.000                     3rd Qu.:5.000
##  Max.   :104.00   Max.   :7.000                     Max.   :9.000
##                                                     NA's   :2
##     hdlngth           skullw         totlngth         taill
##  Min.   : 82.50   Min.   :50.00   Min.   :75.00   Min.   :32.00
##  1st Qu.: 90.67   1st Qu.:54.98   1st Qu.:84.00   1st Qu.:35.88
##  Median : 92.80   Median :56.35   Median :88.00   Median :37.00
##  Mean   : 92.60   Mean   :56.88   Mean   :87.09   Mean   :37.01
##  3rd Qu.: 94.72   3rd Qu.:58.10   3rd Qu.:90.00   3rd Qu.:38.00
##  Max.   :103.10   Max.   :68.60   Max.   :96.50   Max.   :43.00
##
##     footlgth        earconch          eye            chest          belly
##  Min.   :60.30   Min.   :40.30   Min.   :12.80   Min.   :22.0   Min.   :25.00
##  1st Qu.:64.60   1st Qu.:44.80   1st Qu.:14.40   1st Qu.:25.5   1st Qu.:31.00
##  Median :68.00   Median :46.80   Median :14.90   Median :27.0   Median :32.50
##  Mean   :68.46   Mean   :48.13   Mean   :15.05   Mean   :27.0   Mean   :32.59
##  3rd Qu.:72.50   3rd Qu.:52.00   3rd Qu.:15.72   3rd Qu.:28.0   3rd Qu.:34.12
##  Max.   :77.90   Max.   :56.20   Max.   :17.80   Max.   :32.0   Max.   :40.00
##  NA's   :1
```

```
pct_miss(possum)
```

```
## [1] 0.206044
```

La fonction `summary` permet de constater qu'il y a des valeurs manquantes dans la variable `age` (deux valeurs manquantes) et `footlgth` (une valeur manquante), ce qui ne représente que 0,2 % des données. Lors de l'analyse, ces valeurs seront ignorées pour la variable `age` , et remplacées par la moyenne pour la variable `footlgth` .

Vérification de la présence de données aberrantes

```
library(dlookr)
diagnose_outlier(possum)
```

```
##    variables outliers_cnt outliers_ratio outliers_mean with_mean without_mean
## 1       case            0      0.0000000           NaN 52.500000    52.500000
## 2       site            0      0.0000000           NaN  3.625000     3.625000
## 3        age            0      0.0000000           NaN  3.833333     3.833333
## 4    hdlngth            3      2.8846154      96.03333 92.602885    92.500990
## 5     skullw            7      6.7307692      62.84286 56.883654    56.453608
## 6   totlngth            0      0.0000000           NaN 87.088462    87.088462
## 7      taill            4      3.8461538      37.12500 37.009615    37.005000
## 8   footlgth            0      0.0000000           NaN 68.459223    68.459223
## 9   earconch            0      0.0000000           NaN 48.130769    48.130769
## 10       eye            1      0.9615385      17.80000 15.046154    15.019417
## 11     chest            1      0.9615385      32.00000 27.000000    26.951456
## 12     belly            2      1.9230769      32.50000 32.586538    32.588235
```

Il semble y avoir des données aberrantes dans les variables `hdlngth`, `eye`, `chest`, `belly`, et particulièrement dans `skullw` (6,73 %) et `taill` (3,85 %). La visualisation des données permettra de mieux cerner ces valeurs pour les variables d'intérêt, et de les remplacer si nécessaire lors de l'analyse.

1.2 Nettoyage des données de `possum`

La variable `case`, qui identifie chaque opossum, n'est d'aucune utilité, car l'identifiant correspond au numéro de ligne du jeu de données. Elle est donc supprimée.

Afin d'améliorer la compréhension des données, la variable `Pop` est renommée et les valeurs des variables `Pop` et `sex` sont clarifiées.

```
# Suppression de la colonne "case"
possum <- possum[ , -1]

# Variable "Pop" renommée pour plus de clarté
names(possum)[2] <- "region"

# Valeurs des facteurs renommées pour plus de clarté
possum$region <- as.character(possum$region)
possum$region[possum$region == "Vic"] <- "Victoria"
possum$region[possum$region == "other"] <- "NewSouthWales_Queensland"
possum$region <- as.factor(possum$region)
levels(possum$region)
```

```
## [1] "NewSouthWales_Queensland" "Victoria"
```

```
possum$sex <- as.character(possum$sex)
possum$sex[possum$sex == "m"] <- "male"
possum$sex[possum$sex == "f"] <- "female"
possum$sex <- as.factor(possum$sex)
levels(possum$sex)
```

```
## [1] "female" "male"
```

```
head(possum)
```

```
##     site   region    sex age hdlngth skullw totlngth taill footlgth earconch
## C3     1 Victoria   male   8    94.1   60.4     89.0  36.0     74.5     54.5
## C5     1 Victoria female   6    92.5   57.6     91.5  36.5     72.5     51.2
## C10    1 Victoria female   6    94.0   60.0     95.5  39.0     75.4     51.9
## C15    1 Victoria female   6    93.2   57.1     92.0  38.0     76.1     52.2
## C23    1 Victoria female   2    91.5   56.3     85.5  36.0     71.0     53.2
## C24    1 Victoria female   1    93.1   54.8     90.5  35.5     73.2     53.6
##      eye chest belly
## C3  15.2  28.0    36
## C5  16.0  28.5    33
## C10 15.5  30.0    34
## C15 15.2  28.0    34
## C23 15.1  28.5    33
## C24 14.2  30.0    32
```

1.3 Exploration du jeu de données `possumsites`

```
dim(possumsites)
```

```
## [1] 7 3
```

```
names(possumsites)
```

```
## [1] "Longitude" "Latitude"  "altitude"
```

```
head(possumsites)
```

```
##             Longitude  Latitude altitude
## Cambarville  145.8833 -37.55000      800
## Bellbird     148.8000 -37.61667      300
## Allyn River  151.4667 -32.11667      300
## Whian Whian  153.3333 -28.61667      400
## Byrangery    153.4167 -28.61667      200
## Conondale    152.5833 -26.43333      400
```

```
row.names(possumsites)
```

```
## [1] "Cambarville" "Bellbird"    "Allyn River" "Whian Whian" "Byrangery"
## [6] "Conondale"   "Bulburin"
```

Le jeu de données de `possumsites` comprend 7 observations et 3 variables, soit la longitude, la latitude et l'altitude. Les sites sont identifiés par leur nom, qui apparaît dans le nom des lignes du jeu de données.

1.4 Nettoyage des données de `possumsites`

La variable `altitude` est supprimée, car elle ne sera pas utile pour l'analyse. À l'inverse, le nom des sites est ajouté comme nouvelle variable, ainsi que le numéro qui leur est associé, numéro utilisé dans le jeu de données `possum`.

```
# Retrait de la variable "altitude"
possumsites <- possumsites[ ,-3]

# Vérification que le nombre de sites correspond dans les deux jeux de données
unique(possum$site)
```

```
## [1] 1 2 3 4 5 6 7
```

```
# Ajout des numéros et des noms des sites
possumsites$site_no <- c(1:7)
possumsites$site_name <- row.names(possumsites)
possumsites
```

```
##             Longitude  Latitude site_no   site_name
## Cambarville  145.8833 -37.55000       1 Cambarville
## Bellbird     148.8000 -37.61667       2    Bellbird
## Allyn River  151.4667 -32.11667       3 Allyn River
## Whian Whian  153.3333 -28.61667       4 Whian Whian
## Byrangery    153.4167 -28.61667       5   Byrangery
## Conondale    152.5833 -26.43333       6   Conondale
## Bulburin     151.4667 -24.55000       7    Bulburin
```

---

2. Visualisation exploratoire des données

Afin d'assurer une cohérence dans la représentation graphique des données, des palettes de couleurs sont définies pour le sexe et la région. Ainsi, les données concernant les femelles seront représentées en jaune, celles concernant les mâles, en bleu, celles de la région de Victoria, en rose, et enfin, celles de la région de New South Wales et Queensland, en vert.

```
# Définition des palettes de couleurs
sex_col <- c("gold1", "dodgerblue3")
region_col <- c("#7fbc41", "#ae377b")
sites_col <- c("#8e0152", "#f1b6da", "#e6f5d0", "#b8e186","#7fbc41", "#4d9221","#27641
9")
num_region_col <- c("black", "white")
```

2.1 Carte des sites de capture des opossums

Les sites de capture sont visualisés dans leur contexte, c'est-à-dire sur la carte de l'Australie et de ses États.

```
# Chargement des biliothèques nécessaires
library(sf)
library(tmap)
library(mapview)
```

```
# Chargement des données de l'Australie
australia <- st_read("georef-australia-state@public/georef-australia-state-millesime.sh
p")
```

```
## Reading layer `georef-australia-state-millesime' from data source
##   `/Users/gdupuis/Documents/TÉLUQ/Automne 2024/SCI 1402_Projet en science des donnée
s/SCI 1402-TN3-Genevieve-Dupuis/georef-australia-state@public/georef-australia-state-mil
lesime.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 9 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 96.81704 ymin: -43.65855 xmax: 167.9969 ymax: -9.222722
## Geodetic CRS:  WGS 84
```

```
australia
```

```
## Simple feature collection with 9 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 96.81704 ymin: -43.65855 xmax: 167.9969 ymax: -9.222722
## Geodetic CRS:  WGS 84
##   year ste_code                        ste_name ste_area_co  ste_type
## 1 2021    ['2']                     ['Victoria']         AUS     state
## 2 2021    ['1']               ['New South Wales']         AUS     state
## 3 2021    ['9']             ['Other Territories']         AUS territory
## 4 2021    ['5']             ['Western Australia']         AUS     state
## 5 2021    ['4']               ['South Australia']         AUS     state
## 6 2021    ['7']            ['Northern Territory']         AUS territory
## 7 2021    ['8'] ['Australian Capital Territory']         AUS territory
## 8 2021    ['6']                     ['Tasmania']         AUS     state
## 9 2021    ['3']                   ['Queensland']         AUS     state
##   ste_iso3166                       geometry
## 1         VIC MULTIPOLYGON (((149.9058 -3...
## 2         NSW MULTIPOLYGON (((159.0634 -3...
## 3        <NA> MULTIPOLYGON (((167.9475 -2...
## 4          WA MULTIPOLYGON (((117.8939 -3...
## 5          SA MULTIPOLYGON (((133.5444 -3...
## 6          NT MULTIPOLYGON (((132.825 -10...
## 7         ACT MULTIPOLYGON (((149.2318 -3...
## 8         TAS MULTIPOLYGON (((144.0492 -4...
## 9         QLD MULTIPOLYGON (((142.6369 -1...
```

2.1.1 Préparation des données

Les données de l'Australie sont au format *shapefile* d'ESRI et comprennent 9 éléments d'une géométrie de type multipolygone. Les États où ont été capturés les opossums sont isolés dans de nouveaux objets spatiaux, en regroupant New South Wales et Queensland comme une seule région. Seuls le nom des États et la géométrie sont conservés, et les crochets et apostrophes autour des noms des États sont supprimés.

```
# Régions où ont été capturés les opossums
victoria <- australia[1, c("ste_name", "geometry")]
nswales_queensland <- australia[c(2, 9), c("ste_name", "geometry")]

# Suppression des crochets et apostrophes
victoria$ste_name[victoria$ste_name == "['Victoria']"] <- "Victoria"
nswales_queensland$ste_name[nswales_queensland$ste_name == "['New South Wales']"] <- "Ne
w South Wales"
nswales_queensland$ste_name[nswales_queensland$ste_name == "['Queensland']"] <- "Queensl
and"
```

Pour être visualisées sur une carte, les données des sept sites sont converties en objet spatial, où les coordonnées deviennent une géométrie de type point. Comme les coordonnées sont au format longitude et latitude, le datum WGS 84 est utilisé pour définir le système de coordonnées de référence (code EPSG 4326).

```
# Conversion des données de "possumsites" en objet spatial
possumsites_points <- st_as_sf(x = possumsites,
                               coords = c("Longitude", "Latitude"),
                               crs = 4326)
possumsites_points
```

```
## Simple feature collection with 7 features and 2 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 145.8833 ymin: -37.61667 xmax: 153.4167 ymax: -24.55
## Geodetic CRS:  WGS 84
##             site_no   site_name                   geometry
## Cambarville       1 Cambarville    POINT (145.8833 -37.55)
## Bellbird          2    Bellbird    POINT (148.8 -37.61667)
## Allyn River       3 Allyn River POINT (151.4667 -32.11667)
## Whian Whian       4 Whian Whian POINT (153.3333 -28.61667)
## Byrangery         5   Byrangery POINT (153.4167 -28.61667)
## Conondale         6   Conondale POINT (152.5833 -26.43333)
## Bulburin          7    Bulburin    POINT (151.4667 -24.55)
```

Dans le jeu de données `possum` , chaque site est associé à une région. Cette information est d'abord extraite, ce qui permet de voir que les deux premiers sites sont associés à la région de Victoria, et les cinq autres font partie de la région qui regroupe New South Wales et Queensland. La région est ensuite ajoutée à l'objet spatial.

```
# Extraction des sites et de leur région correspondante
possum_regions <- possum[ , c("site", "region")]
```

```
# Une seule ligne par site conservée
library(dplyr)
possum_regions_unique <- possum_regions |>
  group_by(site) |>
  filter(row_number() == 1)
possum_regions_unique
```

```
## # A tibble: 7 × 2
## # Groups:   site [7]
##    site region
##   <dbl> <fct>
## 1     1 Victoria
## 2     2 Victoria
## 3     3 NewSouthWales_Queensland
## 4     4 NewSouthWales_Queensland
## 5     5 NewSouthWales_Queensland
## 6     6 NewSouthWales_Queensland
## 7     7 NewSouthWales_Queensland
```

```
# Ajout des régions à "possumsites_points"
possumsites_regions <- merge(x = possumsites_points, y = possum_regions_unique,
                         by.x = "site_no", by.y = "site")
possumsites_regions
```

```
## Simple feature collection with 7 features and 3 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 145.8833 ymin: -37.61667 xmax: 153.4167 ymax: -24.55
## Geodetic CRS:  WGS 84
##   site_no   site_name                   region                   geometry
## 1       1 Cambarville                 Victoria    POINT (145.8833 -37.55)
## 2       2    Bellbird                 Victoria    POINT (148.8 -37.61667)
## 3       3 Allyn River NewSouthWales_Queensland POINT (151.4667 -32.11667)
## 4       4 Whian Whian NewSouthWales_Queensland POINT (153.3333 -28.61667)
## 5       5   Byrangery NewSouthWales_Queensland POINT (153.4167 -28.61667)
## 6       6   Conondale NewSouthWales_Queensland POINT (152.5833 -26.43333)
## 7       7    Bulburin NewSouthWales_Queensland    POINT (151.4667 -24.55)
```

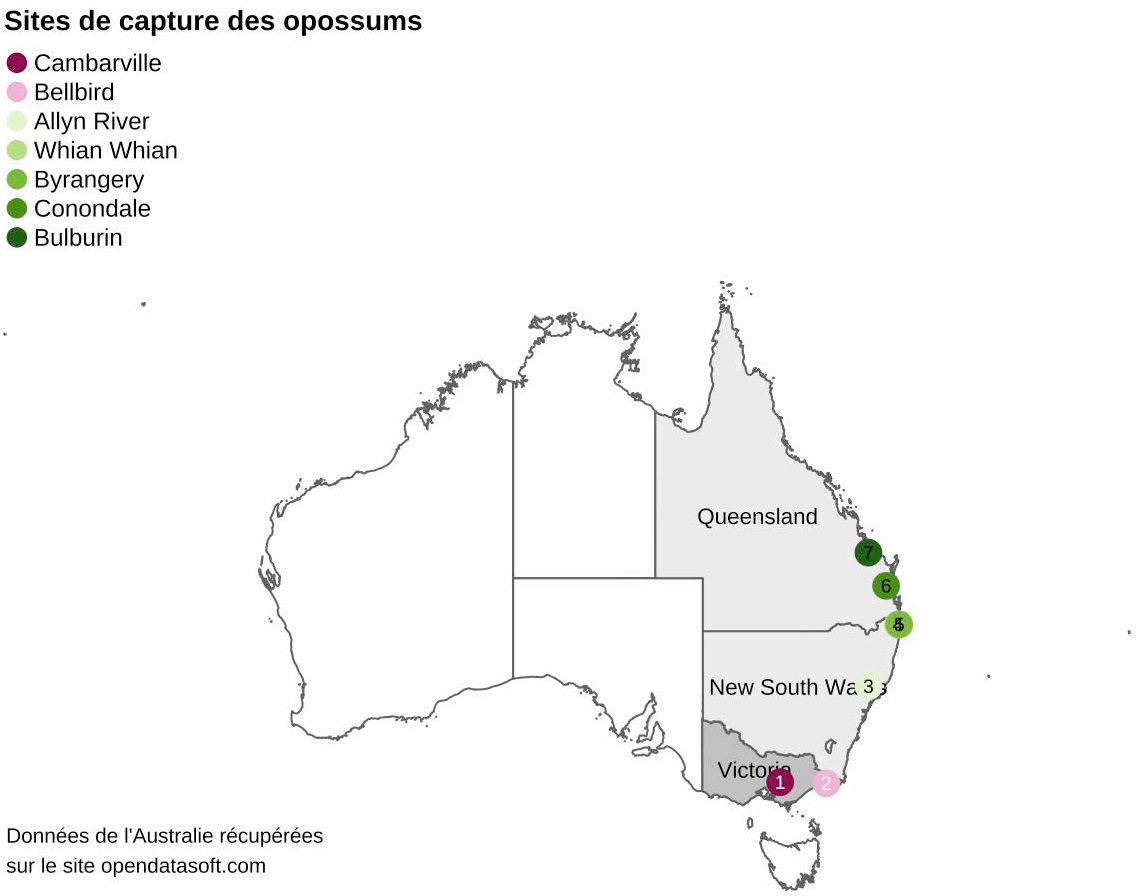
Enfin, pour que les sites apparaissent suivant l'ordre de leur numéro dans la légende de la carte, la variable `site_name` est convertie en facteur, dont les niveaux correspondent à l'ordre des numéros des sites.

