Statistique descriptive

Maximilien DIALUFUMA VAKAMBI

Guide pour étudiant(e)s

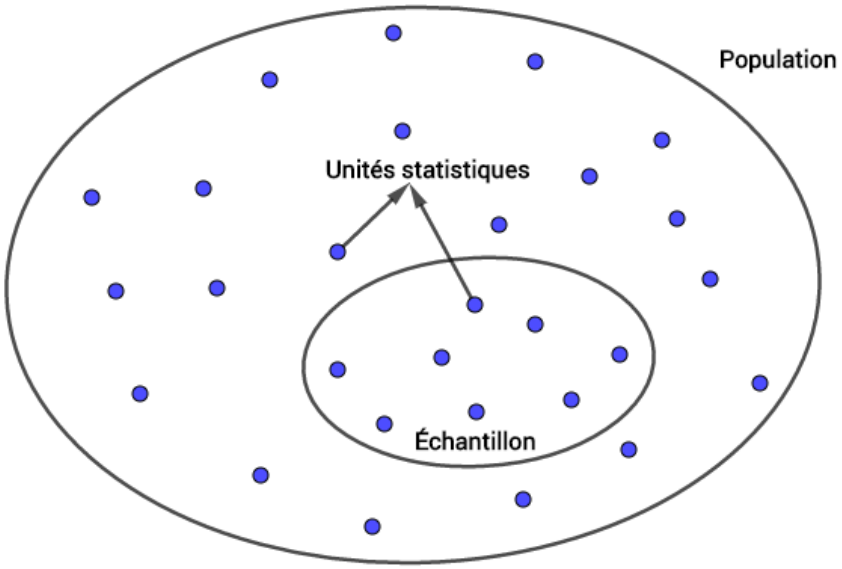

Table des matières

Avant propos

Ce document s'adresse aux étudiants du premier cycle qui veulent apprendre la statistique descriptive. Il part du vocabulaire de la statistique descriptive, aux calculs des mesures statistiques, à l'interprétation de ces mesures et à la visualisation de données selon les types des variables.

Nous initions les étudiants à l'utilisation du logiciel R ("*Rstudio*") comme outil pour l'analyse descriptive des données. Les codes sont fournis à la fin de chaque chapitre. Il n'y a pas des cours préalables pour faire ce cours.

Le contenu de ce document n'engage que l'auteur. Pour toute observation ou erreur, prière de me contacter par e-mail [1].

1. E-mail : maximilien.dialufuma.1@ulaval.ca/mdialufuma@gmail.com

Chapitre 1

Initiation à la statistique

La statistique est définie comme l'ensemble des méthodes à partir desquelles sont *recueillies*, *organisées*, *analysées* et *interpretées* des données. Il faudrait faire la différence entre la statistique et les statistiques. La statistique est une science ou une branche des mathématiques tandis que les statistiques sont des données ou des résultats obtenus à partir d'observations. Le but de la statistique est d'augmenter la compréhension que l'on a d'un phénomène à partir de données issues de celui-ci.

1.1 Définition

Dans cette section, nous définissons les terminologies (ou jargons) de la statistique utile pour la suite des chapitres.

Population

C'est l'ensemble de référence sur lequel porte l'étude dans laquelle les données ont été recueillies.

Individu ou Unité statistique

C'est un élément de la population. L'ensemble des individus qui constitue la population.

C'est l'unité sur laquelle les mesures sont prises. Parfois on l'appelle aussi unité d'échantillonnage.

Unité expérimentale

C'est l'unité à laquelle le traitement est appliqué dans une étude expérimentale.

Échantillon

C'est un sous-groupe de la population, composé des unités statistiques pour lesquelles des observations ont été recueillies.

Lorsque les mesures ont été prises pour tous les individus de la population, on parle de *recensement*.

Variable

C'est une mesure prise sur un individu. Elle représente une caractéristique des individus.

Observation

C'est l'ensemble des valeurs obtenues en mesurant des variables sur un individu de la population.

Paramètre

C'est une mesure qui caractérise la variable dans la population.

En général, les paramètres sont inconnus (Ex : la moyenne de la population, la variance de la population).

Statistique

C'est une mesure qui caractérise la variable dans un échantillon.

