Module 4
Visualisation des données

Sommaire

Dernière mise à jour le 4 août 2020

Introduction

La visualisation des données est un ensemble de méthodes de représentation graphique qui permettent de synthétiser les informations que contiennent ces données pour mettre en évidence les informations clés qu'ils renferment.

> *La visualisation est un aspect essentiel de l'analyse de données. Elle offre une ligne d'attaque frontale, révèle la structure complexe de données qui ne pourraient être comprises d'aucune autre façon. Elle permet de découvrir des résultats inattendus et de remettre en question les conclusions attendues. (William S. Cleveland (in Visualizing Data, Hobart Press, 1993))*

La visualisation de données ne cesse de prendre de la place dans des domaines très variés allant du domaine médical au domaine financier. En effet, nous assistons actuellement à une accumulation de grande quantité de données qui dépasse la capacité humaine de compréhension des informations qu'elles recèlent. Dans ce contexte, nous n'avons plus besoin de données, mais d'informations pertinentes, utiles et compréhensibles. La visualisation des données est un moyen efficace pour répondre à ce besoin.

4.1 Graphiques classiques

4.1.1 Diagramme en bâtons

Un **diagramme en bâtons** est une représentation graphique de la distribution des données à l'aide de segments (Figure 4.1). Dans le cas de variables quantitatives discrètes, les valeurs des individus sont représentées sur l'axe horizontal (l'abscisse) et les valeurs des variables correspondantes sur l'axe vertical (l'ordonnée). À chaque individu correspond un segment dont la hauteur est proportionnelle à la valeur de sa variable représentée. Notons que sur l'axe des ordonnées, on peut également représenter les fréquences ou les

pourcentages. Ceci aura pour effet de conserver l'allure du graphique, mais de modifier ses valeurs. Dans le cas de variables qualitatives, les différentes modalités de la variable qualitative sont représentées sur l'axe horizontal et l'effectif correspondant sur l'axe vertical.

▶ **Exemple 4.1** Commençons par cet exemple tiré d'une étude évaluant la qualité de service et la quantité d'information reçue par le patient lors de son séjour à l'hôpital. La base de données `satisfaction_hopital.csv` comprend les informations de 534 patients ayant séjourné dans des hôpitaux de la région parisienne.

Les lignes de codes suivantes permettent d'interroger la base de données et d'obtenir le diagramme en baton de la figure 4.1.

```
# Lecture des données
MyData <- read.csv(file="satisfaction_hopital.csv", header=TRUE, sep=",")
# Affichage des données
summary(MyData)
MyData$service.c<-factor(MyData$service)
tab.service<-table(MyData$service.c)
barplot(tab.service,main="Répartition des services", xlab="Numéro des services",ylab="Effectifs",ylim=c(0,100),col="blue", cex.axis = 1.5, cex.lab=1.5)
```

```
    service          sexe            age          profession
 Min.   :1.000   Min.   :0.0000   Min.   :18.00   Min.   :1.000
 1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:45.00   1st Qu.:3.000
 Median :5.000   Median :0.0000   Median :60.00   Median :4.000
 Mean   :4.549   Mean   :0.4981   Mean   :58.21   Mean   :4.431
 3rd Qu.:7.000   3rd Qu.:1.0000   3rd Qu.:72.00   3rd Qu.:5.500
 Max.   :8.000   Max.   :1.0000   Max.   :97.00   Max.   :8.000
                                  NA's   :6       NA's   :107
```

La commande `summary(MyData)` permet d'afficher toutes les variables considérées parmi

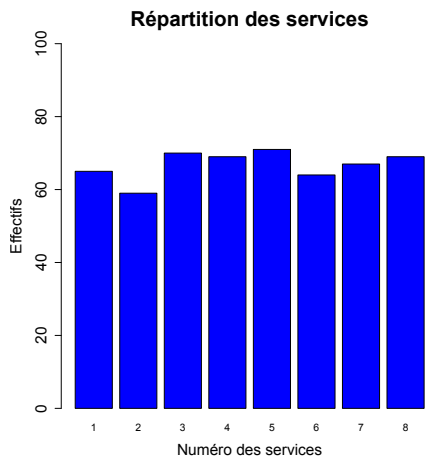

Figure 4.1: Diagramme en bâtons.

lesquelles on retrouve la variable binaire `sexe` et la variable quantitative `age` qui varie entre 18 et 97 ans. On retrouve également la variable `service` qui désigne le service ayant accueilli le patient (qui peut varier entre 1 et 8). Le diagramme en bâton de la figure 4.1 illustre le nombre de patient (sur l'axe des ordonnées) par service (sur l'axe des abscisses).