```
# Affichage des noms des sites par ordre de numéro de site et non par ordre alphabétique
possumsites_regions$site_name <- factor(possumsites_regions$site_name,
                                        levels = row.names(possumsites))
levels(possumsites_regions$site_name)
```

```
## [1] "Cambarville" "Bellbird"    "Allyn River" "Whian Whian" "Byrangery"
## [6] "Conondale"   "Bulburin"
```

2.1.2 Création de la carte

```
tm_shape(australia) + tm_borders() +
  tm_credits("Données de l'Australie récupérées \nsur le site opendatasoft.com",
             position = c("left", "bottom"), size = 0.65) +
  tm_shape(victoria) + tm_polygons(col = "grey76") + tm_text("ste_name", size = 0.7) +
  tm_shape(nswales_queensland) + tm_polygons(col = "grey93") + tm_text("ste_name", size
= 0.7) +
  tm_shape(possumsites_regions) + tm_dots(col = "site_name", palette = sites_col, size =
0.8,
                                          title = "Sites de capture des opossums", title.siz
e = 0.8,) +
  tm_text("site_no", col = "region", palette = num_region_col, size = 0.6, fontface = "b
old",
          legend.col.show = FALSE) +
  tm_layout(frame = FALSE, legend.outside = TRUE, legend.outside.position = "top",
            legend.text.size = 0.8, legend.title.fontface = "bold")
```



La visualisation de la carte de l'Australie rend évident le regroupement des sites de Cambarville et Bellbird en une même région, plus australe que celle des cinq autres sites.

Cependant, il est difficile de distinguer Whian Whian de Byrangery, les sites 4 et 5 respectivement, car les deux points semblent superposés. Leur proximité est vérifiée à l'aide des coordonnées géographiques d'abord, puis par une visualisation rapide à l'aide de la bibliothèque `mapview` .

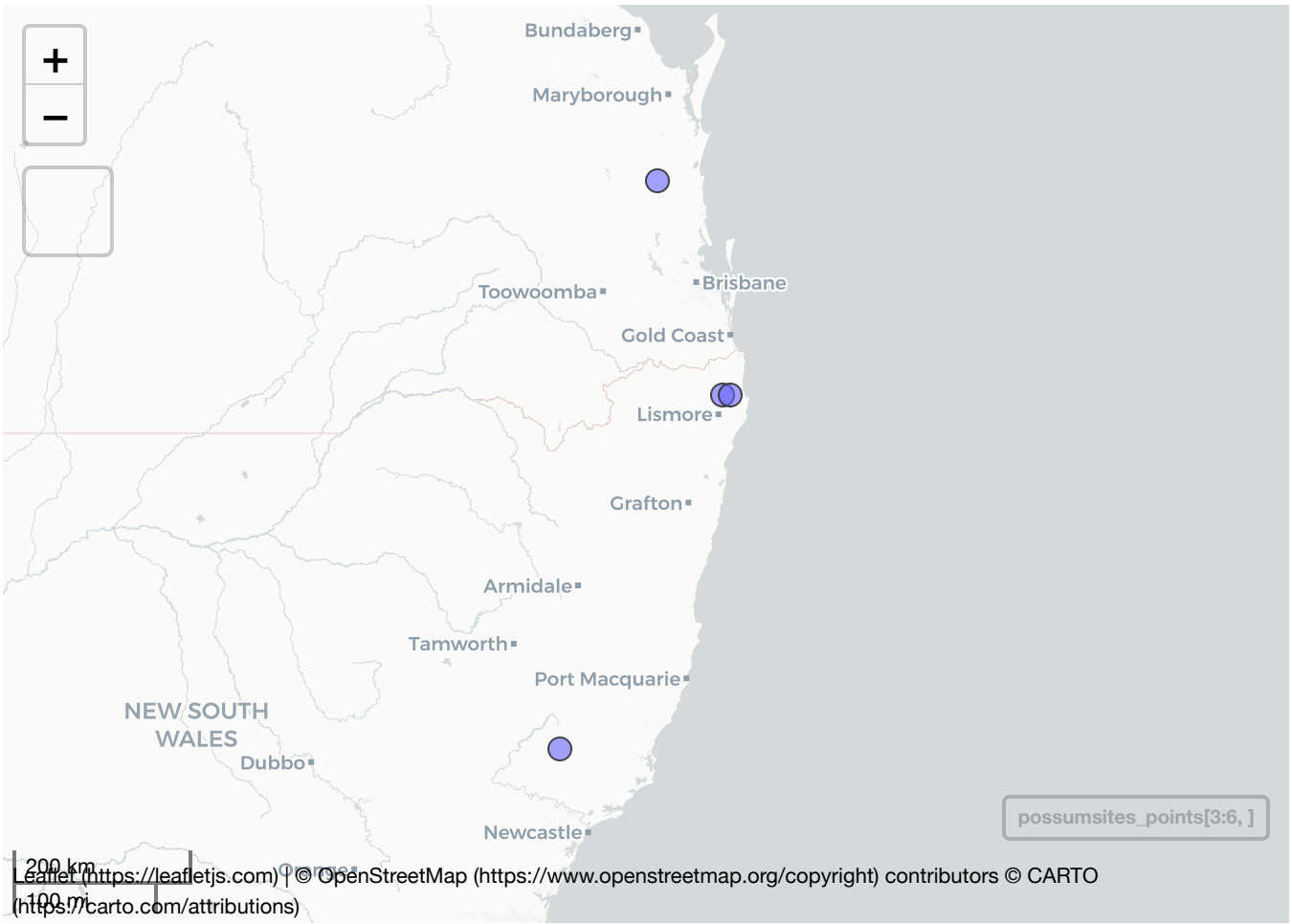
```
# Visualisation rapide des deux sites très rapprochés
possumsites_points$geometry[4:5]
```

```
## POINT (153.3333 -28.61667)
```

```
## POINT (153.4167 -28.61667)
```

```
mapview(possumsites_points[3:6, ], legend = FALSE)
```

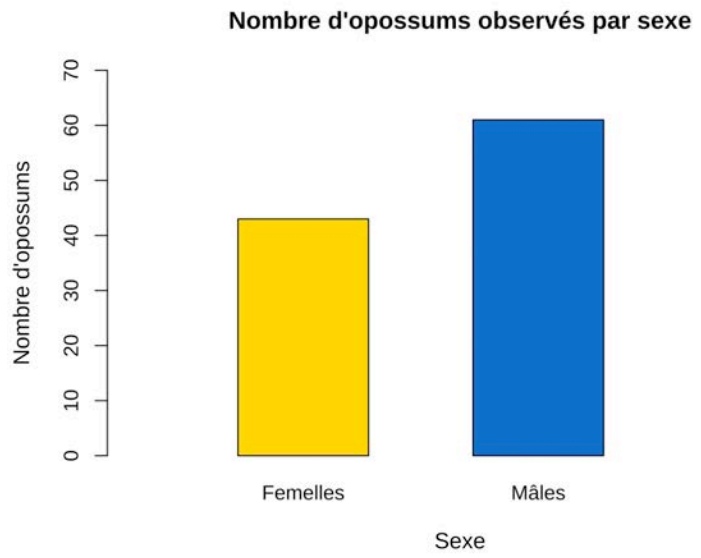
```
## Geometry set for 2 features
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 153.3333 ymin: -28.61667 xmax: 153.4167 ymax: -28.61667
## Geodetic CRS:  WGS 84
```



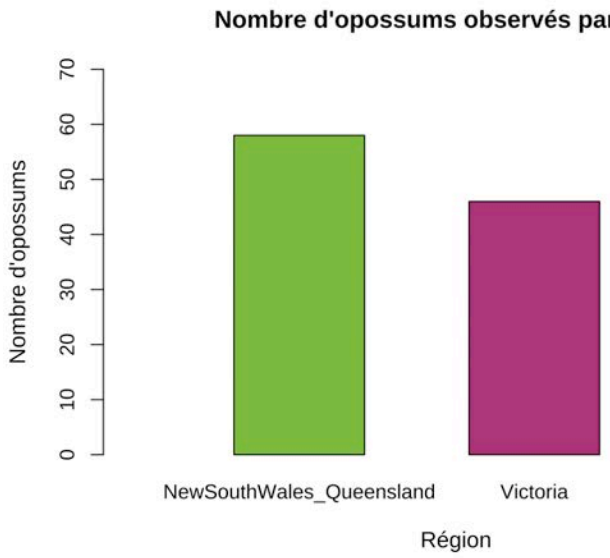
2.2 Visualisation des variables qualitatives et quantitatives discrètes

La répartition des variables qualitatives – soit le sexe et la région – et des variables quantitatives discrètes – le site et l'âge – est représentée graphiquement par des diagrammes en bâtons.

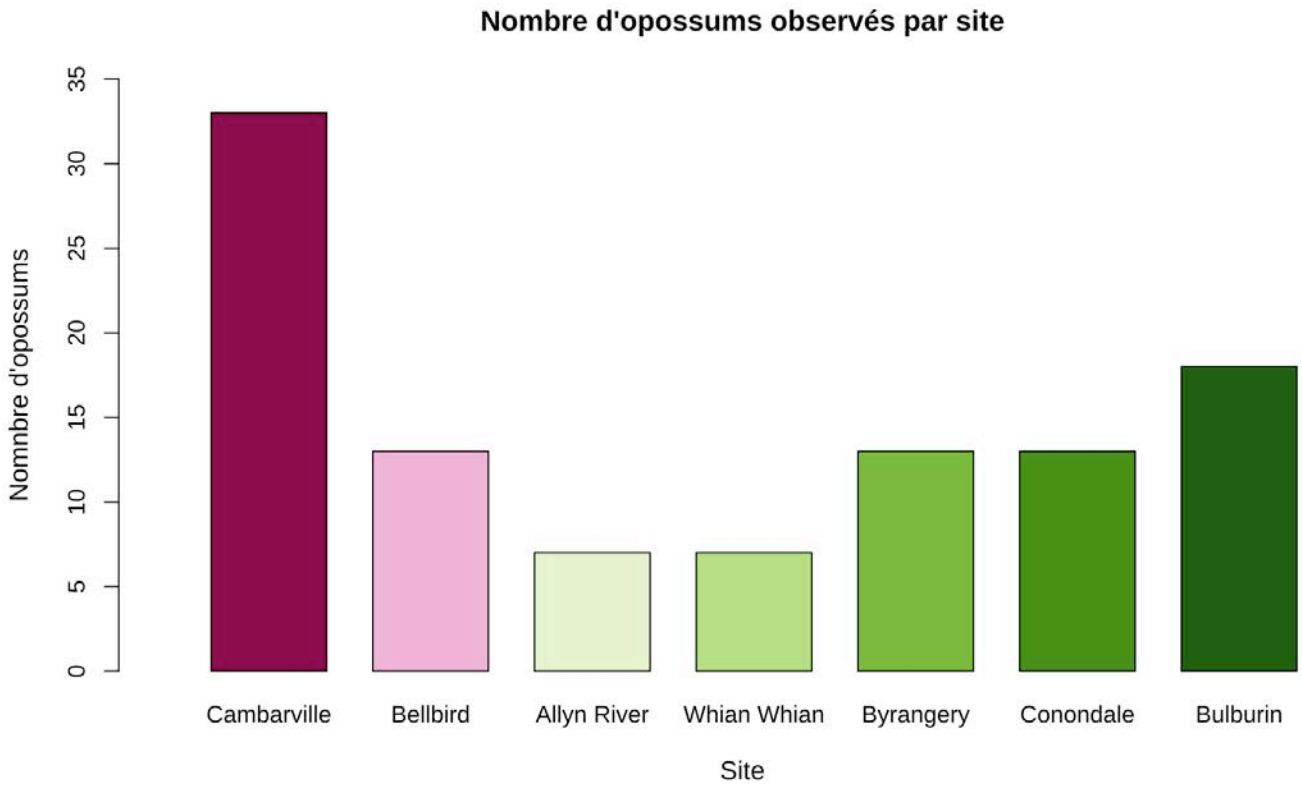
```
# Variable "sex"
barplot(table(possum$sex), main = "Nombre d'opossums observés par sexe",
        cex.main = 1.2,
        names.arg = c("Femelles", "Mâles"),
        xlab = "Sexe", ylab = "Nombre d'opossums", ylim = c(0, 70),
        cex.lab = 1.1, col = sex_col,
        width = 0.2, space = 0.8, xlim = c(0, 1))
```



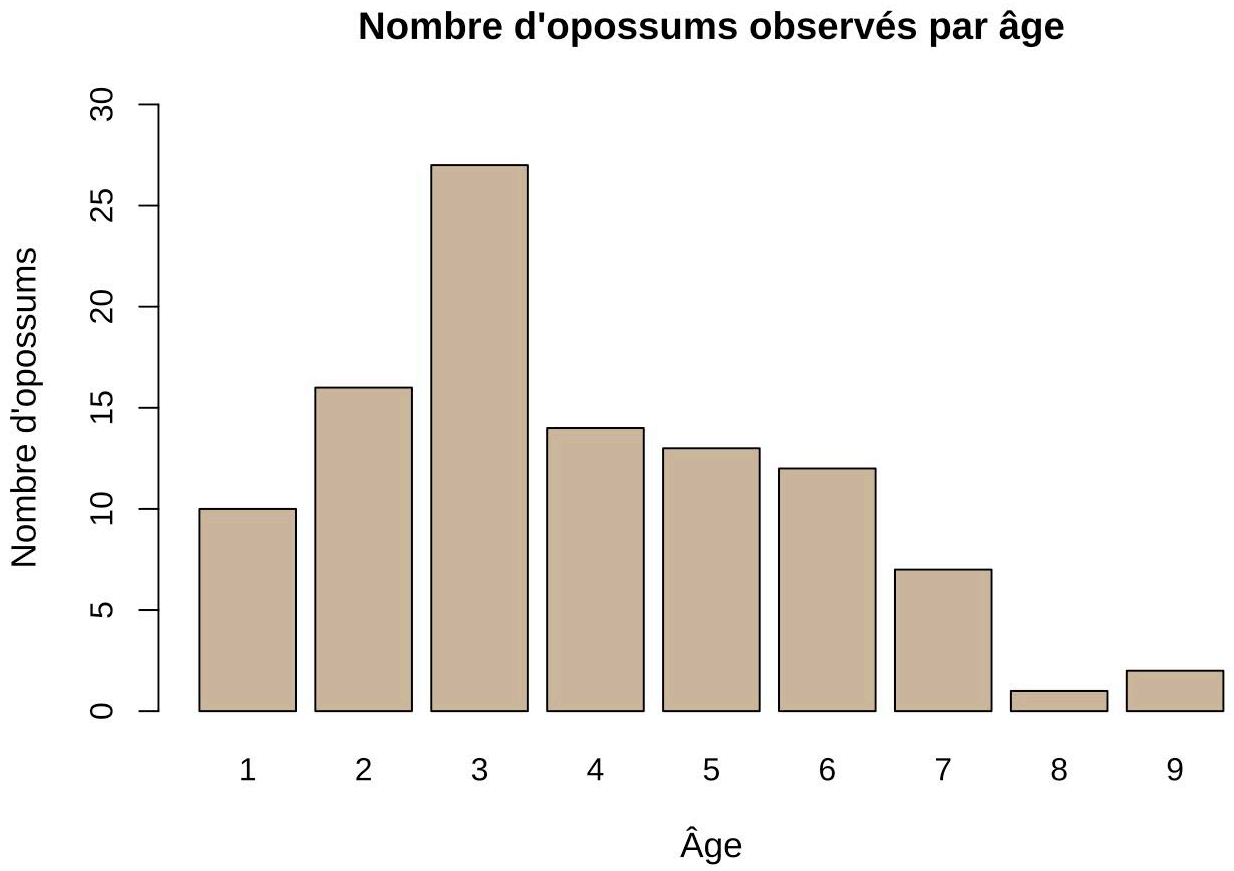

```
# Variable "region"
barplot(table(possum$region), main = "Nombre d'opossums observés par région",
        cex.main = 1.2,
        xlab = "Région", ylab = "Nombre d'opossums", ylim = c(0, 70),
        cex.lab = 1.1, col = region_col,
        width = 0.2, space = 0.8, xlim = c(0, 1))
```




```
# Variable "site"
barplot(table(possum$site), main = "Nombre d'opossums observés par site", cex.main = 1.
2,
        names.arg = row.names(possumsites),
        xlab = "Site", ylab = "Nomnbre d'opossums", ylim = c(0, 35),
        cex.lab = 1.1, width = 0.7, xlim = c(0, 7), space = 0.4,
        col = sites_col)
```