Exemple : la variance d'un échantillon, moyenne d'un échantillon).

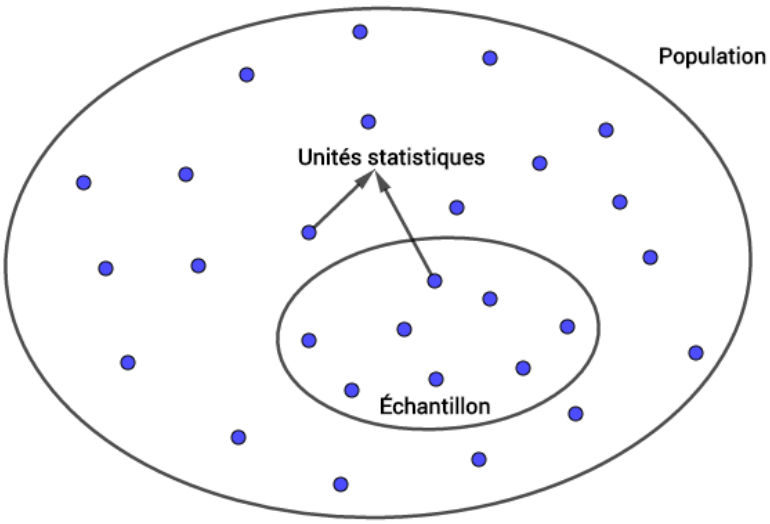

FIGURE 1.1 – Représentation d'une population, des unités statistiques et d'un échantillon.

Prenons un exemple dans lequel nous sortirons certaines définitions présentées ci-dessus.

Exemple 1.1.

Un biologiste étudie une certaine espèce de serpents. Il s'intéresse en particulier au poids des serpents à la naissance. Il obtient les poids de 70 serpents nouveau-nés. Ces poids sont mesurés en grammes.

Dans cet exemple, la *population* qui nous intéresse est l'ensemble de tous les serpents de l'espèce. L'*échantillon* est l'ensemble des 70 serpents obtenus par le biologiste. La *variable statistique* qui nous intéresse est la variable poids à la naissance. Le tableau suivant présente la liste de 70 poids mesurés par le biologiste.

33.90	34.87	34.49	34.16	35.14	34.10	34.41
33.10	35.59	35.20	34.35	33.87	35.18	34.54
35.71	34.74	36.42	34.68	34.05	34.76	35.49
35.44	36.03	34.19	35.49	35.30	34.26	35.36
35.83	33.64	34.83	36.11	35.32	37.37	34.78
36.66	33.91	33.55	34.91	34.17	34.96	34.71
34.31	35.78	34.86	34.19	35.50	32.62	33.20
35.52	34.59	35.32	36.21	35.61	36.14	34.31
34.55	37.61	35.86	33.75	34.77	34.37	33.68
35.70	35.65	35.67	34.20	34.11	35.18	35.07

TABLE 1.1 – Données brutes sur la mesure du poids des serpents à la naissance

1.2 Types de variables statistiques

La plus part des variables statistiques qu'on utilise sont régroupées en deux catégories à savoir : les *variables catégoriques* et les *variables numériques*.

1.2.1 Variables catégoriques (ou qualitatives)

Une variable qualitative admet des valeurs possibles (ou modalités) qu'on appelle les *catégories*. Elle se subdivise en deux groupes :

- **Variables nominales :** Ces sont les variables catégoriques dont les modalités ne peuvent être ordonnées.
  Exemple : le sexe d'une personne, la coleur de ses cheveux, la couleur de ses yeux, son état civil.
- **Variables ordinales :** Ces sont les variables catégoriques dont les modalités peuvent être ordonnées.
  Exemple : Un niveau de satisfaction, un niveau de scolarité, univeau d'apprentissage.

1.2.2 Variables numériques (ou quantitatives)

Ces sont des variables qui peuvent être mesurées numériquement. Elle se subdivise aussi en deux groupes :

- **Variables discrètes :** Ces sont les variables qui ne peuvent prendre qu'un nombre fini de valeurs.
  Exemple : le nombre de machines dans une entreprise, le nombre d'étudiants à la faculté des sciences de l'université de Kinshasa.
- **Variables continues :** Ces sont les variables qui peuvent prendre comme valeurs tous les points d'un intervalle de nombres réels.
  Exemple : le poids d'une personne, sa taille, la durée de vie d'une machine.