4.1.2 Diagramme en boîte

Le **diagramme en boîte** ou boîte à moustache d'un ensemble de données est une représentation graphique de ses caractéristiques statistiques. Il résume les caractéristiques de position de la variable étudiée à savoir la médiane, les quartiles et l'étendue. Il est à noter que ce diagramme peut être horizontal (Figure 4.2) ou bien vertical (Figure 4.4). Cette représentation est composée de deux rectangles et de deux segments. Les limites du rectangle correspondent au premier et troisième quartile, désignés respectivement par Q1 et Q3 sur la Figure 4.2. Les limites des segments correspondent à la valeur minimale et maximale de l'ensemble des données. L'étendue correspond à l'écart entre la valeur minimale et la valeur maximale.

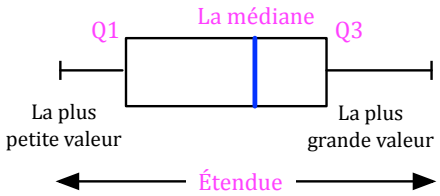

Figure 4.2: Diagramme en boîte.

▶ **Exemple 4.2** Revenons à l'ensemble des données étudié précédemment. Les lignes de codes suivantes permettent de représenter le diagramme en boîte qui décrit la répartition de l'âge (Figure 4.3) et de l'âge selon le sexe (Figure 4.4).

```
boxplot(MyData$age,horizontal=TRUE,col="blue",main="Répartition des âges",ylim=c(0,100))
boxplot(MyData$age~MyData$sexe,names=c("Homme","Femme"),col=c("blue","pink"),main="Répartition des âges par sexe", cex.main = 2, cex.sub=1.5, cex.axis = 1.5, cex.lab= 1.5)
```

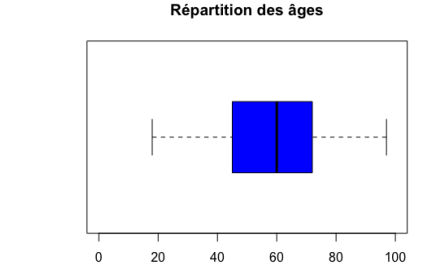

Figure 4.3: Diagramme en boîte de l'age.

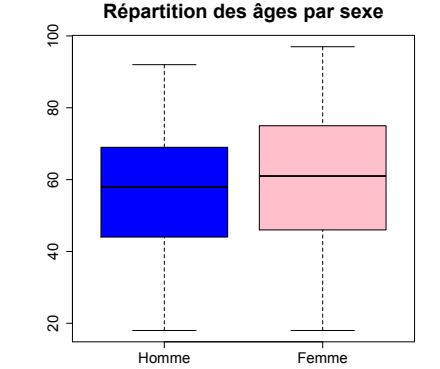

Figure 4.4: Diagramme en boîte de l'age.

Nous pouvons voir graphiquement sur la Figure 4.3 que l'âge minimum est 18 ans, l'âge maximum est 97 ans et la médiane est située autour de 60 ans. Ces valeurs sont confirmées par le résultat (*Outputs*) de la commande `summary` :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  18.00   45.00   60.00   58.21   72.00   97.00       6
```

4.1.3 Diagramme circulaire

Un diagramme circulaire, aussi appelé camembert ou tarte, est une représentation graphique de données qualitatives sous la forme d'un disque partagé en secteurs (Figure 4.5). À chaque modalité étudiée correspond un secteur. Les mesures des secteurs sont proportionnelles aux effectifs représentés.

▶ **Exemple 4.3** Nous nous intéressons dans cet exemple à la variable `profession` qui prend une parmi les 8 valeurs suivantes : (1) Agriculteur exploitant (2) Artisan, commerçant, chef d'entreprise (3) Cadre, profession intellectuelle ou artistique, profession libérale (4) Profession intermédiaire de l'enseignement, de la santé, du travail social ou de la fonction publique, technicien (5) Employé (6) Ouvrier (7) Étudiant, militaire, chômeur sans avoir jamais travaillé et (8) autre.

Les lignes de codes suivantes permettent de générer le diagramme circulaire de la Figure 4.5 qui contient les étiquettes de chaque secteur (Ligne de code 2) auxquelles nous pouvons ajouter les pourcentages (Lignes de code 4).