```
# Variable "age"
barplot(table(possum$age), main = "Nombre d'opossums observés par âge",
        cex.main = 1.2,
        xlab = "Âge", ylab = "Nombre d'opossums", ylim = c(0, 30),
        cex.lab = 1.1, col = "bisque3")
```

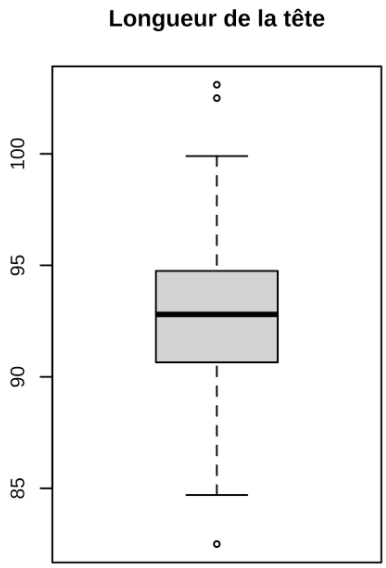


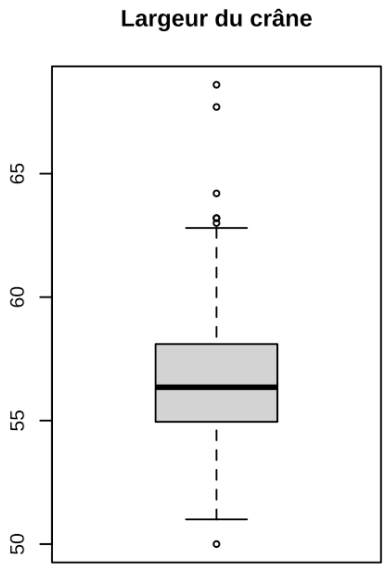
2.3 Visualisation des variables quantitatives continues

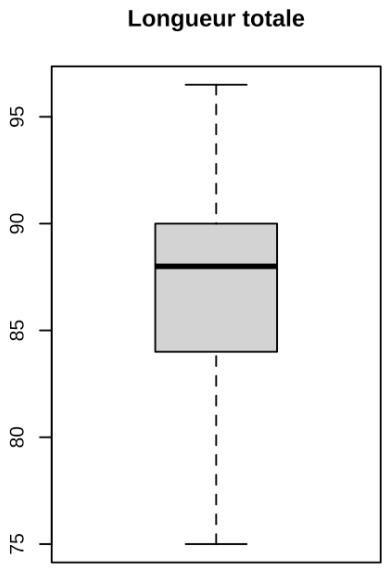
La répartition des variables quantitatives continues est représentée à l'aide de diagrammes en boîte.

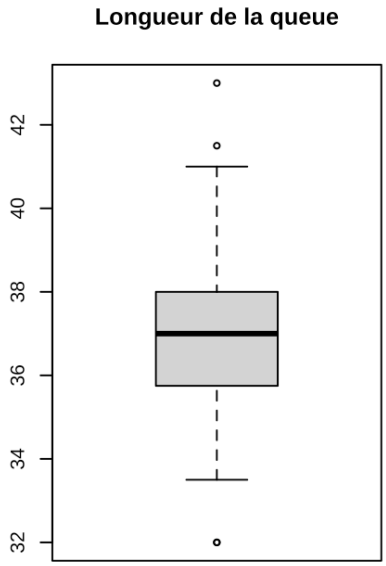
```
# Création d'un vecteur des noms de variables en français
variables <- c("Longueur de la tête", "Largeur du crâne", "Longueur totale",
               "Longueur de la queue", "Longueur du pied", "Longueur de la conque de l'o
reille",
               "Largeur de l'œil", "Circonférence de la poitrine", "Circonférence du ven
tre")
```

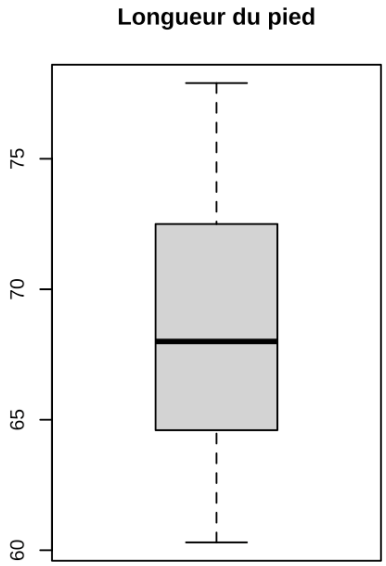
```
# Représentation des 9 variables
par(mfrow = c(3, 3))
for (i in 1:9) {
    boxplot(possum[ , i + 4], main = variables[i])
}
```

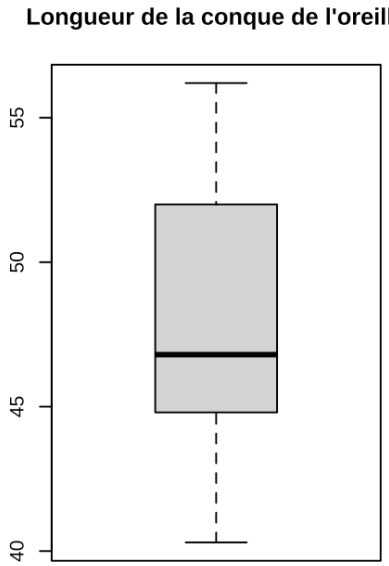

Longueur de la tête
85
90
95
100

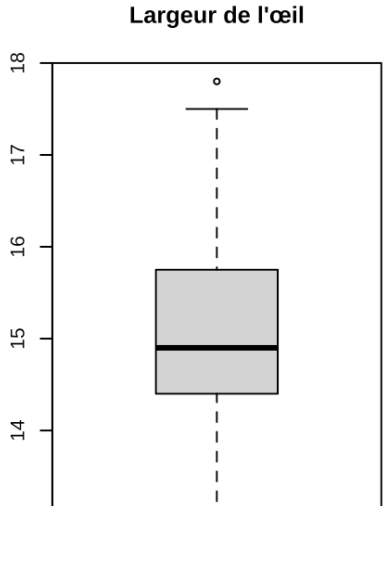

Largeur du crâne
50
55
60
65

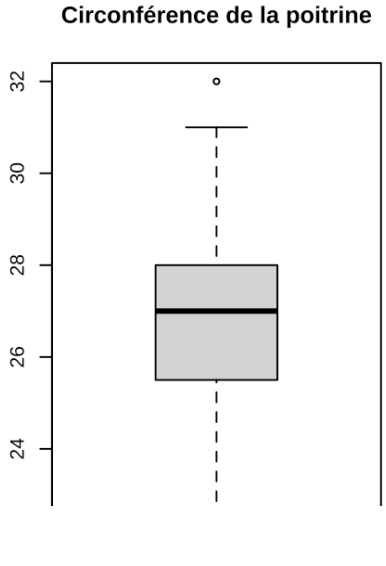

Longueur totale
75
80
85
90
95

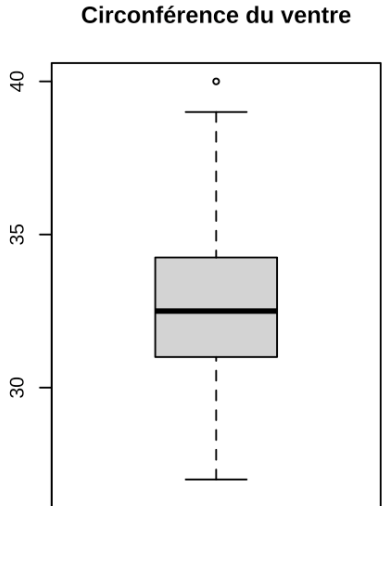

Longueur de la queue
32
34
36
38
40
42

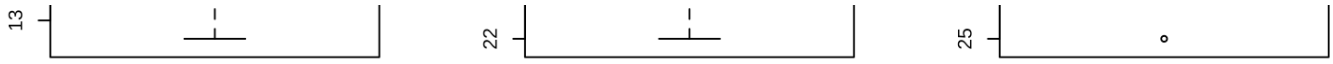

Longueur du pied
60
65
70
75


Longueur de la conque de l'oreille
40
45
50
55


Largeur de l'œil
14
15
16
17
18


Circonférence de la poitrine
24
26
28
30
32


Circonférence du ventre
30
35
40

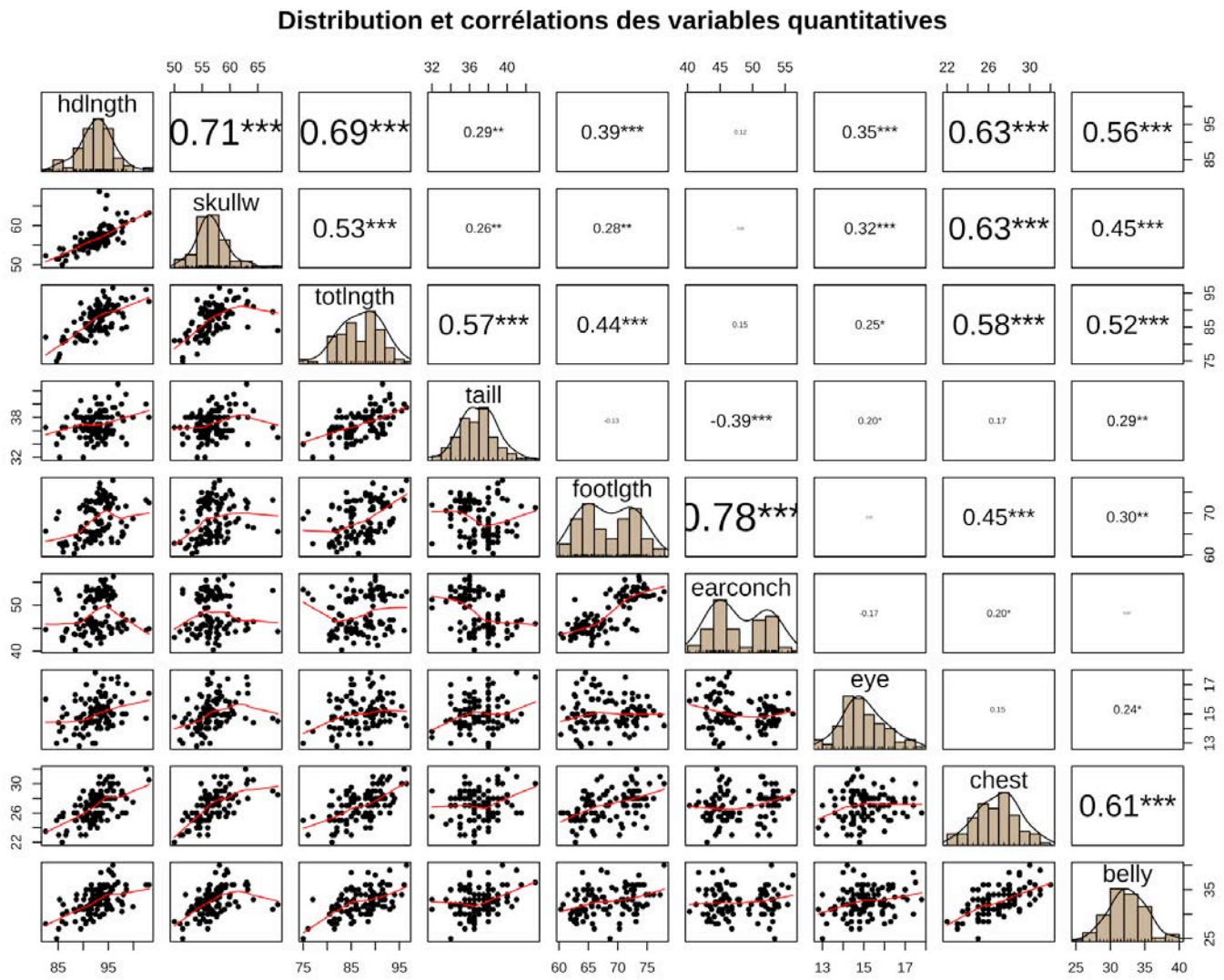


2.4 Visualisation des relations entre les variables quantitatives continues

Pour avoir une meilleure idée des données morphométriques dans leur ensemble, les variables quantitatives continues sont représentées par paires, sur le même graphique.

```
library(psych)
```

```
par(mfrow = c(1, 1))
pairs.panels(possum[5:13],
             ellipses = FALSE, method = "pearson",
             hist.col = "bisque3", cex.cor = 1.8,
             scale = TRUE, stars = TRUE,
             main = "Distribution et corrélations des variables quantitatives")
```



Ce graphique permet de visualiser rapidement la distribution de chaque variable et de faire ressortir certaines corrélations, notamment entre la longueur du pied et la longueur de la conque de l'oreille, entre la longueur de la tête et la largeur du crâne, et entre la longueur de la tête et la longueur totale.

---

3. Analyse des données

3.1 Analyse d'un dimorphisme sexuel

L'objectif de cette section est de déterminer s'il existe un dimorphisme sexuel chez les opossums observés. Plus précisément, l'hypothèse selon laquelle la circonférence du ventre serait plus grande chez les femelles en raison de leur marsupium sera étudiée.

3.1.1 Portrait moyen d'un opossum en fonction du sexe

Les moyennes de chaque variable sont comparées entre les femelles et les mâles.

```
possum_means <- possum |>
  group_by(sex) |>
  summarise(across(hdlngth:belly, ~ mean(.x, na.rm = TRUE)))