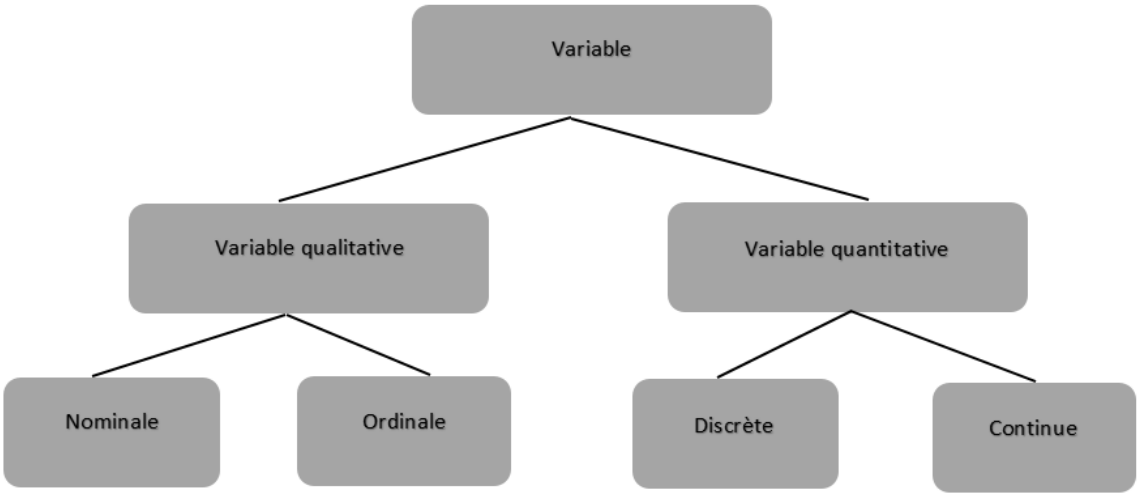

FIGURE 1.2 – Types de varaibles

1.3 Méthodes de collecte des données

La collecte des données est une opération qui peut porter sur l'ensemble des unités statistiques (population) ou sur un groupe des individus (échantillon). Cependant, faire la collecte des données est un travail qui peut rencotrer plusieurs contraintes : le coût, manque de personnel, faible taux de réponse, le délai d'exploitation et la qualité des résultats, etc...

L'alternative dans ce cas est de constitué un sous-groupe *représentatif* extrait de la population qui recouvre les caractéristiques des paramètres que l'on souhaite estimer.

1.3.1 Répresentivité d'un échantillon

L'échantillonnage probabiliste utilisé aujourd'hui a été développé dans un article de Neyman [1] publié en 1934. Il définit un échantillon représentatif comme étant sélectionné selon un *plan de sondage aléatoire* sous le contrôle d'un statisticien. Il montre également comment caractériser la qualité des estimations calculées à partir de l'échantillon à l'aide de la théorie des probabilités.

Un échantillon est dit *représentatif* s'il renferme toutes les caractéristiques d'une population. Ce qui revient à dire que chaque élément de la population a une même chance d'appartenir à un même échantillon. Les unités statistiques doivent être

1. Jerzy Neyman (16 avril 1894 - 5 août 1981) est considéré comme un des grands fondateurs de la statistique moderne. Il a contribué très largement à la théorie des probabilités, vérifiant les hypothèses, les intervalles de confiance et d'autres parties des statistiques.

tirées au hasard (échantillon aléatoire).

Pour obtenir un échantillon représentatif, on suppose que la population est homogène face au caractère faisant l'objet de l'étude statistique.

Il existe deux types de méthode :

- **Les méthodes aléatoires ;**
- **Les méthodes non aléatoires ou à choix raisonné.**

Les méthodes aléatoires

Les méthodes aléatoires sont les plus utilisées en pratiques, car elles sont basées sur l'échantillonage. Cette approche consiste à sélectionner au *hasard* certaines unités de la population et à estimer la caractéristique d'intérêt à l'aide des seuls individus échantillonnés. Les méthodes d'échantillonnage sont beaucoup utilisées par des institutions (Institut National de Statistique, Banque Centrale du Congo) pour obtenir de l'information sur des populations cibles.

