Module 3
Analyse exploratoire des données

Sommaire

Dernière mise à jour le 24 mars 2021

Introduction

L'analyse exploratoire des données consiste...

> *...plutôt que de modéliser directement les données, on s'attachera donc dans un premier temps à les décrire à l'aide de résumés numériques et graphiques. L'idée est de caractériser la forme d'une distribution et d'identifier les éventuelles valeurs influentes. (Tukey, 1977).*

3.1 Qu'est-ce qu'une donnée ?

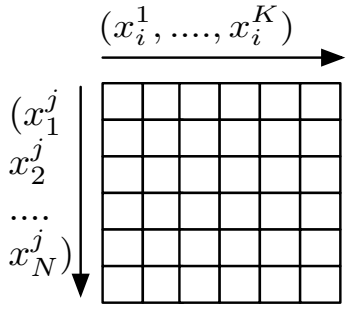
Le mot **donnée** se définit de différentes façons dans la littérature selon les domaines et les champs d'application. Mentionnons quelques exemples :
Une donnée est...

- un **enregistrement** caractérisé par un ensemble de champs (terminologie des bases de données).
- un **individu** défini par un ensemble de caractéristiques ou de paramètres ou de variables (terminologie issue de la statistique).
- une **instance** caractérisée par un ensemble d'attributs (terminologie orientée objet en informatique).
- un **point** ou un **vecteur** caractérisé par ses coordonnées dans un espace vectoriel (terminologie de l'algèbre).

Représentation des données

Les données sont généralement représentées sous la forme d'un tableau rectangulaire (ou matrice) à N lignes représentant les individus et K colonnes correspondant aux variables. On note M la matrice de dimension (N, K) contenant les données.

$$M = \begin{pmatrix} x_1^1 & . & . & x_1^K \\ x_2^1 & . & . & x_2^K \\ . & . & . & . \\ x_N^1 & . & . & x_N^K \end{pmatrix}$$



où x_i^j est la valeur de l'individu i pour la variable j.
On notera $\mathbf{x}_i = (x_i^1, ..., x_i^K)'$ le vecteur des variables de l'individu i et $\mathbf{x}^j = (x_1^j, ..., x_n^j)'$ le vecteur des individus de la variable j.

▶ **Exemple 3.1** Commençons par cet exemple tiré d'une étude dont l'objectif est d'évaluer la consommation de la crème glacée aux États-Unis. Nous disposons de la base de données `icecream` fournit dans la consigne du module 3. Les lignes de codes suivantes permettent d'interroger la base de données :

```
# Lecture des données
MyData<-read.csv("icecream.csv", header=T)
# Affichage des 6 premières lignes de l'objet "MyData"
head(MyData)
```

La base de données comprend 30 individus (désignés par `X`) décrits par les 4 variables suivantes :

1. Consommation (`cons`) : Consommation de crème glacée par habitant
2. Revenu (`income`) : revenu familial hebdomadaire en dollars
3. Prix (`price`) : prix de la crème glacée par pinte (environ 0,473 litre) en dollars
4. Température (`temp`) : température moyenne en degrés Fahrenheit

```
  X  cons income price temp
1 1 0.386     78 0.270   41
2 2 0.374     79 0.282   56
3 3 0.393     81 0.277   63
4 4 0.425     80 0.280   68
5 5 0.406     76 0.272   69
6 6 0.344     78 0.262   65
```

3.2 Type de données

La détermination du type de chaque variable est une étape nécessaire avant leur analyse. Cette étape permet de décider des méthodes d'analyse appropriées.

3.2.1 Variables qualitatives

Une variable est dite **qualitative** si ses valeurs ne sont pas mesurables. Le sexe, la profession, l'état matrimonial sont quelques exemples de variables qualitatives. Les valeurs d'une variable qualitative sont appelées **modalités**.
Une variable qualitative est dite **ordinale** si ses modalités suivent une relation d'ordre. Par exemple, une pathologie peut prendre la valeur `légère`, `modérée` ou `sévère`. Ces valeurs peuvent être ordonnées :
`légère` $<$ `modérée` $<$ `sévère`.

Une variable qualitative est dite **nominale** si ses modalités ne sont pas ordonnées naturellement. Par exemple, dans une population de personnes actives, la profession est une variable nominale.