possum_means
```

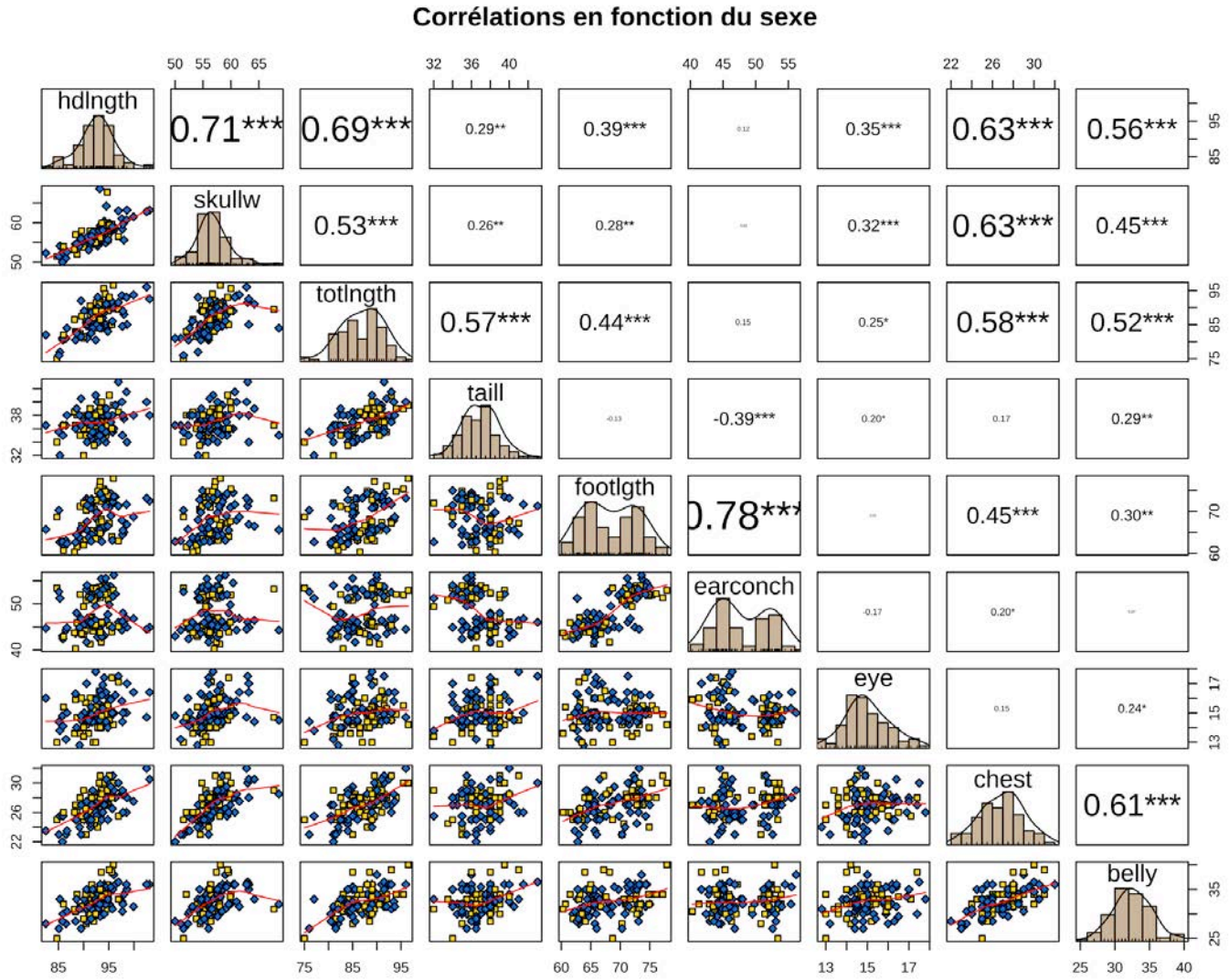
```
## # A tibble: 2 × 10
##   sex    hdlngth skullw totlngth taill footlgth earconch   eye chest belly
##   <fct>    <dbl>  <dbl>    <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1 female    92.1   56.6     87.9  37.1     69.1     48.6  14.8  27.3  32.9
## 2 male      92.9   57.1     86.5  36.9     68.0     47.8  15.2  26.8  32.4
```

Ce tableau ne suggère aucun dimorphisme sexuel, car les moyennes diffèrent très peu entre les femelles et les mâles.

3.1.2 Visualisation des données en fonction du sexe

De possibles corrélations entre les caractéristiques morphologiques et le sexe sont explorées à l'aide d'une représentation graphique, où les données des femelles apparaissent en jaune, et celles des mâles, en bleu.

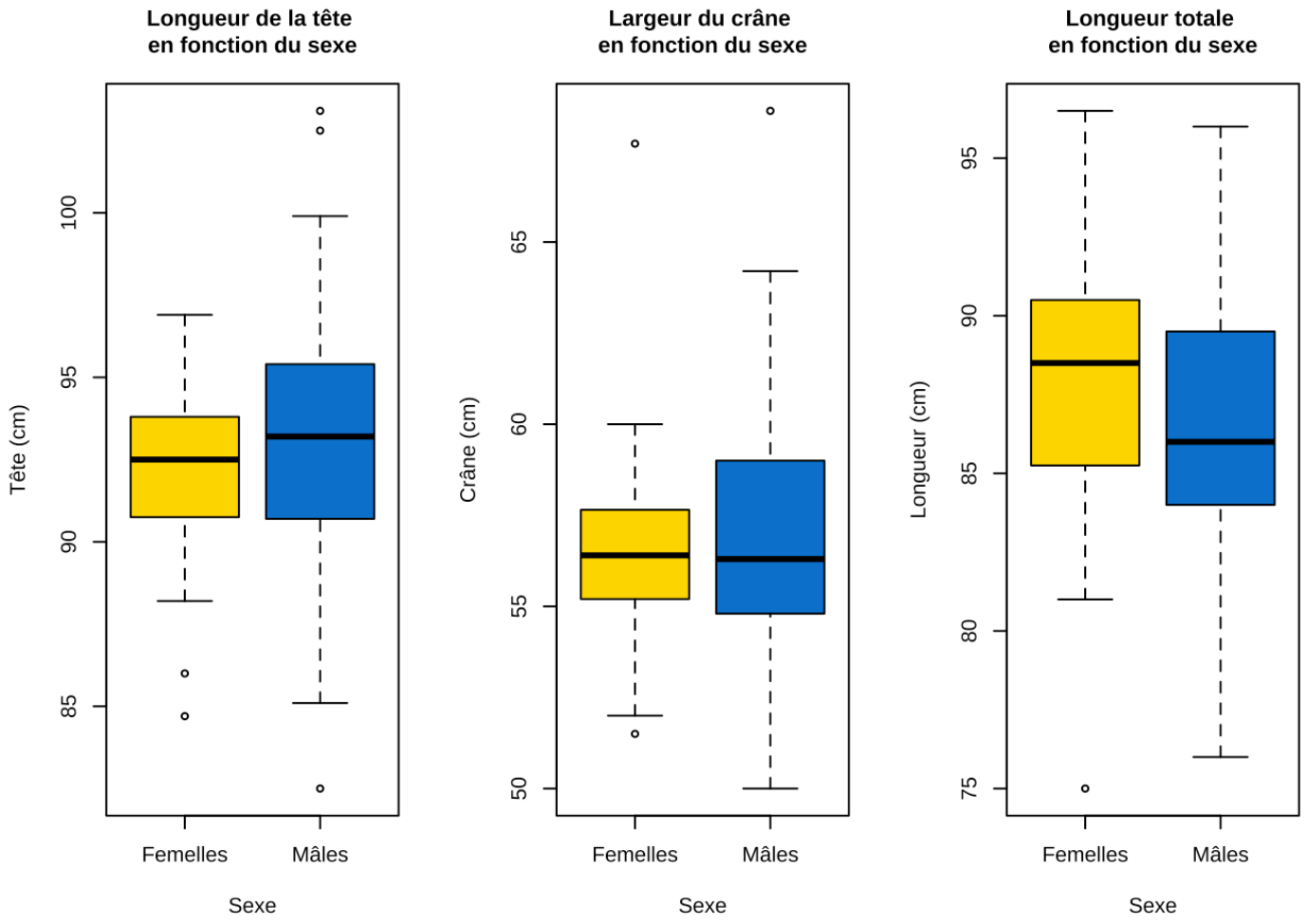
```
pairs.panels(possum[5:13],
             ellipses = FALSE, method = "pearson",
             hist.col = "bisque3", cex.cor = 1.8,
             scale = TRUE, stars = TRUE,
             pch = 21 + as.numeric(possum$sex),
             bg = sex_col[possum$sex],
             main = "Corrélations en fonction du sexe")
```



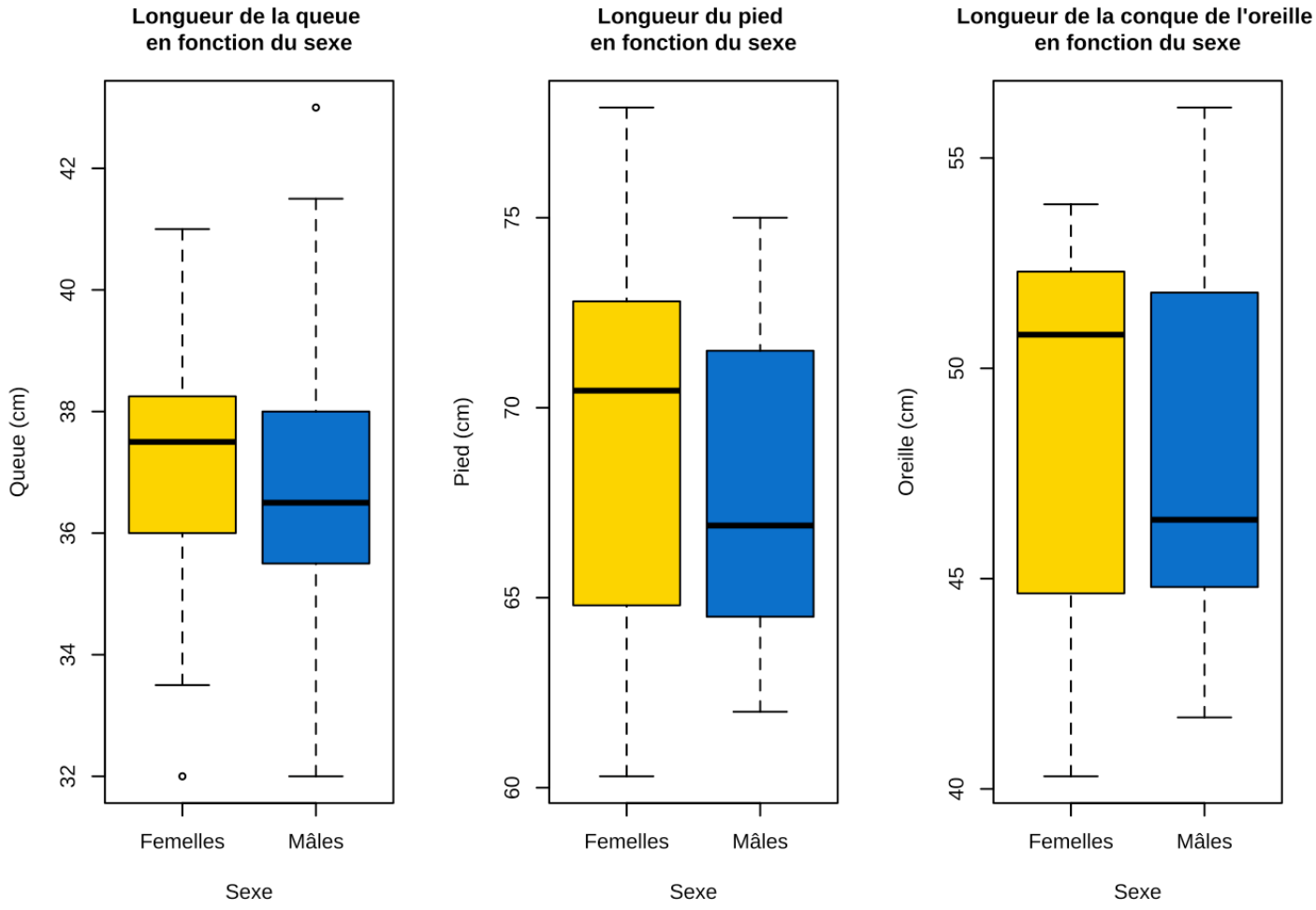
Ce graphique ne permet pas de repérer de corrélation évidente entre le sexe et l'une ou l'autre des caractéristiques morphologiques. La répartion des variables quantitatives continues en fonction du sexe est donc visualisée séparément pour chaque variable.

```
# Définition d'une fonction
box_graph_sex <- function(data, variable, title, ylab) {
  boxplot(variable ~ data$sex, names = c("Femelles", "Mâles"),
          col = sex_col,
          main = paste(title, "\nen fonction du sexe"),
          ylab = paste(ylab, "(cm)"), xlab = "Sexe", cex.main = 1.05)
}
```

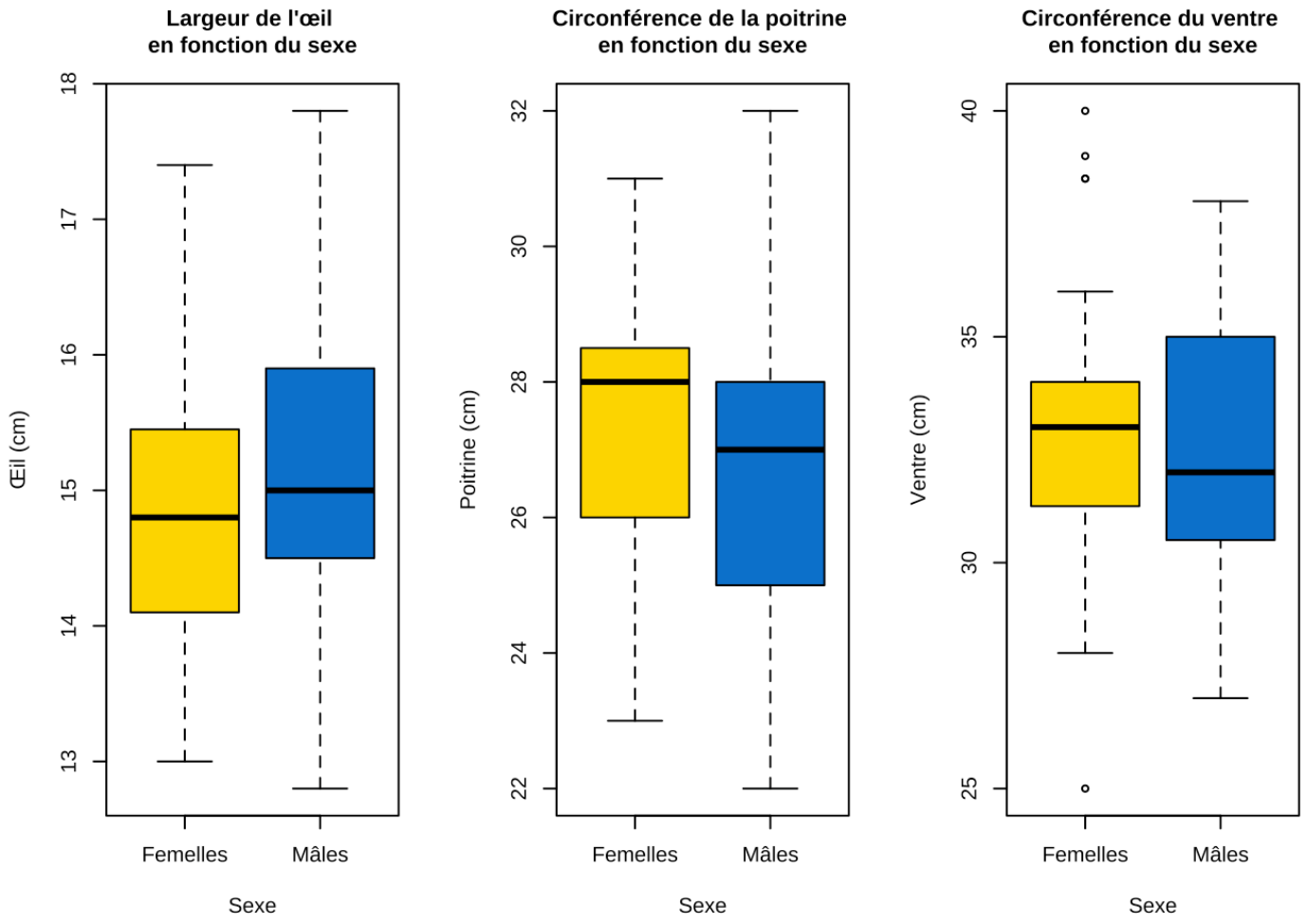
```
# Diagrammes en boîte pour chaque variable en fonction du sexe
par(mfrow = c(1, 3))
box_graph_sex(possum, possum$hdlngth, variables[1], "Tête")
box_graph_sex(possum, possum$skullw, variables[2], "Crâne")
box_graph_sex(possum, possum$totlngth, variables[3], "Longueur")
```




```
box_graph_sex(possum, possum$taill, variables[4], "Queue")
box_graph_sex(possum, possum$footlgth, variables[5], "Pied")
box_graph_sex(possum, possum$earconch, variables[6], "Oreille")
```




```
box_graph_sex(possum, possum$eye, variables[7], "Œil")
box_graph_sex(possum, possum$chest, variables[8], "Poitrine")
box_graph_sex(possum, possum$belly, variables[9], "Ventre")
```



Malgré quelques variations, la visualisation des données ne permet toujours pas de croire qu'il existe un dimorphisme sexuel chez les opossums observés.

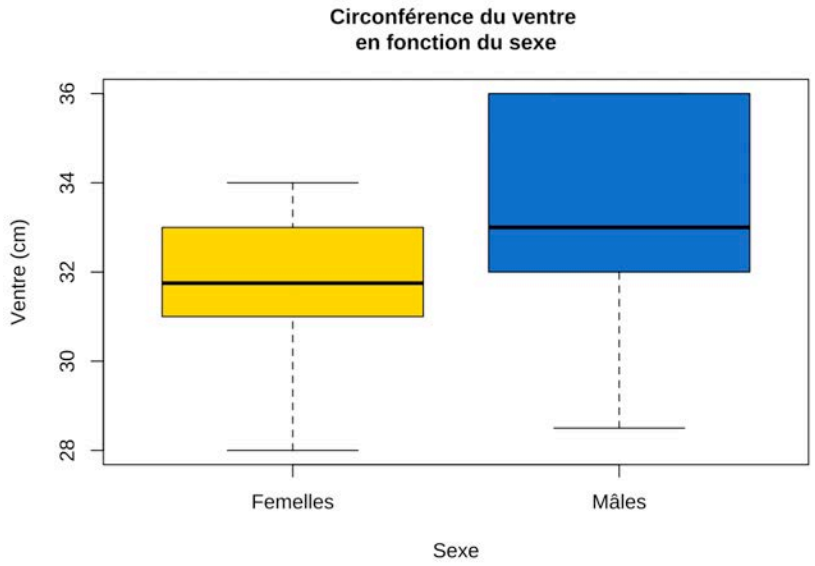
3.1.3 Opossums en âge de reproduction

L'hypothèse de départ étant de comparer la taille des ventres des femelles et des mâles en raison de la poche marsupiale des femelles, une nouvelle question se pose : une différence de la circonférence du ventre est-elle observable seulement pour les opossums en âge de reproduction?

D'après le site *Animal Diversity Web*, une base de données de l'Université du Michigan sur la biologie des animaux, l'âge de maturité sexuelle du *Trichosurus caninus* (https://animaldiversity.org/accounts/Trichosurus_caninus/) est de 2 à 3 ans, et celui du *Trichosurus cunninghami* (https://animaldiversity.org/accounts/Trichosurus_cunninghami/) est aussi de 2 à 3 ans pour les mâles, et de 2 à 5 ans pour les femelles. Les observations qui correspondent à des opossums de 2 et 3 ans sont donc isolées et la circonférence du ventre en fonction du sexe est visualisée à l'aide d'un diagramme en boîte.

```
# Sous-ensemble des opossums en âge de reproduction
sex_maturity <- subset(possum, age == c(2, 3), na.rm = TRUE)
```

```
# Visualisation de la variable "belly"
box_graph_sex(sex_maturity, sex_maturity$belly, variables[9], "Ventre")
```



Il ressort de ce diagramme que les mâles ont le ventre un peu plus gros que les femelles. Comme cela est possiblement dû au fait que les mâles en âge de reproduction sont globalement plus gros que les femelles, la moyenne de chaque caractéristique morphologique est comparée entre les deux groupes.