Comme c'est un cours de statistique descriptive, on ne pourrait aller loin sur la *théorie de l'échantillonage* [7, 9], car elle nécessite un cours préable de la *statististique inférentielle* [1, 6, 8], pour la construction d'estimateurs des caractéristiques de la population à l'aide des données de l'échantillon et l'étude de leurs propriétés statistiques (biais, variance).

Néamoins nous pourrons citer quelques méthodes pour sélectionner des échantillons aléatoires d'unités suivant un plan de sondage.

- **Plans de sondage simples :**
  - à probabilités égales ;
  - à probabilités inégales.
- **Plans de sondage complexes :**
  - stratifié
  - en grappe
  - plusieurs degrés

Les méthodes non aléatoires ou à choix raisonné

Dans ces méthodes, l'échantillon obtenu est constitué d'unités statistiques qui n'ont pas été tirés au hasard. Nous pourrons citer certaines méthodes de :

- Convenance ;
- Échantillonnage par réseau (ou boule de neige) ;
- Quotas ;
- Volontariat.

1.3.2 Questionnement de la statistique

Dans une étude statistique, trois questions sont souvent posées en vue de mener une bonne analyse.

- **Qui ?**
  - Quelles sont les unités statistiques à l'étude ?
  - Combien y en a-t-il ?
  - Population ou échantillon ?

- **Quoi ?**
  - Quelles sont les variables et les unités de celles-ci ?
  - Comment sont-elles mesurées ?

- **Pourquoi ?**
  - Quels sont les objectifs poursuivis dans l'étude ?
  - À quelles questions veut-on répondre ?
  - Quelle est la portée recherchée des conclusions ?

Ces questions permettent de faire un bon diagnostique pour une étude statistique.

Exercice 1.1.

1. Selon le domaine de votre étude :

   (i) donnez deux exemples de variables quantitatives continues ;
   (ii) donnez deux exemples de variables quantitatives discrètes ;
   (iii) donnez trois exemples de variables qualitatives ordinales ;
   (iv) donnez trois exemples de variables qualitatives nominales.

2. Déterminez la nature (quantitative continue, quantitative discrète, qualitative ordinale et qualitative nominale) de chacune des variables suivantes :

   (i) Le salaire d'un ministre ;
   (ii) Les couleurs de l'arc-en-ciel ;
   (iii) Le poids d'un bébé à la naissance ;
   (iv) Le nombre des carrefours dans la ville de Kinshasa ;
   (v) La mention (S, D, GD, LPGD) d'un étudiant pour une année académique ;
   (vi) Le nombre des étudiants à l'univsersité de Kinshasa ;
   (vii) Le niveau d'appprentissage d'une personne (Excellent, bon,ni bon/ni mauvais ,mauvais, médiocre) d'un produit dans le marché ;
   (ix) La longueur d'un arbre.

Chapitre 2

Représentation des variables

Dans ce chapitre, il est question de manipuler (ou organiser) les données sous la forme d'un tableau statistique et obtenir sa représentation graphique.

Nous considérons une suite des valeurs prises par une variale Y appelée *série statistique* sur les unités d'observations. On note $y_1, \ldots, y_n$ les valeurs distincts de la variable Y et n le nombre d'observation.

2.1 Définitions

Effectif ou Fréquence absolue

C'est le nombre de fois qu'une valeur observée (ou modalité) apparait dans une serie statistique. On le note n_i, avec $i = 1, \ldots, I$.

Effectif cumulé

C'est la somme des effectifs précédents. On le note N_i, avec $i = 1, \ldots, I$.

$$N_i = \sum_{k=1}^{i} n_k = 1,\ k = 1, \ldots, i$$

Fréquence relative

C'est l'effectif divisé par le nombre d'unités d'observation.

$$f_i = \frac{n_i}{n},\ i = 1, \ldots, I \tag{2.1}$$

Une propriété de la fréquence relative est

$$\sum_{i=1}^{n} f_i = 1 \tag{2.2}$$

En effet,

$$\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{n} n_i = \frac{n}{n} = 1$$

En général, la fréquence relative s'exprime toujours en pourcentage (%).