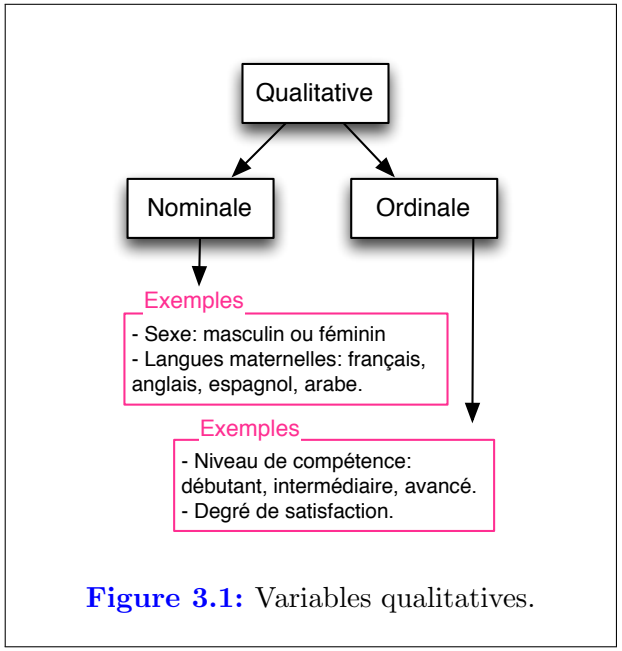


Figure 3.1: Variables qualitatives.

3.2.2 Variables quantitatives

Une variable quantitative est dite **discrète** si elle ne peut prendre que des valeurs qui peuvent être énumérées. Dans **Exemple 3.1**, les quatre variables étudiées `cons`, `income`, `price` et `temp` sont quantitatives.

La variable quantitative est dite **continue** si ses valeurs potentielles ne peuvent pas être énumérées.

Les variables binaires sont des variables quantitatives discrètes qui possèdent des propriétés particulières. Nous distinguons deux types de données binaires :

- Données **symétriques** : une variable binaire est dite symétrique si ses deux modalités ont la même importance, c'est-à-dire si celles-ci peuvent être indifféremment codées par 0 ou 1. Par exemple, la variable sexe est une variable symétrique parce qu'elle peut être codée par 0 ou 1 pour masculin de même que pour féminin sans aucune différence.
- Données **asymétriques** : une variable binaire est dite asymétrique si les deux modalités n'ont pas la même importance. Par exemple, le résultat d'un examen médical ne peut pas être codé par 0 si l'examen est positif et 1 si l'examen est négatif vu l'importance du résultat attendu de l'examen.

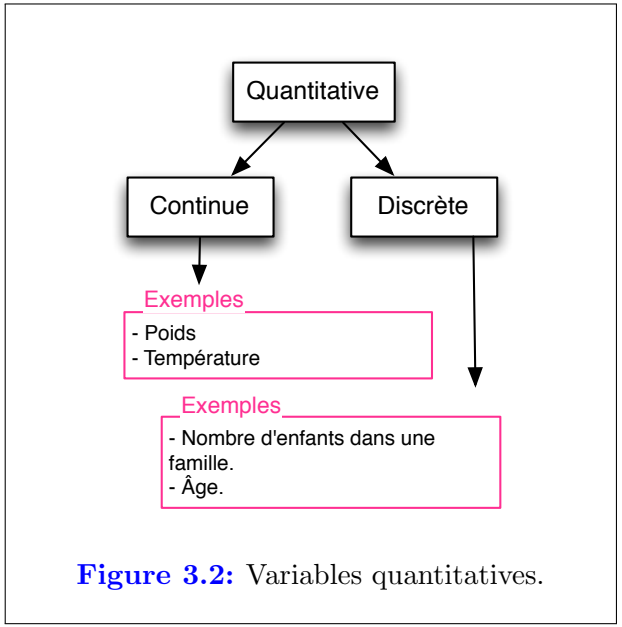


Figure 3.2: Variables quantitatives.

3.3 Description des données

3.3.1 Description des données qualitatives

Soit un individu décrit par une variable qualitative x pouvant prendre c modalités $(a_1, a_2, ..., a_i, ...a_c)$.

L'effectif

L'effectif, aussi appelé fréquence absolue, est le nombre d'individus n_i dont la variable x présente la modalité a_i.

La fréquence

La fréquence de la modalité est le rapport entre l'effectif et le nombre total d'individus.

$$f_j = \frac{n_i}{N}$$

N est le nombre total d'individus.