```
# Comparaison de la taille globale des opossums de 2 et 3 ans
sm_means <- sex_maturity |>
  group_by(sex) |>
  summarise(across(hdlngth:belly, ~ mean(.x, na.rm = TRUE)))
sm_means
```

```
## # A tibble: 2 × 10
##   sex    hdlngth skullw totlngth taill footlgth earconch   eye chest belly
##   <fct>    <dbl>  <dbl>    <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1 female    91.3   56.0     86    36.6     67.7     47.7  14.7  26.9  31.7
## 2 male      95.2   58.7     89.1  37.7     69.4     47.7  15.2  27.8  33.5
```

Effectivement, les mâles de 2 et 3 ans sont en moyenne globalement plus gros que les femelles du même âge. L'âge de la maturité sexuelle ne donne donc aucune information supplémentaire par rapport à l'hypothèse. De plus, comme le nombre d'observations est peu élevé, l'analyse statistique sera effectuée sur toutes les observations, sans tenir compte de l'âge.

3.1.4 Analyse statistique

L'analyse statistique servira à tester les hypothèses statistiques suivantes concernant les moyennes de la circonférence du ventre des deux sexes :

$H_0 : \mu_{\text{ventre femelles}} \leq \mu_{\text{ventre mâles}}$

$H_a : \mu_{\text{ventre femelles}} > \mu_{\text{ventre mâles}}$

$\alpha = 0.05$

L'analyse se fera à l'aide d'un test *t*.

Préparation des données pour l'analyse

L'exploration initiale des données avait montré la présence probable de données aberrantes dans la variable `belly` . La méthode de l'écart-type est utilisée pour les détecter.

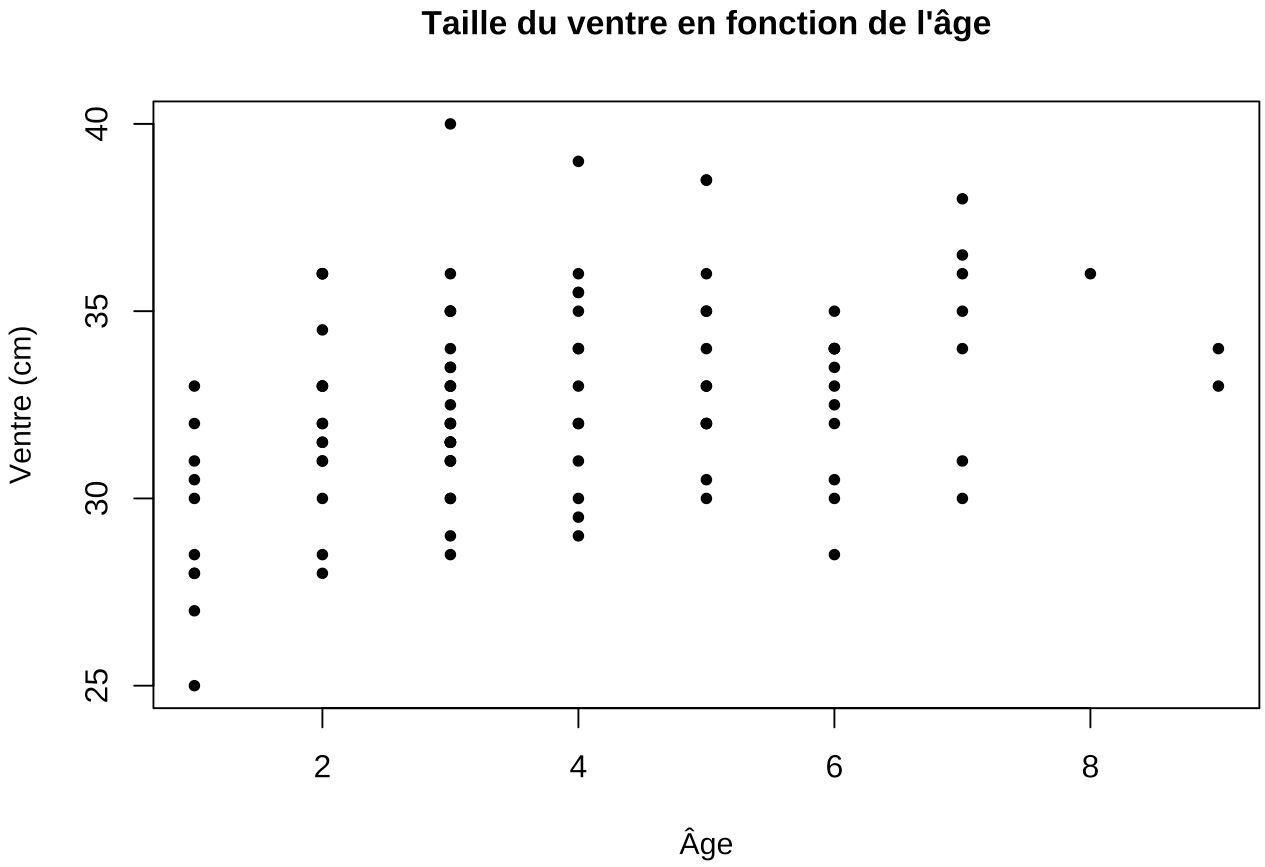
```
# Détection des valeurs aberrantes avec la méthode de l'écart-type
lower_belly <- mean(possum$belly) - (1.5 * sd(possum$belly))
upper_belly <- mean(possum$belly) + (1.5 * sd(possum$belly))
print(paste("Limite inférieure :", round(lower_belly, digits = 2)))
print(paste("Limite supérieure : ", round(upper_belly, digits = 2)))

index_belly_outliers <- which(possum$belly < lower_belly | possum$belly > upper_belly)
print(paste("Nombre de valeurs hors de l'écart-type :", length(index_belly_outliers)))
```

```
## [1] "Limite inférieure : 28.44"
## [1] "Limite supérieure :  36.73"
## [1] "Nombre de valeurs hors de l'écart-type : 11"
```

Cette méthode détecte 11 valeurs en dehors de l'écart type, ce qui est beaucoup plus que les 2 valeurs identifiées par la fonction `diagnose` . Une visualisation des données permettra d'avoir une meilleure idée de la distribution des valeurs.

```
plot(belly ~ age, data = possum, main = "Taille du ventre en fonction de l'âge",
     pch = 20, xlab = "Âge", ylab = "Ventre (cm)", cex.lab = 0.95, cex.main = 1.05)
```



Le graphique montre que les valeurs sont en fait assez rapprochées. Par conséquent, seules la valeur minimale et la valeur maximale sont remplacées par la moyenne.

```
# Identification des deux valeurs extrêmes
which.min(possum$belly)
which.max(possum$belly)
```

```
## [1] 39
## [1] 21
```

```
# Remplacement des valeurs extrêmes par la moyenne
possum[c(21, 39), 13]
possum[c(21, 39), 13] <- mean(possum$belly)
possum[c(21, 39), 13]
```

```
## [1] 40 25
## [1] 32.58654 32.58654
```

Les données sont maintenant séparées par sexe, en ne conservant que la variable d'intérêt pour l'analyse statistique.

```
females <- possum[possum$sex == "female", "belly"]
males <- possum[possum$sex == "male", "belly"]
print(paste("Nombre de femelles :", length(females)))
print(paste("Nombre de mâles :", length(males)))
```

```
## [1] "Nombre de femelles : 43"
## [1] "Nombre de mâles : 61"
```

Les deux groupes obtenus sont de taille inégale, avec un ratio approximatif de trois mâles pour deux femelles.

Vérification des suppositions

Le test *t* sur deux groupes indépendants requiert l'indépendance des observations, la normalité des résidus et l'homogénéité des variances entre les deux groupes.

Toutes les mesures morphologiques ont été prélevées sur des opossums différents, capturés au hasard sur différents sites australiens. L'indépendance des observations est donc respectée.

La normalité des résidus et l'homoscédasticité sont vérifiées à l'aide de méthodes graphiques, ainsi qu'avec un test formel pour la normalité des résidus.

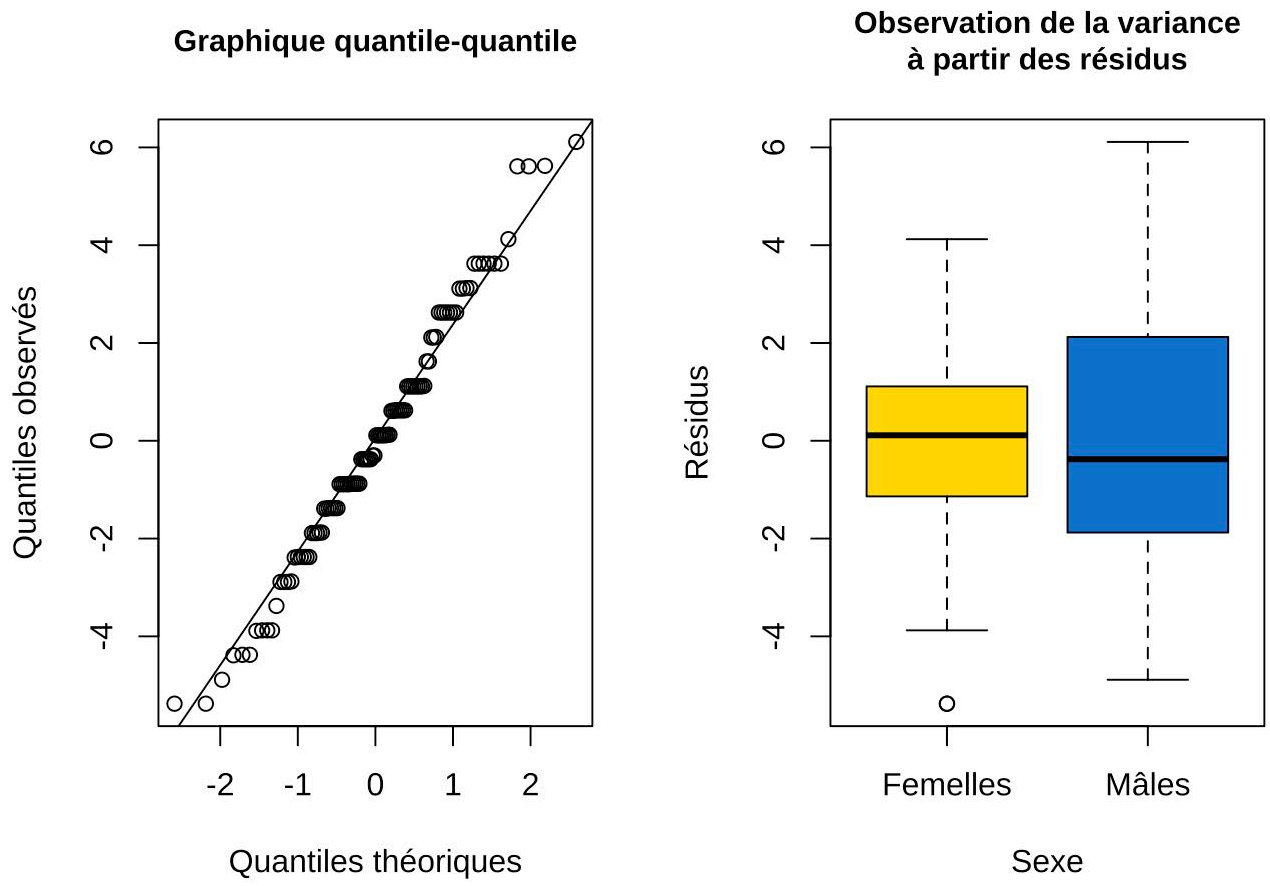
```
# Calcul des résidus
res_females <- females - mean(females)
res_males <- males - mean(males)
res_sex <- c(res_females, res_males)

# Test formel Anderson-Darling
library(nortest)
ad.test(res_sex)
```

```
##
##  Anderson-Darling normality test
##
## data:  res_sex
## A = 0.39588, p-value = 0.3648
```

```
# Graphique quantile-quantile
par(mfrow = c(1, 2))
qqnorm(res_sex, xlab = "Quantiles théoriques", ylab = "Quantiles observés",
       main = "Graphique quantile-quantile", cex.main = 0.95)
qqline(res_sex)

# Homoscédasticité
possum$res_sex <- res_sex
boxplot(res_sex ~ sex, data = possum, names = c("Femelles", "Mâles"),
        col = sex_col,
        main = "Observation de la variance\nà partir des résidus",
        ylab = "Résidus", xlab = "Sexe", cex.main = 0.95)
```



Le test d'Anderson-Darling, avec une valeur de $P = 0{,}3648$, et le graphique quantile-quantile indiquent que les résidus suivent une distribution normale. Cependant, les variances sont hétérogènes, possiblement parce que les deux groupes sont inégaux en taille et le nombre d'observations, peu élevé. Par conséquent, l'analyse se fera avec un test *t* de Welch.

```
t.test(x = females, y = males, data = possum, var.eqal = FALSE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  females and males
## t = 1.0181, df = 95.01, p-value = 0.1556
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3225532        Inf
## sample estimates:
## mean of x mean of y
##  32.88775  32.37705
```

Interprétation des résultats

Avec $P(t_{95} \geq 1{,}02) = 0{,}1556$, la probabilité d'observer une valeur de t =1,02 (ou plus grande) dans une population où H_0 est vraie est suffisamment élevée pour ne pas rejeter l'hypothèse nulle. Ainsi, il n'est pas possible de conclure que la taille du ventre des femelles est plus grande que celle du ventre des mâles en raison de leur marsupium.

Par ailleurs, d'après la visualisation exploratoire des données, aucune autre caractéristique morphologique ne semble varier significativement en fonction du sexe. Par conséquent, aucune autre analyse statistique ne sera effectuée concernant le dimorphisme sexuel.

3.2 Analyse d'un dimorphisme géographique

L'objectif de cette section est de confirmer qu'il existe un dimorphisme géographique chez les opossums observés, c'est-à-dire des traits morphologiques qui diffèrent entre les individus de la région de Victoria et celle de New South Wales et Queensland, tel que l'ont conclu les chercheurs Lindenmayer et son équipe.

3.2.1 Portrait moyen d'un opossum en fonction de la région

Les moyennes de chaque variable sont comparées entre les opossums des deux régions.

```
region_means <- possum |>
  group_by(region) |>
  summarise(across(hdlngth:belly, ~ mean(.x, na.rm = TRUE)))

region_means
```

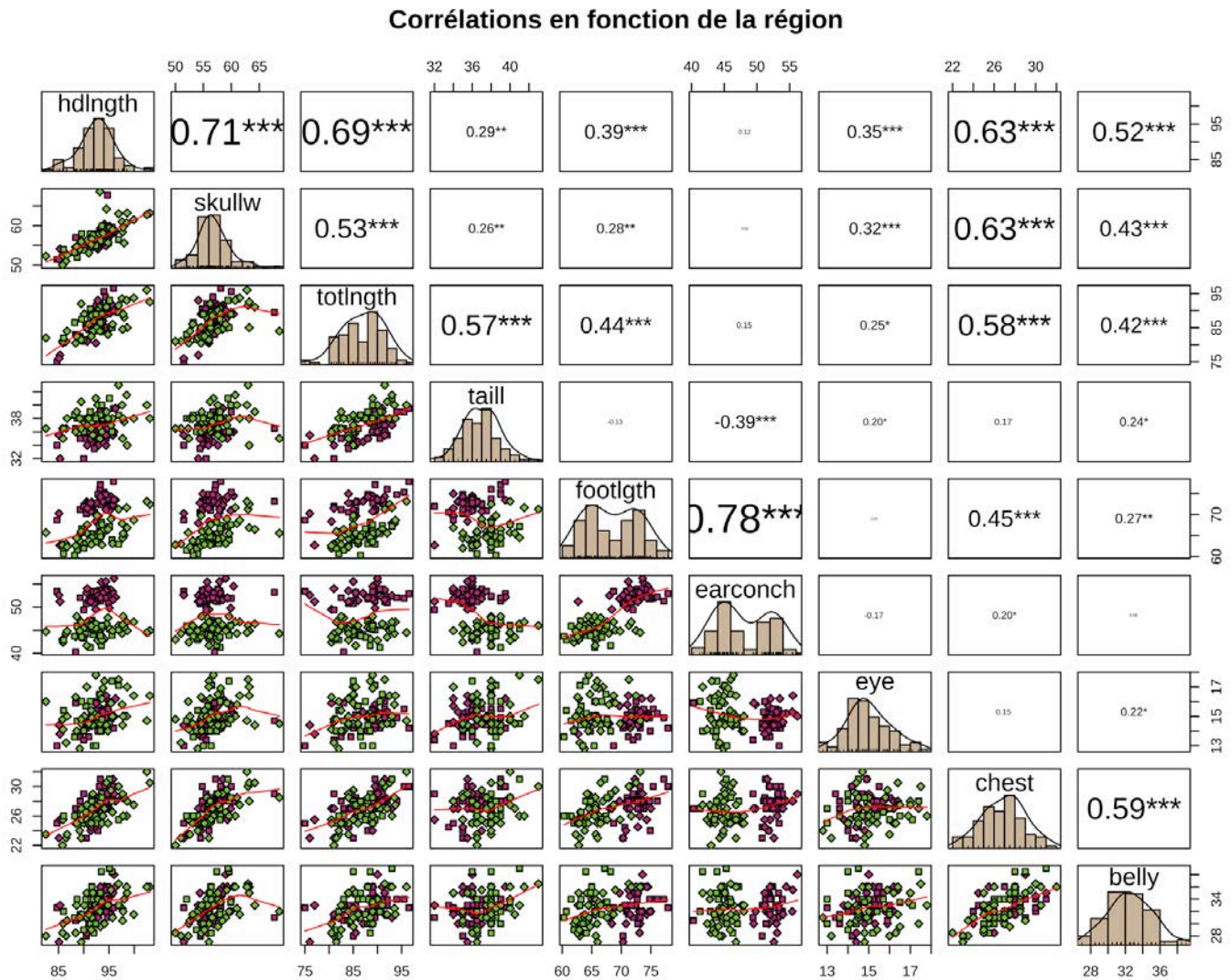
```
## # A tibble: 2 × 10
##   region       hdlngth skullw totlngth taill footlgth earconch   eye chest belly
##   <fct>          <dbl>  <dbl>    <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1 NewSouthWal…    92.6   57.1     86.8  37.9     65.4     44.9  15.2  26.6  32.5
## 2 Victoria        92.6   56.7     87.5  35.9     72.4     52.2  14.9  27.4  32.7
```

La comparaison des moyennes de chaque variable entre les régions fait ressortir des écarts de taille pour la longueur du pied et la conque de l'oreille, avec des différences de moyennes de 7 cm et 7,3 cm respectivement. Dans une moindre mesure, la moyenne de la longueur de la queue varie de 2 cm entre les deux groupes. Pour toutes les autres variables, la différence est minime, voire inexistante, variant entre 0 et 1 cm.

3.2.2 Visualisation des données en fonction de la région

De possibles corrélations entre les caractéristiques morphologiques et la région sont maintenant explorées à l'aide d'une représentation graphique, où les données de la population de Victoria sont représentées en rose, et celles de la population de New South Wales et Queensland, en vert.