```
1 #Label des professions uniquement
2 lbls <- c("Agriculteur exploitant", "Artisan", "Commerçant", "Chef d'↩
    entreprise", "Cadre", "Profession intermédiaire", "Employé", "↩
    Ouvrier","Étudiant")
3 #Label des professions avec les pourcentages
4 lbls <- paste(lbls, "\n", table(MyData$profession), labels = paste(↩
    round(prop.table(table(MyData$profession))*100), "%", sep = ""))
5 #Diagramme circulaire
6 pie(table(MyData$profession),main="Répartition des professions",labels↩
    =lbls)
```

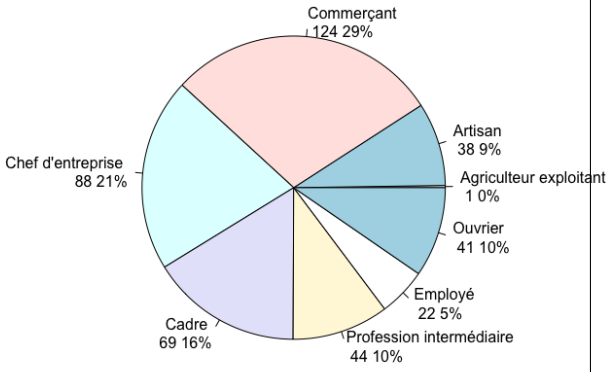

Figure 4.5: Diagramme circulaire des différentes professions.

4.2 Graphiques avancés

4.2.1 Superposition de graphique

La superposition de graphiques de base permet d'afficher simultanément plusieurs graphiques afin d'avoir des informations complémentaires

▶ **Exemple 4.4** Soit l'ensemble de données IRIS correspondant à 3 espèces de fleurs (Iris setosa, Iris virginica, Iris versicolor). Ces espèces sont caractérisées par 4 variables : la longueur et la largeur des sépales et la longueur et la largeur des pétales. Toutes ces variables sont quantitatives (exprimées en millimètres). La base de données comprend 50 échantillons par espèce. La Figure 4.6 illustre la variation de la longueur des pétales en fonction de la longueur des sépales. Cette figure peut être obtenue par les lignes 1 à 8 du code suivant. Nous pourrons également superposées des informations complémentaires par l'ajout de la droite de régression de la longueur des pétales sur la longueur des sépales.