Les effectifs cumulés et les fréquences cumulées

Les effectifs et les fréquences cumulés sont des informations complémentaires pouvant être utiles dans le cas de variables quantitatives discrètes. Ces quantités sont respectivement définies selon les équations suivantes :

$$N_i = \sum_{j=1}^{i} n_j \quad \text{et} \quad F_i = \sum_{j=1}^{i} f_j$$

N_i peut être interprété comme étant le nombre d'individus dont l'effectif est inférieur ou égal à n_i et F_i comme leur fréquence.

► **Exemple 3.2** Cet exemple provient d'une étude portant sur l'investigation de la couleur des yeux et des cheveux et du genre de 592 étudiants. La base de données étant prédéfinie dans R, son utilisation ne nécessite pas un téléchargement préalable.

Les lignes de codes suivantes permettent d'interroger la base de données (Lignes 1-4).

La base de données comprend 592 échantillons décrits par 3 variables :

- La variable `Sex` prend la valeur `Male` ou `Female`
- La variable `Hair` prend la valeur `Black, Brown, Red` ou `Blond`

- La variable `Eye` prend la valeur `Brown, Blue, Hazel` ou `Green`

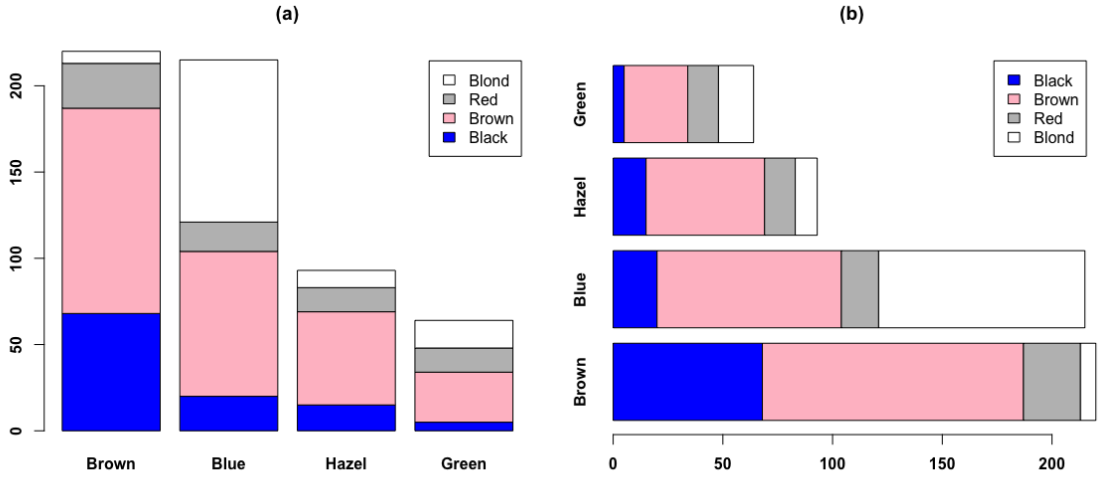
```
#Exemple 3.2
# Lecture des données
data(HairEyeColor)
# Affichage du sommaire de certaines statistiques descriptives
HairEyeColor
a <- as.table( apply(HairEyeColor, c(1,2), sum) )
# représentation graphique verticale
barplot(a, main="Eye en fonction de Hair",
legend = rownames(a), cex.axis=1, font.axis=2)
# représentation graphique horizontale
barplot(a, main="Eye en fonction Hair",
legend = rownames(a), cex.axis=1, font.axis=2, horiz=TRUE)
```

```
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Nous pouvons également, présenter les données sous forme d'un tableau (Ligne 6), dans lequel les lignes représentent les différentes modalités de la première variable, les colonnes

celles de la deuxième variable et les cases du tableau contiennent les effectifs correspondants. Ainsi l'effectif d'étudiants ayant des `Hair=Black` et `Eye=Brown` est égal à 68. Cette valeur est reportée sur les graphiques suivants en bleu foncé sur la première colonne à gauche figure (a) et sur la première ligne en bas de la figure (b).