```
pairs.panels(possum[5:13],
             ellipses = FALSE, method = "pearson",
             hist.col = "bisque3", cex.cor = 1.8,
             scale = TRUE, stars = TRUE,
             pch = 21 + as.numeric(possum$sex),
             bg = region_col[possum$region],
             main = "Corrélations en fonction de la région")
```

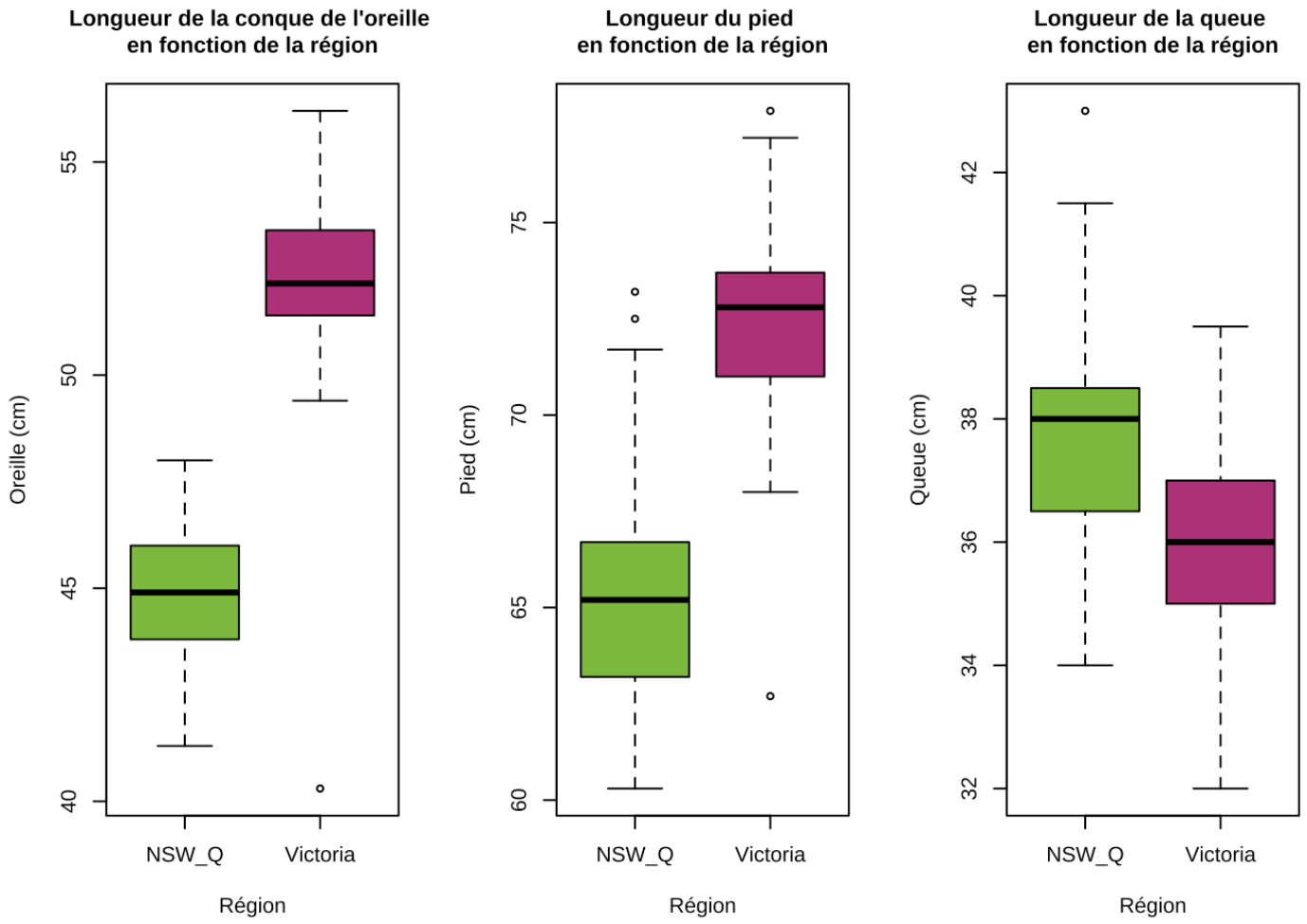


Cette visualisation permet de corroborer des différences marquées entre les deux populations d'opossums. Comme dans le tableau des moyennes, la longueur du pied et la longueur de la conque de l'oreille sont les deux caractéristiques morphologiques pour lesquelles les deux groupes semblent les plus distincts. Le graphique fait d'ailleurs ressortir une forte corrélation entre ces deux variables (corrélation déjà identifiée à la section 2.4). Encore une fois, la différence pour la longueur de la queue est moins évidente, mais elle est tout de même suggérée.

La répartition des trois caractéristiques morphologiques mises en évidence précédement est représentée graphiquement selon la région, sous forme de diagrammes en boîte.

```
# Définition d'une fonction
box_graph_rg <- function(data, variable, title, ylab) {
  boxplot(variable ~ data$region,
          col = region_col,
          names = c("NSW_Q", "Victoria"),
          main = paste(title, "\nen fonction de la région"),
          ylab = paste(ylab, "(cm)"), xlab = "Région", cex.main = 1.05)
}
```

```
# Visualisation des trois variables d'intérêt en fonction de la région
par(mfrow = c(1, 3))
box_graph_rg(possum, possum$earconch, variables[6], "Oreille")
box_graph_rg(possum, possum$footlgth, variables[5], "Pied")
box_graph_rg(possum, possum$taill, variables[4], "Queue")
```



Si les opossums de la région de Victoria ont de plus grandes oreilles et de plus longs pieds que leurs voisins de New South Wales et Queensland, ils ont en revanche la queue un peu plus courte. Par ailleurs, il est intéressant de noter qu'une valeur de longueur de la conque de l'oreille est nettement inférieure aux autres pour la région de Victoria.

3.2.3 Analyses statistiques

Des test d'hypothèse sur deux groupes seront effectuées pour vérifier si la différence entre les opossums des deux régions est significative, et non seulement due à l'échantillon, pour les trois caractéristiques identifiées par la visualisation des données.

Les hypothèses statistiques sont les suivantes, et seront testées séparément pour chaque caractéristique :

$H_0 : \mu_{\text{oreillle/pied/queue Victoria}} = \mu_{\text{oreille/pied/queue NSW\_Queensland}}$

$H_a : \mu_{\text{oreillle/pied/queue Victoria}} \neq \mu_{\text{oreille/queue/pied NSW\_Queensland}}$

$\alpha = 0.05$

Préparation des données

L'exploration initiale des données avait fait ressortir la présence d'une donnée manquante pour la variable `footlgth`, et la présence probable de données aberrantes dans la variable `taill`.

Pour éviter de supprimer l'ensemble des mesures de l'opossum pour lequel la longueur de pied n'a pas été consignée, la valeur manquante est remplacée par la moyenne de la région à laquelle cet opossum appartient.

```
# Détection de la valeur manquante
which(is.na(possum$footlgth))
```

```
## [1] 41
```

```
# Identification de la région liée à l'observation
possum[41, 2]
```

```
## [1] Victoria
## Levels: NewSouthWales_Queensland Victoria
```

```
# Remplacement de la valeur manquante par la moyenne de la région
possum[41, 9] <- region_means[2, 6]
possum[41, 9]
```

```
## [1] 72.39778
```

La méthode de l'écart-type est maintenant utilisée pour repérer les valeurs aberrantes dans la variable `taill`.

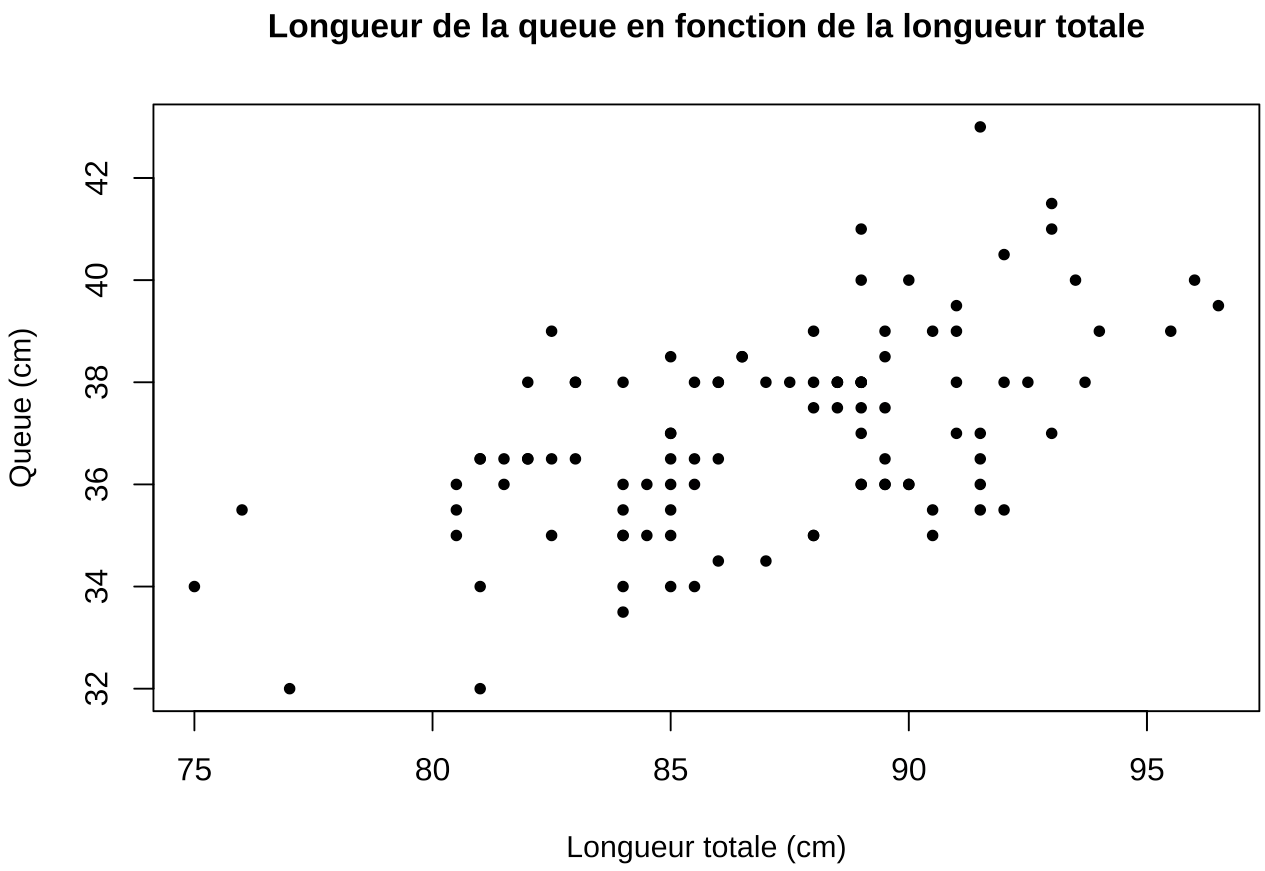
```
# Détection des valeurs aberrantes avec la méthode de l'écart-type
lower_tail <- mean(possum$taill) - (1.5 * sd(possum$taill))
upper_tail <- mean(possum$taill) + (1.5 * sd(possum$taill))
print(paste("Limite inférieure :", round(lower_tail, digits = 2)))
print(paste("Limite supérieure : ", round(upper_tail, digits = 2)))

index_tail_outliers <- which(possum$taill < lower_tail | possum$taill > upper_tail)
print(paste("Nombre de valeurs hors de l'écart-type :", length(index_tail_outliers)))
```

```
## [1] "Limite inférieure : 34.07"
## [1] "Limite supérieure :  39.95"
## [1] "Nombre de valeurs hors de l'écart-type : 17"
```

Encore une fois, la méthode de l'écart type identifie beaucoup plus de données aberrantes que la fonction `diagnose` ne l'avait fait, soit 17 valeurs plutôt que 4. Une visualisation des données pourrait aider à diagnostiquer les valeurs problématiques.

```
par(mfrow = c(1, 1))
plot(taill ~ totlngth, data = possum,
     main = "Longueur de la queue en fonction de la longueur totale",
     pch = 20, xlab = "Longueur totale (cm)", ylab = "Queue (cm)", cex.lab = 0.95, cex.m
ain = 1.05)
```



La visualisation fait ressortir une forte corrélation entre la longueur de la queue et la longueur totale, et les mesures semblent toutes plausibles. Elles seront donc toutes conservées telles quelles pour l'analyse statistique.

Vérification des suppositions

La supposition d'indépendance des données a déjà été vérifiée. La normalité des résidus et l'homoscédasticité sont vérifiées à l'aide de méthodes graphiques, pour chaque caractéristique analysée.

3.2.3.1 Analyse statistique de la longueur de la conque de l'oreille

```
# Séparation des données par région en ne gardant que la variable d'intérêt
victoria_ear <- possum[possum$region == "Victoria", "earconch"]
nswq_ear <- possum[possum$region == "NewSouthWales_Queensland", "earconch"]
print(paste("Nombre d'opossums à Victoria :", length(victoria_ear)))
print(paste("Nombre d'opossums à New South Wales et Queensland :", length(nswq_ear)))
```

```
## [1] "Nombre d'opossums à Victoria : 46"
## [1] "Nombre d'opossums à New South Wales et Queensland : 58"
```

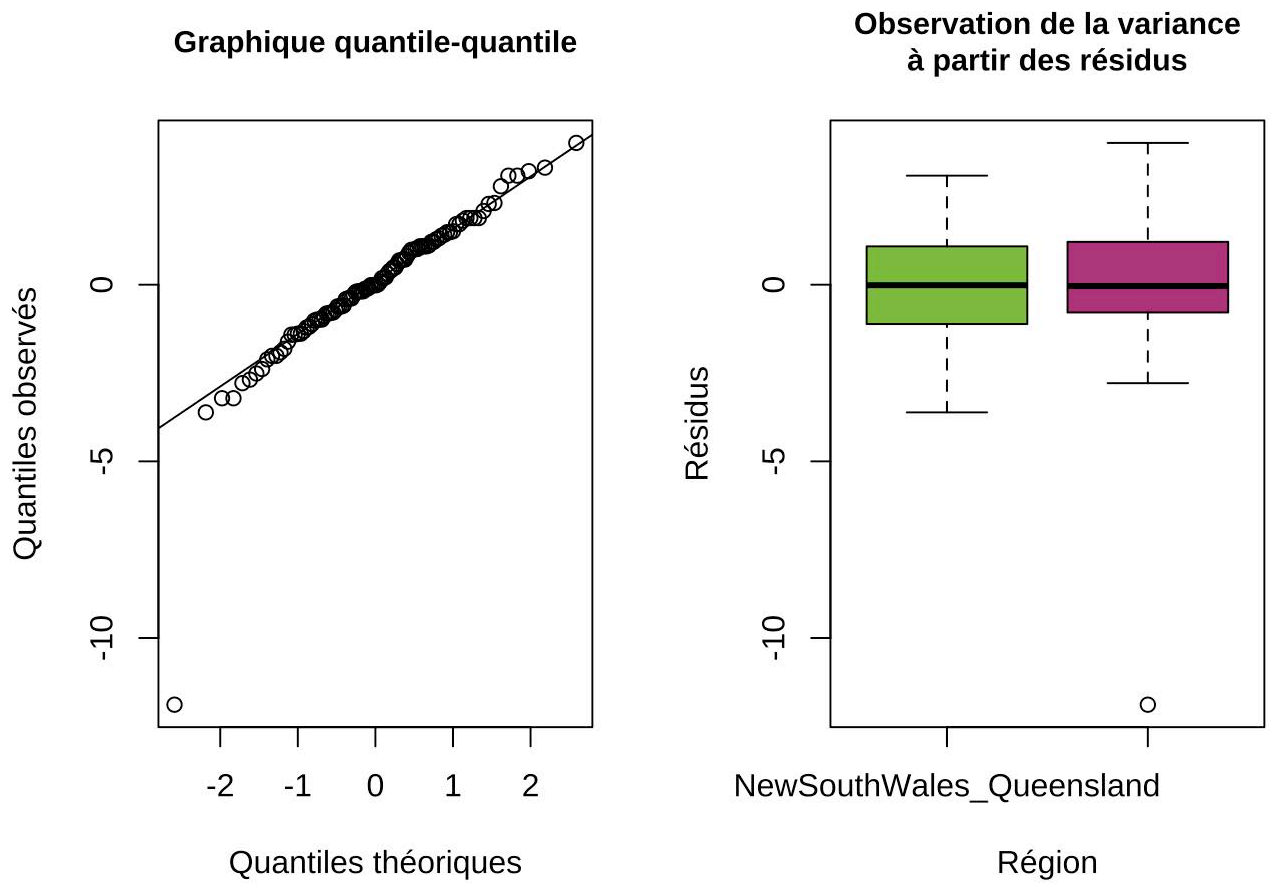
Cette fois encore, les deux groupes sont de taille inégale, mais la variation est moindre qu'entre les groupes liés au sexe.

```
# Calcul des résidus pour la conque de l'oreille
res_victoria_ear <- victoria_ear - mean(victoria_ear)
res_nswq_ear <- nswq_ear - mean(nswq_ear)
res_ear <- c(res_victoria_ear, res_nswq_ear)
```

```
# Vérification des suppositions à l'aide de méthodes graphiques
par(mfrow = c(1, 2))

## Graphique quantile-quantile
qqnorm(res_ear, xlab = "Quantiles théoriques", ylab = "Quantiles observés",
       main = "Graphique quantile-quantile", cex.main = 0.95)
qqline(res_ear)

# Homoscédasticité
possum$res_ear <- res_ear
boxplot(res_ear ~ region, data = possum,
        col = region_col,
        main = "Observation de la variance\nà partir des résidus",
        ylab = "Résidus", xlab = "Région", cex.main = 0.95)
```



La vérification des suppositions indique que les résidus suivent une distribution normale et que l'homoscédasticité est respectée. L'analyse peut donc se faire avec un test *t* de Student.