```
1  data(iris)
2  iris[1:150, ]
3  plot(iris$Petal.Length,iris$Sepal.Length,
4       las=1, cex = 0.8, pch = 4, col = "blue",
5       main = "Longueur pétales vs. sépales",
6       col.main = "blue",
7       ylab= "Longueur des sépales (cm)",
8       xlab = "Longueur des pétales (cm)")
9  reg=lm(iris$Sepal.Length~iris$Petal.Length)
10 abline(reg,lty = 2, col = "red")
11 abline(h = mean(iris$Sepal.Length),
12        lty = 3, lwd = 2, col = "darkgreen")
13 legend(x = 1, y = 7.9,
14        legend = c("Données observées", "Droite de régression",
15                   "Moyenne de la longuer des sépales"),
16        cex = 0.8, pch = c(4, NA, NA), lty = c(NA, 1, 3), lwd = c(1, 1,↩
              2), col = c("blue", "red", "darkgreen"))
```

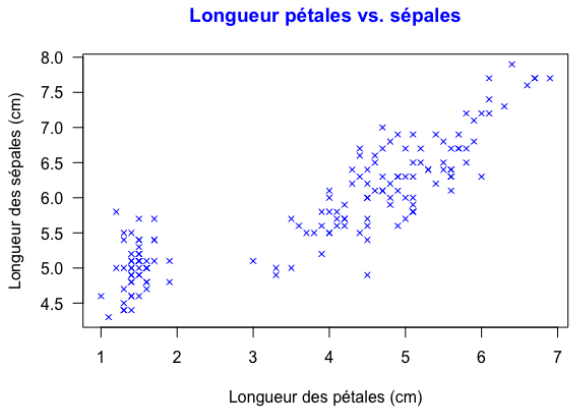

Figure 4.6: Variation de la longueur de pétales en fonction de celles des sépales.

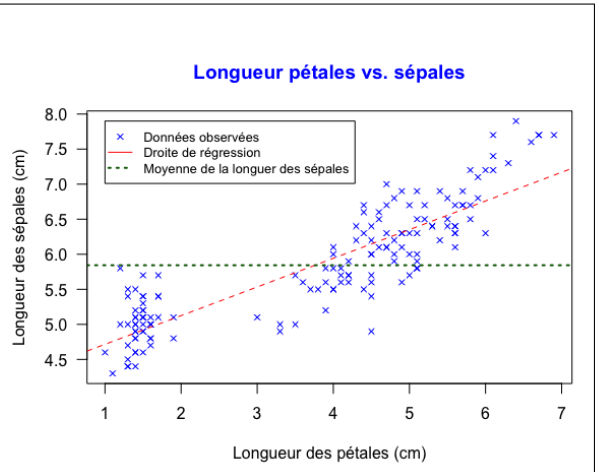

Figure 4.7: Superposition de plusieurs informations.

4.2.2 Graphique radar

Le graphique en radar est aussi appelé graphique polaire ou diagramme en étoile ou diagramme de Kiviat ou diagramme d'araignée. Ce diagramme permet d'afficher des données multidimensionnelles sous la forme d'un graphique à deux dimensions de plus de trois variables. Ce graphique comprend autant d'axes que de variables dont le nom est indiqué autour du graphique. Les axes partent tous du point central. Des segments de droite (lignes) relient les données d'une catégorie (ou classe) formant une forme polygonale. Le graphique radar a l'avantage de représenter plusieurs classes décrites par des données multidimensionnelles sur un même graphique.

▶ **Exemple 4.5** Les données IRIS sont caractérisées par 4 variables : la longueur et la largeur des sépales, la longueur et la largeur des pétales. La Figure 4.8 illustre la variation de ces variables en fonction de l'espèce sur un graphique à deux dimension. Cette figure peut être obtenue par les lignes de codes suivants :

```
install.packages('fmsb')
library(fmsb)
plot.new()
colors_border=c( rgb(0.2,0.5,0.5,0.9), rgb(0.8,0.2,0.5,0.9), rgb(0.7,0.5,0.1,0.9))
colors_in=c( rgb(0.2,0.5,0.5,0.4), rgb(0.8,0.2,0.5,0.4), rgb(0.7,0.5,0.1,0.4))
# Choisir 3 lignes(1,51,101)  de Classes differentes
radarchart( iris[c(1,51,101), 1:4]  , axistype = 1 ,
            pcol = colors_border , pfcol = colors_in ,
            plwd = 4 , plty = 1, maxmin = FALSE,
            cglcol ="grey", cglty = 1, axislabcol = "grey",
            caxislabels = seq(0,20,5), cglwd = 0.8,
            vlcex = 0.8 , title = "Diagramme radar")
```

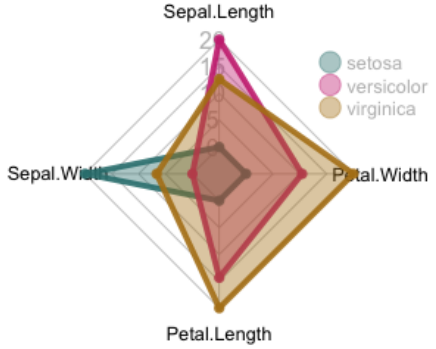

Figure 4.8: Diagramme radar des données IRIS.