```
a
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
```



3.3.2 Description des données quantitatives

Nous distinguons trois grandes familles de description des données quantitatives : les caractéristiques de tendances centrales, de dispersion et de formes. La description des

données est souvent illustrée par des graphiques (diagramme en bâtons, camembert, boîte à moustaches). Ces graphiques seront traités dans le module Visualisation des données.

Caractéristiques de tendances centrales

Les caractéristiques de tendances centrales indiquent l'ordre de grandeur des données et de leur valeur centrale, c'est-à-dire la position autour de laquelle se rassemblent ces valeurs.

1- La moyenne arithmétique

La moyenne arithmétique μ est la somme des valeurs de la variable j pour tous les individus i, $i = 1, ..., N$.

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_i^j \tag{3.1}$$

La moyenne arithmétique est sensible aux valeurs aberrantes.

2- La moyenne arithmétique pondérée

Lorsque les variables n'ont pas la même importance, on attribue un poids à chacune d'entre elles. Dans ce cas, on calcule la moyenne arithmétique pondérée :

$$\lambda_j = \frac{\sum_{i=1}^{N} w_i x_i^j}{\sum_{i=1}^{N} w_i} \tag{3.2}$$

3- La médiane

Soit un ensemble de N données rangées par ordre croissant. La médiane est la valeur de la variable qui partage l'ensemble des données en deux parties de même effectif.

— Si N est impair ($N = 2n + 1$), alors la médiane est la donnée de rang n.

— Si N est pair ($N = 2n$), alors la médiane est la donnée de rang n ou de rang $n+1$ ou bien la moyenne des deux.

Dans le cas d'une distribution normale, la médiane et la moyenne sont égales.

4- Les quartiles

Soit un ensemble de N données rangées par ordre croissant. Les quartiles sont les valeurs Q1, Q2, Q3 de la variable qui partagent l'effectif en quatre sous-ensembles de même effectif. Le premier quartile (Q1) est la plus petite donnée de cet ensemble telle qu'au moins un quart des données sont inférieures ou égales à Q1. Le troisième quartile (Q3) est la plus petite donnée de cet ensemble telle qu'au moins les trois quarts des données de l'ensemble de données ordonnées sont inférieurs ou égaux à Q3. Le deuxième quartile Q2 correspond à la médiane.

▶ **Exemple 3.3** Revenons à l'exemple tiré de l'étude de la consommation de la crème glacée aux États-Unis. Les lignes de codes suivantes permettent d'afficher certaines statistiques descriptives. Nous pouvons ainsi voir que pour la variable `price`, le prix minimum est 0.2600 et le prix maximum est 0.2920. Les trois quartiles sont respectivement, 0.2685, 0.27770 et 0.2815.

```
# Lecture des données
MyData<-read.csv("icecream.csv", header=T)
# Affichage du sommaire de certaines statistiques descriptives
summary(MyData)
```

```
       X              cons            income          price             temp
 Min.   : 1.00   Min.   :0.2560   Min.   :76.00   Min.   :0.2600   Min.   :24.00
 1st Qu.: 8.25   1st Qu.:0.3113   1st Qu.:79.25   1st Qu.:0.2685   1st Qu.:32.25
 Median :15.50   Median :0.3515   Median :83.50   Median :0.2770   Median :49.50
 Mean   :15.50   Mean   :0.3594   Mean   :84.60   Mean   :0.2753   Mean   :49.10
 3rd Qu.:22.75   3rd Qu.:0.3912   3rd Qu.:89.25   3rd Qu.:0.2815   3rd Qu.:63.75
 Max.   :30.00   Max.   :0.5480   Max.   :96.00   Max.   :0.2920   Max.   :72.00
```

Ces statistiques peuvent également être mesurées en utilisant les lignes de codes suivants.

```
mean(MyData$price)     # moyenne
median(MyData$price)   # Médiane
quantile(MyData$price, c(0.25, 0.5, 0.75), type = 1) #Quantiles
```

```
> mean(MyData$price)    # moyenne
[1] 0.2753
> median(MyData$price)  # Médiane
[1] 0.277
> quantile(MyData$price, c(0.25, 0.5, 0.75), type = 1) # Les trois quantiles
25%   50%   75%
0.268 0.277 0.282
```

Caractéristiques de dispersion

Comme leur nom l'indique, les caractéristiques de dispersion servent à évaluer la variabilité des données et à résumer l'éloignement de l'ensemble des individus par rapport à leur tendance centrale.

1- L'étendue

L'étendue est l'écart entre la plus grande et la plus petite des valeurs. Cette caractéristique, qui dépend des valeurs aberrantes, est par conséquent peu fiable.

2- La variance et écart-type

La variance est la moyenne des carrés des écarts à la moyenne.

$$var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2$$

avec μ la moyenne arithmétique de l'ensemble de données.

L'écart-type σ_x est la racine carrée positive de la variance :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Les caractéristiques de dispersion peuvent être directement obtenues à partir de fonction statistiques prédéfinies dans R.

▶ **Exemple 3.4** Revenons à l'exemple tiré de l'étude de la consommation de la crème glacée aux États-unis. Les lignes de codes suivantes permettent d'obtenir la variation standard (var), l'écart type (sd) et l'étandue (range).