```
t.test(earconch ~ region, data = possum, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  earconch by region
## t = -19.028, df = 102, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group NewSouthWales_Queensla
nd and group Victoria is not equal to 0
## 95 percent confidence interval:
##  -8.031318 -6.515009
## sample estimates:
## mean in group NewSouthWales_Queensland                mean in group Victoria
##                               44.91379                                 52.18696
```

Interprétation des résultats

Le test *t* montre que la probabilité d'observer une valeur de t = -19,03 dans une population où H_0 est vraie est très faible ($P < 0{,}0001$). L'hypothèse nulle est donc rejetée, ce qui permet de conclure qu'il existe une différence significative de longueur de la conque de l'oreille entre les opossums des deux régions.

3.2.3.2 Analyse statistique de la longueur du pied

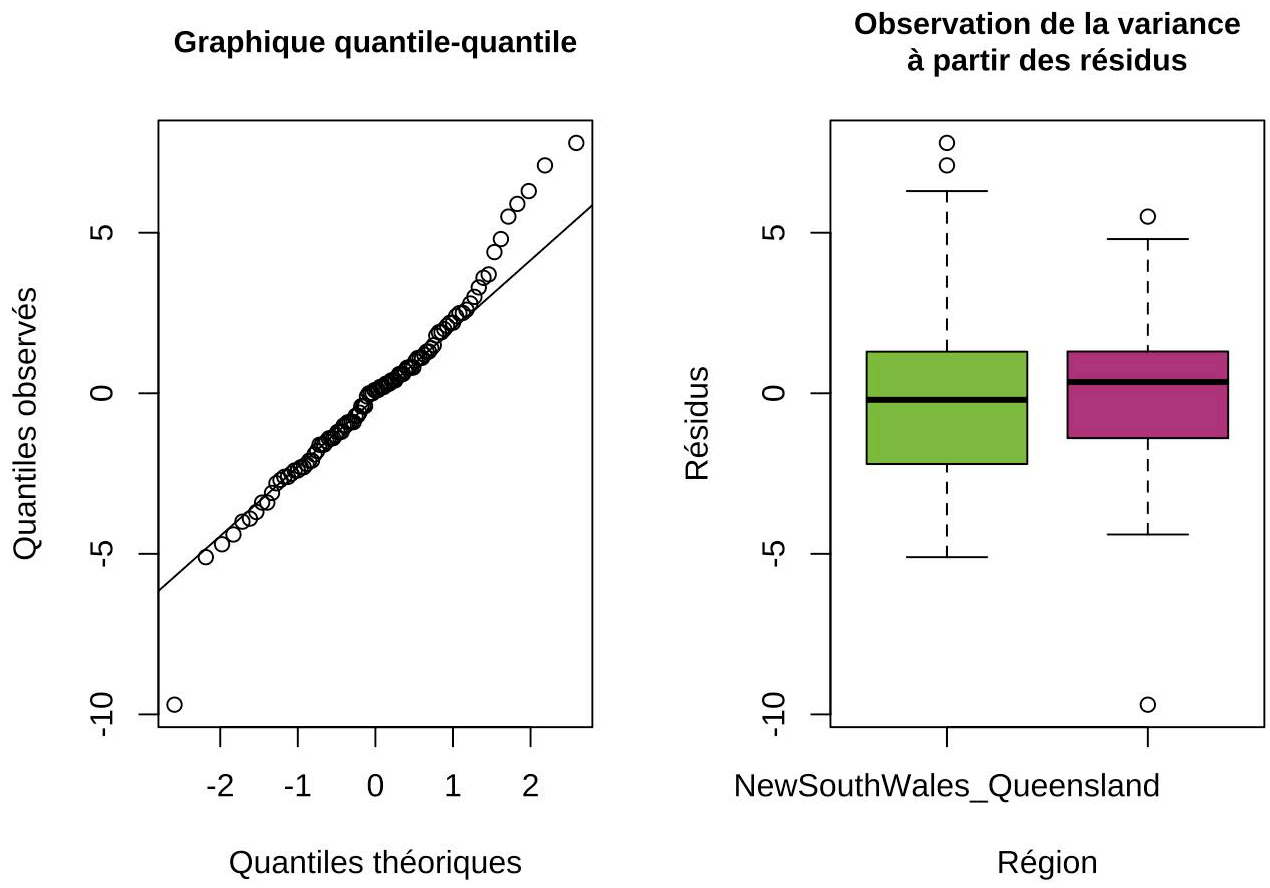
```
# Séparation des données par région en ne gardant que la variable d'intérêt
victoria_foot <- possum[possum$region == "Victoria", "footlgth"]
nswq_foot <- possum[possum$region == "NewSouthWales_Queensland", "footlgth"]
```

```
# Calcul des résidus pour la longueur du pied
res_victoria_foot <- victoria_foot - mean(victoria_foot)
res_nswq_foot <- nswq_foot - mean(nswq_foot)
res_foot <- c(res_victoria_foot, res_nswq_foot)
```

```
# Vérification des suppositions à l'aide de méthodes graphiques
par(mfrow = c(1, 2))

## Graphique quantile-quantile
qqnorm(res_foot, xlab = "Quantiles théoriques", ylab = "Quantiles observés",
       main = "Graphique quantile-quantile", cex.main = 0.95)
qqline(res_foot)

# Homoscédasticité
possum$res_foot <- res_foot
boxplot(res_foot ~ region, data = possum,
        col = region_col,
        main = "Observation de la variance\nà partir des résidus",
        ylab = "Résidus", xlab = "Région", cex.main = 0.95)
```



Le graphique quantile-quantile indique que les résidus suivent assez bien une distribution normale. La supposition d'homoscédasticité est aussi respectée, l'analyse peut donc se faire ici aussi à l'aide d'un test *t* de Student.

```
t.test(footlgth ~ region, data = possum, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  footlgth by region
## t = -13.234, df = 102, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group NewSouthWales_Queensla
nd and group Victoria is not equal to 0
## 95 percent confidence interval:
##  -8.042603 -5.946056
## sample estimates:
## mean in group NewSouthWales_Queensland                mean in group Victoria
##                                65.40345                                72.39778
```

Interprétation des résultats

Le résultat du test *t* pour la longueur du pied est similaire au test précédent, avec une probabilité très faible ($P < 0{,}0001$) d'observer une valeur de $t = -13{,}23$ dans une population où H_0 est vraie. L'hypothèse nulle est donc rejetée encore une fois, ce qui permet de conclure qu'il existe aussi une différence significative de longueur du pied entre les opossums des deux régions.

3.2.3.3 Analyse statistique de la longueur de la queue

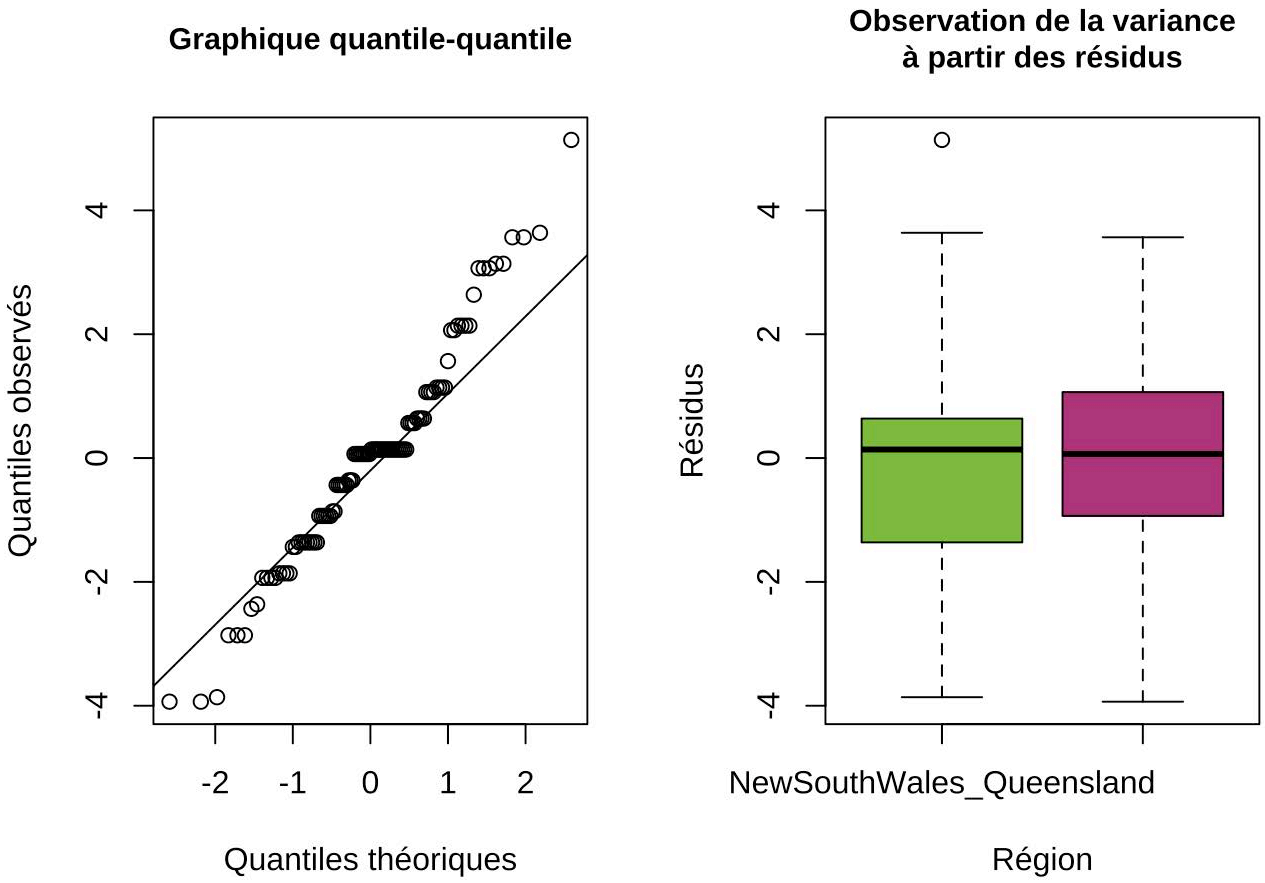
```
# Séparation des données par région en ne gardant que la variable d'intérêt
victoria_tail <- possum[possum$region == "Victoria", "taill"]
nswq_tail <- possum[possum$region == "NewSouthWales_Queensland", "taill"]
```

```
# Calcul des résidus pour la longueur de la queue
res_victoria_tail <- victoria_tail - mean(victoria_tail)
res_nswq_tail <- nswq_tail - mean(nswq_tail)
res_tail <- c(res_victoria_tail, res_nswq_tail)
```

```
# Vérification des suppositions
par(mfrow = c(1, 2))

## Graphique quantile-quantile
qqnorm(res_tail, xlab = "Quantiles théoriques", ylab = "Quantiles observés",
       main = "Graphique quantile-quantile", cex.main = 0.95)
qqline(res_tail)

# Homoscédasticité
possum$res_tail <- res_tail
boxplot(res_tail ~ region, data = possum,
        col = region_col,
        main = "Observation de la variance\nà partir des résidus",
        ylab = "Résidus", xlab = "Région", cex.main = 0.95)
```



La vérification des suppositions à l'aide des graphiques confirme le respect de l'homoscédasticité, mais la distribution des résidus ne suit pas tout à fait la droite théorique. Des vérifications supplémentaires sont donc faites à l'aide de tests formels.

```
# Test formel d'Anderson-Darling
ad.test(res_tail)
```

```
##
##  Anderson-Darling normality test
##
## data:  res_tail
## A = 1.4222, p-value = 0.001062
```

```
# Test formel de Cramer-von Mises
cvm.test(res_tail)
```

```
##
##  Cramer-von Mises normality test
##
## data:  res_tail
## W = 0.28123, p-value = 0.0005311
```

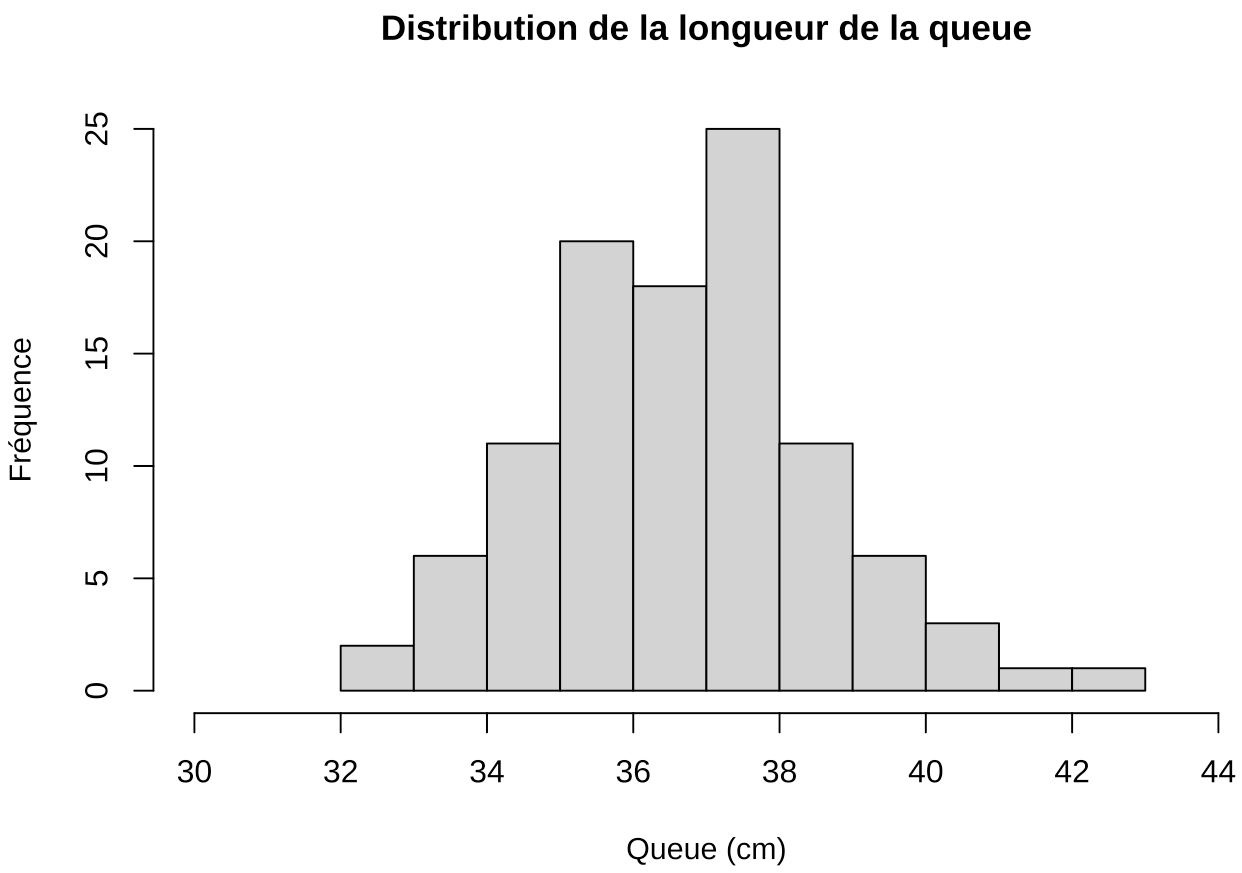
```
# Test formel de Shapiro-Wilk
shapiro.test(res_tail)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_tail
## W = 0.96931, p-value = 0.0162
```

Aucun test formel ne confirme le respect de la normalité des résidus, les valeurs de *P* étant toutes inférieures au seuil de signification, fixé à 0,05.

La distribution de la variable est représentée par un histogramme.

```
hist(possum$taill, main = "Distribution de la longueur de la queue",
        cex.main = 1.1,
        xlab = "Queue (cm)", ylab = "Fréquence", xlim = c(30, 44), ylim = c(0, 25),
        cex.lab = 0.95)
```



La distribution n’est pas symétrique, alors une transformation des données pourrait s’avérer utile. Cependant, après essais, ni la transformation logarithmique des données, ni la transformation au carré n’a permis de respecter la supposition de normalité des résidus. Par conséquent, l’analyse se fera à l’aide d’un test non paramétrique, celui de Mann-Whitney.

```
# Test de Mann-Whitney
wilcox.test(taill ~ region, data = possum)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  taill by region
## W = 2112.5, p-value = 2.954e-07
## alternative hypothesis: true location shift is not equal to 0
```

Interprétation des résultats

Le test effectué donne une valeur de probabilité bien inférieure au seuil de signification ($P < 0{,}0001$), l'hypothèse nulle est donc rejetée. Cela permet de conclure que, malgré une différence moins marquée que pour les autres caractéristiques (longueur de la conque de l'oreille et longueur du pied), la différence de longueur de queue chez les opossums des deux régions est bien significative.

3.3 Conclusions

La section d'analyse des données cherchait à vérifier s'il existe un dimorphisme chez les opossums étudiés, dimorphisme sexuel d'une part, et dimorphisme géographique d'autre part. Aucune différence significative entre les femelles et les mâles n'ayant été observée à partir des données morphométriques, l'hypothèse du dimorphisme sexuel est rejetée. À l'inverse, les mesures de trois caractéristiques morphologiques diffèrent selon la région, ce qui confirme l'existence d'un dimorphisme géographique entre les deux populations d'opossums étudiés. Ces caractéristiques sont la longueur de la conque de l'oreille et celle du pied, deux caractéristiques plus petites chez les opossums du nord, et la longueur de la queue, plus courte chez les opossums du sud.

Il est à noter que ces conclusions concordent avec les résultats publiés par David Lindenmayer en 1995, dans son article *Morphological Variation Among Populations of the Mountain Brushtail Possum, Trichosurus-Caninus Ogilby (Phalangeridae, Marsupialia)* (https://www.publish.csiro.au/zo/ZO9950449).

---

4. Prédiction de l'espèce

L'objectif de cette section est de développer un modèle de classification permettant de prédire l'espèce à laquelle appartient un opossum, à partir des mesures de ses caractéristiques morphologiques les plus significatives.

L'espèce est d'abord ajoutée aux données, en associant le *Trichosurus cunninghami* à la région de Victoria, et le *Trichosurus caninus* à la région de New South Wales et Queensland.