```
var(MyData$cons)        # Variation standard
sd(MyData$cons)         # Écart type
range(MyData$cons)      # Étendue
```

Caractéristiques de formes

De nombreux phénomènes physiques se distinguent par des variables quantitatives qui suivent une loi normale. Ces principales caractéristiques de formes sont le coefficient d'asymétrie et l'aplatissement de Fisher.

Asymétrie

Le coefficient d'asymétrie (*skewness* en anglais) correspond au moment d'ordre trois de la variable centrée réduite. Pour une distribution uniforme x, le coefficient d'asymétrie est donné par la formule :

$$\gamma_1 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

où E désigne l'espérance de x, μ la moyenne et σ l'écart type. Lors que cette espérance existe,

$$\gamma_1 = \frac{\mu_3}{{\mu_2}^{3/2}} \tag{3.3}$$

avec μ_i les moments centrés d'ordre i. Le moment centré d'un échantillon de donnée est fournit par :

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$$

Trois cas se présentent :

- $\gamma_1 = 0$ si la distribution de la variable est symétrique
- $\gamma_1 > 0$ si la distribution de la variable est décalée vers la gauche et la moyenne est supérieure à la médiane
- $\gamma_1 < 0$ si la distribution de la variable est décalée vers la droite gauche et la moyenne est inférieure à la médiane

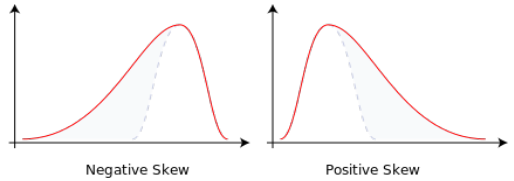


Figure 3.3: Formes de distribution (image extraite de wikepedia).

▶ **Exemple 3.5** Nous considérons dans cet exemple la base de données disponible dans R dénommée `faithful` qui contient des données concernant les éruptions d'un geyser situé dans le parc de Yellowstone aux États-Unis. `faithful` contient 272 échantillons décrits par deux attributs : `eruption` qui indique la durée de l'éruption et `waiting` qui indique l'intervalle de temps entre deux éruptions.

Dans la suite, nous allons calculer le coefficient d'asymétrie de la variable `eruption` selon deux méthodes. La méthode directe qui utilise la fonction `skewness` de R et la méthode indirecte qui passe par le calcul des moments.

```
1 library(e1071)                    # load e1071
2 summary(faithful)                 # Sommaire des attributs
3 eruptions = faithful$eruptions    # Durée de l'éruption
4 hist(eruptions)                   # Histogramme de la durée
5 skewness(eruptions)               # Application de fonction skewness
```

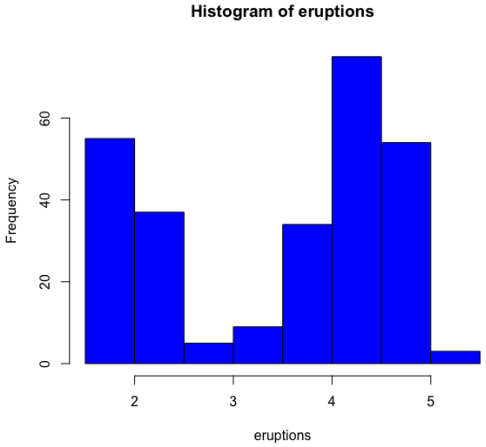


Figure 3.4: Distribution de la variable `eruptions`.

```
    eruptions        waiting
 Min.   :1.600   Min.   :43.0
 1st Qu.:2.163   1st Qu.:58.0
 Median :4.000   Median :76.0
 Mean   :3.488   Mean   :70.9
 3rd Qu.:4.454   3rd Qu.:82.0
 Max.   :5.100   Max.   :96.0
[1] -0.4135498
```

Le coefficient d'asymétrie est égal à -0.41 ; il est donné par la fonction `skewness` (Ligne 5 du code R). Cette valeur étant non nulle ce qui laisse prédire que la distribution de la variable `eruption` est non symétrique. De plus, cette valeur est négative, donc la moyenne est inférieure à la médiane (`Mean`=3.488 < `Median`=4.000) et l'allure de la distribution est plus décalée vers la droite. Ceci est confirmé par l'histogramme de cette variable donné par la figure 3.4.
Le coefficient d'asymétrie peut être aussi calculer à travers des moments d'ordre 2 et 3 et en utilisant l'équation 3.3. Les lignes de codes suivants aboutissent presque au même résultat pour le calcul du coefficient d'asymétrie qui est égale à -0.47.

```
1 library(e1071)                      # load e1071
2 summary(faithful)                   # Sommaire des attributs
3 duration = faithful$eruptions       # Durée de l'éruption
4 mu2 <- moment(duration, order=2, center=TRUE)   # Moment d'ordre 2
5 mu3 <- moment(duration, order=3, center=TRUE)   # Moment d'ordre 3
6 sk1 = mu3/mu2^(3/2)
```

```
[1] -0.4737556
```

Aplatissement

Le coefficient d'aplatissement (*kurtosis* en anglais) correspond au quotient du moment d'ordre quatre de la variable centrée réduite par la puissance quatrième de l'écart type.

Il est donné par l'équation suivante :

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma_x}\right)^4\right]$$

- Si $\gamma_2 = 0$ alors la distribution est dite **musokurtique**.
- Si $\gamma_2 > 0$, alors la distribution est plus concentrée que la normale ; elle est dite **leptokurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.
- Si $\gamma_2 < 0$, alors la distribution est plus aplatie que la normale. Elle est dite **platikurtique**. Dans le domaine de la finance, ce coefficient sert à déterminer des valeurs anormales plus fréquentes.

3.4 Préparation des données

La préparation des données consiste à relever les données aberrantes et manquantes pour cibler leur traitement.

3.4.1 Données aberrantes

Les données aberrantes (*outliers* en anglais) sont des valeurs extrêmes par rapport à l'ensemble des données à analyser (Figure 3.5). Elles sont souvent causées par une erreur commise lors de leur acquisition ou de leur transcription. Cependant, dans certains cas, elles peuvent correspondre à des observations réelles mais particulières. Les données aberrantes ne doivent pas faire l'objet d'un rejet systématique. En effet, leur rejet peut entraîner, dans certains cas, une perte d'informations réelles. De plus, le rejet des valeurs aberrantes peut avoir des conséquences statistiques non négligeables.

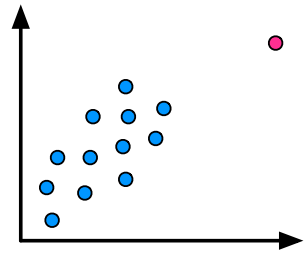


Figure 3.5: Données aberrantes.

Détection des valeurs aberrantes

Il existe plusieurs méthodes de détection des données aberrantes. L'utilisation de la déviation standard décrite dans ce qui suit est celle la plus couramment utilisée. Les données aberrantes sont définies comme étant celles dont la valeur se situe à l'extérieur des limites calculées à l'aide de l'écart-type :

$$L_{inf} = \mu - \epsilon\sigma \quad \text{et} \quad L_{sup} = \mu + \epsilon\sigma \tag{3.4}$$

ou μ est la moyenne de la variable, σ est l'écart type et ϵ est la pondération de l'écart type. La valeur de ϵ est souvent fixée entre 1,5 et 3.

▶ **Exemple 3.6** Nous traitons dans cet exemple la base de données disponible dans R dénommé `Cars` et qui contient 50 voitures caractérisées chacune par deux variables : `dist` qui mesure la distance parcourue par chaque voiture entre le moment d'un freinage et l'immobilisation complète de la voiture en question. La variable `speed` mesure la vitesse de déplacement de la voiture tout juste avant le freinage. Nous considérons les 30 premiers échantillons de cette base de données (Figure 3.6) pour former `Cars1` auxquels nous ajouterons 5 échantillons avec des valeurs aberrantes pour avoir `Cars2` (Figure 3.7).