```
# Ajout de l'espèce aux données
possum$species <- as.character(possum$region)
possum$species[possum$region == "Victoria"] <- "cunninghami"
possum$species[possum$region == "NewSouthWales_Queensland"] <- "caninus"
possum$species <- as.factor(possum$species)
table(possum$species)
```

```
##
##     caninus cunninghami
##          58          46
```

4.1 Arbre de décision

Avant de construire un modèle de classification, le gain d'information et le rapport de gain sont calculés pour chacune des neuf caractéristiques morphologiques.

```
# Chargement des bibliothèques nécessaires
library(rpart)
library(FSelector)
```

```
# Calcul du gain d'information
information.gain(species ~ hdlngth + skullw + totlngth + taill + footlgth + earconch + e
ye + chest + belly, data = possum, unit = "log2")
```

```
##          attr_importance
## hdlngth       0.09621865
## skullw        0.00000000
## totlngth      0.00000000
## taill         0.19491552
## footlgth      0.61011255
## earconch      0.92005712
## eye           0.00000000
## chest         0.00000000
## belly         0.00000000
```

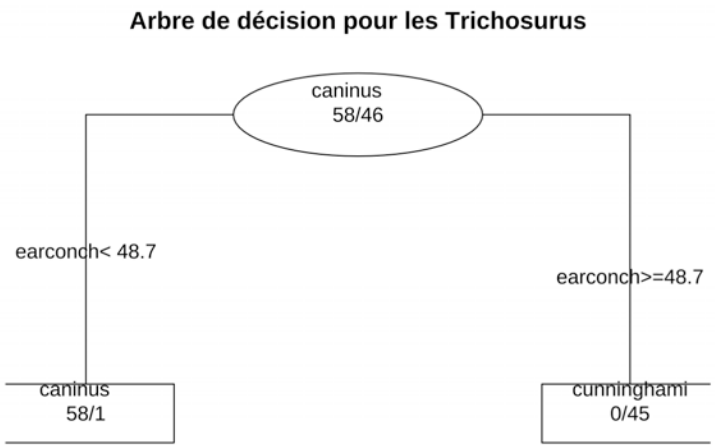
```
# Calcul du rapport de gain
gain.ratio(species ~ hdlngth + skullw + totlngth + taill + footlgth + earconch + eye + c
hest + belly, data = possum, unit = "log2")
```

```
##          attr_importance
## hdlngth        0.1975656
## skullw         0.0000000
## totlngth       0.0000000
## taill          0.2042211
## footlgth       0.6102754
## earconch       0.9322807
## eye            0.0000000
## chest          0.0000000
## belly          0.0000000
```

Ces calculs confirment que c'est la longueur de la conque de l'oreille qui a le plus grand pouvoir discrimant entre les deux espèces d'opossums, suivie de la longueur du pied. La longueur de la queue a moins d'importance, et la longueur de la tête en a peu, mais presque autant que la longueur de la queue. Comme la longueur de la tête n'a pas fait l'ojbet d'une analyse statistique, elle ne sera pas incluse dans le modèle de classification.

En tenant compte des trois caractéristiques les plus importantes, un arbre de décision est dessiné.

```
tree_possum <- rpart(species ~ earconch + footlgth + taill, method = "class", data = pos
sum)
plot(tree_possum, uniform = TRUE, margin = 0.1, main = "Arbre de décision pour les Trich
osurus")
text(tree_possum, use.n = TRUE, fancy = TRUE, pretty = 0, all = TRUE)
```



Cet arbre confirme l'importance de la mesure de la conque de l'oreille, qui à elle seule arrive à classer 103 des 104 opossums dans la bonne espèce! L'observation mal classifiée correspond certainement à la faible valeur observée dans le diagramme en boîte de la section 3.2.2.

4.2 Classificateur bayésien

Toujours à partir des trois caractéristiques morphologiques identifiées par l'analyse et le calcul du gain d'information, un classificateur bayésien est développé. Les données sont d'abord divisées en une base d'entraînement comprenant 80 % des observations, et une base de test comprenant le pourcentage restant des observations.

```
# Chargement des bibliothèques nécessaires
library(caTools)
library(e1071)
library(caret)
```

```
# Séparation des données en base d'entraînement et base de test
set.seed(14)
sample_possum <- sample.split(possum$species, SplitRatio = 0.8)
possum_train <- subset(possum, sample_possum == TRUE)
possum_test <- subset(possum, sample_possum == FALSE)
table(possum_train$species)
```

```
##
##     caninus cunninghami
##          46          37
```

```
table(possum_test$species)
```

```
##
##     caninus cunninghami
##          12           9
```

Il y a 83 observations dans la base d'entraînement, et 21 dans la base de test. Les observations sont divisées proportionnellement selon l'espèce.

Le modèle est maintenant construit sur les données d'entraînement.

```
# Classificateur bayésien
nB_model_possum <- naiveBayes(species ~ footlgth + earconch + taill, data = possum_trai
n)
nB_model_possum
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     caninus cunninghami
##   0.5542169   0.4457831
##
## Conditional probabilities:
##              footlgth
## Y                 [,1]     [,2]
##   caninus     65.40652 2.645365
##   cunninghami 72.75940 1.934703
##
##              earconch
## Y                 [,1]     [,2]
##   caninus     44.95217 1.592167
##   cunninghami 52.12162 2.496568
##
##              taill
## Y                 [,1]     [,2]
##   caninus     38.04348 1.778970
##   cunninghami 35.95946 1.578406
```

Les deux classes ne sont pas équiprobables, et les probabilités conditionnelles ont été calculées pour chaque caractéristique. Le classificateur peut maintenant être utilisé pour faire des prédictions sur les données de test.

```
# Prédictions sur la base de test
nB_predict_possum <- predict(nB_model_possum, possum_test)
nB_predict_possum
```

```
##  [1] cunninghami cunninghami cunninghami cunninghami cunninghami cunninghami
##  [7] cunninghami cunninghami caninus     caninus     caninus     caninus
## [13] caninus     caninus     caninus     caninus     caninus     caninus
## [19] caninus     caninus     caninus
## Levels: caninus cunninghami
```

Pour évaluer le modèle, la matrice de confusion est affichée, et le taux de bonne classification est calculé.

```
# Matrice de confusion
nB_confusion <- table(true = possum_test$species, pred = nB_predict_possum)
nB_confusion
```

```
##              pred
## true          caninus cunninghami
##   caninus          12           0
##   cunninghami       1           8
```

```
# Taux de bonne classification
nB_accuracy <- sum(diag(nB_confusion)) / sum(nB_confusion) * 100
print(paste("Taux de bonne classification :", round(nB_accuracy, digits = 2), "%"))
```

```
## [1] "Taux de bonne classification : 95.24 %"
```

```
confusionMatrix(nB_confusion)
```

```
## Confusion Matrix and Statistics
##
##              pred
## true          caninus cunninghami
##   caninus          12           0
##   cunninghami       1           8
##
##                Accuracy : 0.9524
##                  95% CI : (0.7618, 0.9988)
##     No Information Rate : 0.619
##     P-Value [Acc > NIR] : 0.0005888
##
##                   Kappa : 0.9014
##
##  Mcnemar's Test P-Value : 1.0000000
##
##             Sensitivity : 0.9231
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.8889
##              Prevalence : 0.6190
##          Detection Rate : 0.5714
##    Detection Prevalence : 0.5714
##       Balanced Accuracy : 0.9615
##
##        'Positive' Class : caninus
##
```

Les 12 *Trichosurus caninus* ont été classés comme *caninus*, et 8 des 9 *Trichosurus cunninghami* ont été classés comme *cunninghami*, ce qui correspond à un taux de bonne classification d'un peu plus de 95 %. Les statistiques obtenues à l'aide de la fonction `confusionMatrix` confirment que le modèle est bon. En fait, avec si peu d'observations, il n'est pas possible de faire mieux, puisqu'une seule des observations de la base de test a été mal classée.

4.3 Estimateur des *K* plus proches voisins

Pour comparer, un nouveau modèle de classification est développé à l'aide de l'estimateur des *K* plus proches voisins. Les mêmes bases d'entraînement et de test que pour le classificateur bayésien sont utilisées, mais les données des trois caractéristiques morphologiques d'intérêt sont normalisées.

```
# Chargement de la bibliothèque nécessaire
library(class)
```

```
# Normalisation des données
train_scale <- scale(possum_train[, 8:10])
test_scale <- scale(possum_test[, 8:10])
head(train_scale)
```

```
##           taill  footlgth  earconch
## C3  -0.56312920 1.3340766 1.5413761
## C5  -0.31048205 0.8752912 0.7405739
## C15  0.44745942 1.7011049 0.9832413
## C24 -0.81577636 1.0358661 1.3229755
## C27 -0.05783489 0.9211698 1.3957757
## C28 -0.05783489 0.8523520 1.1531084
```

```
head(test_scale)
```

```
##          taill  footlgth  earconch
## C10  1.2871517 1.6815745 0.9217705
## C23 -0.3186019 0.7134905 1.2339831
## C26 -0.3186019 0.8235000 0.9457869
## C32  1.2871517 2.0776089 0.7776724
## C39 -1.3891043 0.4274657 0.8977542
## AD1  0.2166493 0.6694867 0.8977542
```

Le modèle est d'abord développé simplement, c'est-à-dire avec une valeur de *K* égale à 1.

```
# Développement du modèle
knn_model_possum <- knn(train = train_scale, test = test_scale, cl = possum_train$specie
s, k = 1)
# Matrice de confusion
knn_confusion <- table(possum_test$species, knn_model_possum)
knn_confusion
```

```
##              knn_model_possum
##               caninus cunninghami
##   caninus          12           0
##   cunninghami       0           9
```

```
# Taux de classification
knn_accuracy <- sum(possum_test$species == knn_model_possum) / length(possum_test$specie
s) * 100
print(paste("Taux de bonne classification :", knn_accuracy, "%"))
```

```
## [1] "Taux de bonne classification : 100 %"
```

Avec ce modèle, les 21 opossums sont classés correctement, ce qui n'est pas souhaitable. La valeur de *K* est donc variée, et le modèle est réévalué à chaque variation.

```
# Évaluation du modèle et choix de la valeur de K
# K = 3
knn3_model_possum <- knn(train = train_scale, test = test_scale, cl = possum_train$speci
es,
                      k = 3)
knn_accuracy <- sum(possum_test$species == knn3_model_possum) / length(possum_test$speci
es) * 100
print(paste("Taux de bonne classification avec k = 3 :", round(knn_accuracy, digits =
2), "%"))

# K = 5
knn5_model_possum <- knn(train = train_scale, test = test_scale, cl = possum_train$speci
es,
                      k = 5)
knn_accuracy <- sum(possum_test$species == knn5_model_possum) / length(possum_test$speci
es) * 100
print(paste("Taux de bonne classification avec k = 5 :", round(knn_accuracy, digits =
2), "%"))

# K = 7
knn7_model_possum <- knn(train = train_scale, test = test_scale, cl = possum_train$speci
es,
                      k = 7)
knn_accuracy <- sum(possum_test$species == knn7_model_possum) / length(possum_test$speci
es) * 100
print(paste("Taux de bonne classification avec k = 7 :", round(knn_accuracy, digits =
2), "%"))

# K = 15
knn15_model_possum <- knn(train = train_scale, test = test_scale, cl = possum_train$spec
ies,
                      k = 15)
knn_accuracy <- sum(possum_test$species == knn15_model_possum) / length(possum_test$spec
ies) * 100
print(paste("Taux de bonne classification avec k = 15 :", round(knn_accuracy, digits =
2), "%"))
```

```
## [1] "Taux de bonne classification avec k = 3 : 95.24 %"
## [1] "Taux de bonne classification avec k = 5 : 100 %"
## [1] "Taux de bonne classification avec k = 7 : 100 %"
## [1] "Taux de bonne classification avec k = 15 : 100 %"
```

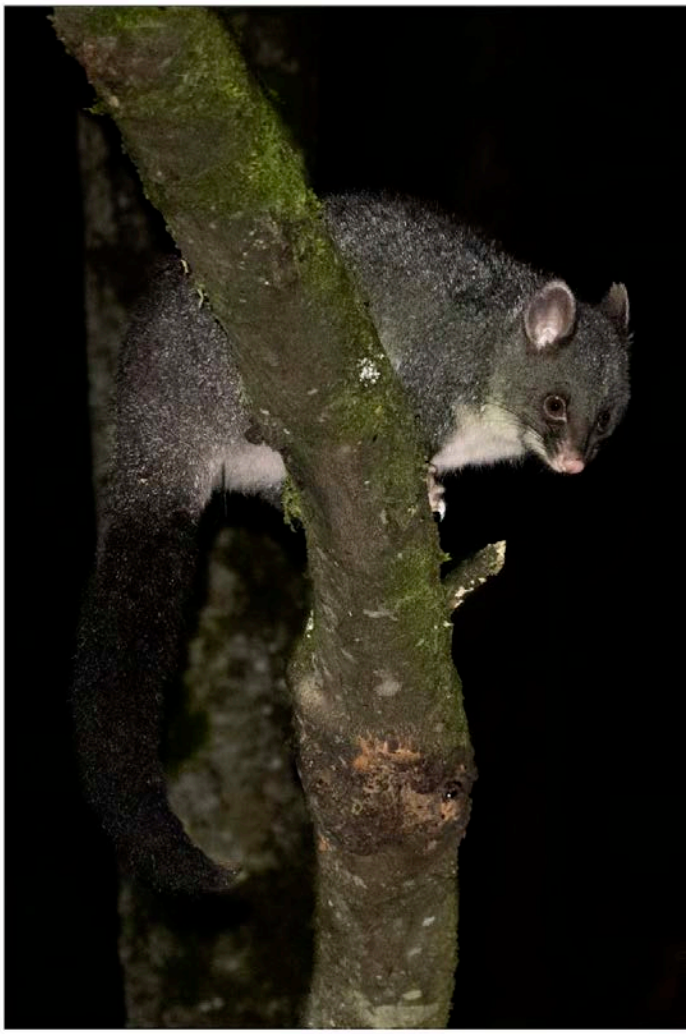
Avec une valeur de *K* égale à 3, le taux de classification obtenu est le même qu'avec le classificateur bayésien, soit 95,24 %. Cependant, avec toute autre valeur de *K*, la classification est parfaite, avec un taux de 100 %. Il faudrait certainement un bien plus grand nombre d'échantillons d'apprentissage que les 83 de ce projet pour que cet estimateur soit réellement performant.

Conclusion

À partir des données morphométriques de 104 opossums de l'est de l'Australie, des *mountain brushtail possums* capturés sur sept sites différents par le scientifique David Lindenmayer et son équipe, ce projet visait à vérifier l'existence d'un dimorphisme entre les opossums étudiés, d'une part en fonction du sexe, et d'autre part en fonction de l'habitat. L'analyse des caractéristiques morphologiques en fonction du sexe n'a fait ressortir aucune différence significative de taille entre les femelles et les mâles, et l'hypothèse d'un dimorphisme sexuel a été rejetée. En revanche, l'analyse des caractéristiques morphologiques en fonction de l'habitat, c'est-à-dire la région de Victoria, regroupant deux sites, et celle de New South Wales et Queensland, regroupant les cinq autres sites, a confirmé l'existence d'un dimorphisme géographique. En effet, des différences significatives entre les populations ont été observées pour trois caractéristiques, soit une oreille plus grande et un pied plus long chez les opossums de la région australe, mais une queue plus courte chez les opossums de cette même région.

À partir de ces caractéristiques, différentes techniques d'apprentissage automatique ont été appliquées pour prédire l'espèce à laquelle appartient un opossum en fonction de ses mesures morphologiques, l'espèce étant liée à la région. De fait, à la suite d'une seconde étude, Lindenmayer et ses collègues ont classé les opossums en deux espèces distinces de *Trichosurus*, soit *T. caninus* pour la région plus nordique de New South Wales et Queensland, et *T. cunninghami* pour la région plus australe de Victoria. Un taux de bonne classification d'environ 95 % a été obtenu avec un classificateur bayésien, ainsi qu'avec un estimateur des K plus proches voisins, avec une valeur de K égale à 3. Compte tenu du nombre limité d'observations, ces résultats sont jugés satisfaisants.

Ce projet s'est basé en grande partie sur les travaux de Lindenmayer, mais il serait intéressant d'étudier la réelle influence de l'habitat sur les caractéristiques morphologiques des opossums. Les *Trichosurus cunninghami* ont-ils developpé des oreilles et des pieds plus grands que leurs cousins *caninus* en raison de l'environnement de Victoria (végétation, altitude, climat, etc.), ou la présence de cette espèce à cet endroit n'est-elle que fortuite, et ses caractéristiques particulières, tributaires d'une autre variable?



À gauche : *Trichosurus caninus* photographié par Brett Vercoe (https://www.inaturalist.org/observations/45661421).

À droite : *Trichosurus cunninghami* photographié par jkmalkoha (https://www.inaturalist.org/observations/233901560).