```
cars1 <- cars[1:30, ]
cars_outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218))  # Ajouter des outliers.
cars2 <- rbind(cars1, cars_outliers)
plot(cars1$speed, cars1$dist, xlim=c(0, 28), ylim=c(0, 230), main="Without Outliers", xlab="speed", ylab="dist", pch="*", col="red", cex=2)  #Graphique sans les outliers
plot(cars2$speed, cars2$dist, xlim=c(0, 28), ylim=c(0, 230), main="With Outliers", xlab="speed", ylab="dist", pch="*", col="red", cex=2) #Graphique avec les outliers
Mean(cars2$dist)
Lower = mean(cars2$dist) - (1.5*sd(cars2$dist))  #Limite inférieure
Upper = mean(cars2$dist) + (1.5*sd(cars2$dist))  #Limite supérieure
```

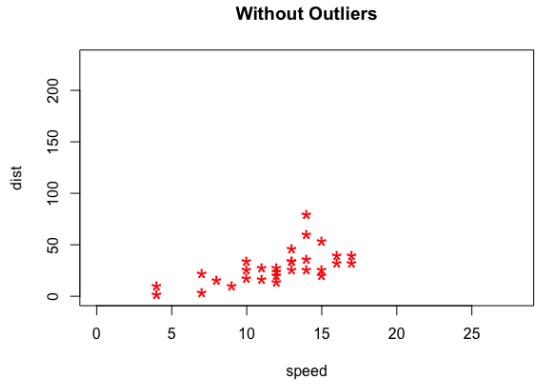


Figure 3.6: Variation de la distance en fonction de la vitesse dans `Cars1`.

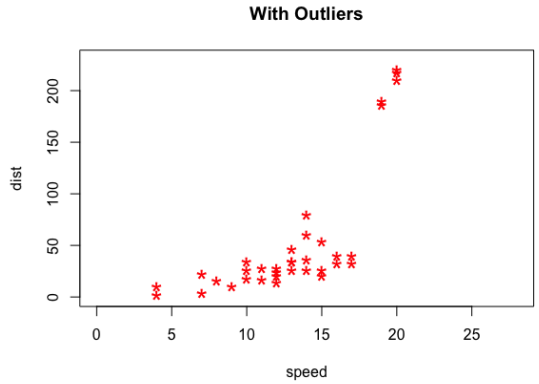


Figure 3.7: Variation de la distance en fonction de la vitesse dans `Cars2`.

Les lignes de codes précédentes permettent de former les deux bases de données et de déterminer formellement selon les relations 3.4 les bornes inférieure et supérieure pour la détection de outliers (Lignes 7 et 8). La valeur de ϵ est fixé à 1.5. Ces bornes évaluées respectivement à -43.09 et 150.69 permettent de supprimer tous les échantillons, considérés comme aberrantes, puisque leur valeur de distance est supérieure à 150.69.

```
> Lower
[1] -43.09596
> Upper
[1] 150.696
> cars2$dist
 [1]   2  10   4  22  16  10  18  26  34  17  28  14  20  24  28  26  34  34  46
 26  36  60  80  20  26  54  32  40  32  40 190 186 210 220 218
```

Traitement des valeurs aberrantes (*Outliers*)

De ce fait, on peut traiter ces valeurs aberrantes pour éviter leur rejet de différentes manières :

1. Les valeurs aberrantes peuvent être causées par une erreur de saisie. Dans ce cas, il faut retourner à la source d'information afin de les corriger.
2. S'il n'est pas possible de retourner à la source d'information, les valeurs aberrantes peuvent être supprimées et remplacées par des valeurs imaginaires. Ces dernières correspondent à des valeurs interpolées à partir de l'ensemble des données par exemple, la moyenne, la médiane et les valeurs des plus proches voisins.
3. Finalement, on peut garder les valeurs aberrantes et adopter des méthodes qui diminuent leur impact au cours des analyses.

3.4.2 Données manquantes

Les données manquantes (*missed data* en anglais) sont des données incomplètes, c'est-à-dire des données pour lesquelles certaines variables sont inconnues. Les données man-

quantes ne peuvent pas être ignorées systématiquement. Leur traitement dépend de leur proportion par rapport à l'ensemble des données. Si cette proportion est faible, les données manquantes sont retirées. Sinon, elles peuvent être remplacées par des valeurs interpolées. Aussi, certaines méthodes d'analyse peuvent être réalisées malgré des données manquantes. La commande `!is.na` permet de supprimer les échantillons ayant des données manquantes.