Fréquence relative cumulée

C'est la somme des fréquences relatives précédentes. On la note F_i, avec $i = 1, \ldots, I$.

$$F_i = \sum_{k=1}^{i} f_k = 1,\ k = 1, \ldots, i$$

De même la fréquence relative s'exprime aussi en pourcentage.

2.2 Variable qualitative nominale

Considérons une variable Y répresentant l'état civil d'une personne ayant les modalités [Célibataire (C), Marié(e) (M), Divorcé(e) (D), Autres (A)] dans la série statistique suivante avec $n = 24$.

D	C	M	M	C	A	D	C
D	M	A	M	D	C	A	M
D	A	M	C	A	M	A	M

TABLE 2.1 – Série statistique sur l'état civil

2.2.1 Tableau des données

y_i	n_i	f_i
A	6	0.25
C	5	0.21
D	5	0.21
M	8	0.33
Total	$n = \sum_{i=1}^{4} n_i = 24$	1

TABLE 2.2 – Tableau des données

2.2.2 Représentation graphique

Pour une variable qualitative nominale, la répresentation graphique peut se faire en utilisant un *diagramme en barres* pour les effectifs et un *diagramme en secteurs (ou camembert)* pour les fréquences relatives.

Diagramme en barres

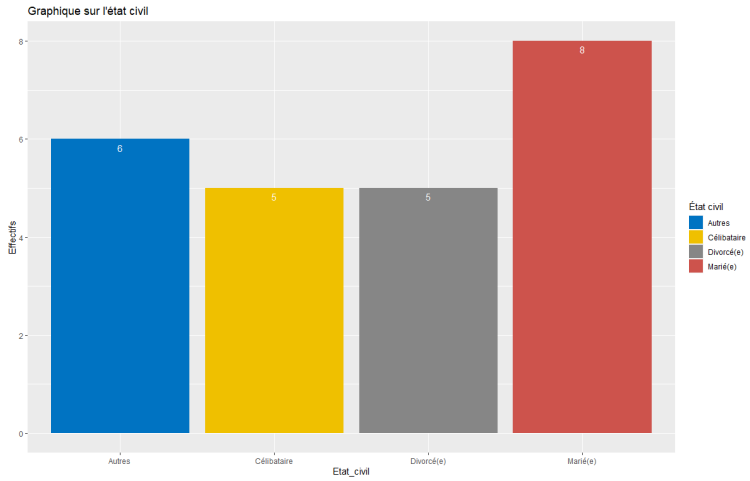

FIGURE 2.1 – Diagramme en barres pour la variable état civil

Diagramme en secteurs (ou camembert)

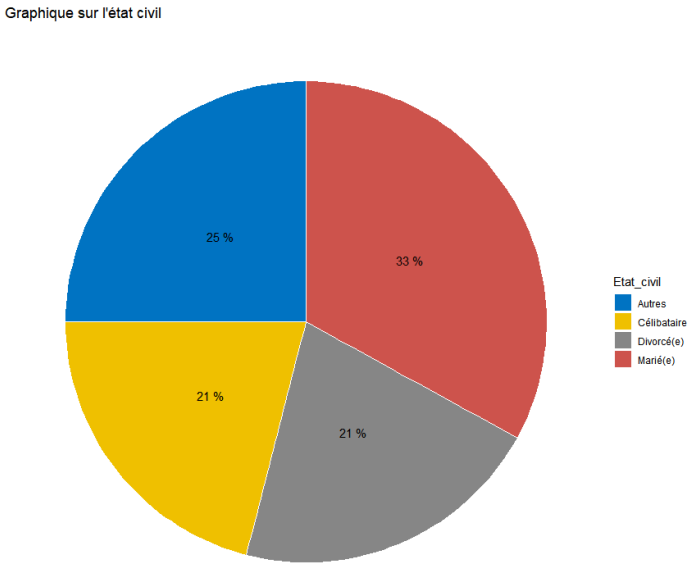

FIGURE 2.2 – Diagramme en secteurs pour la variable état civil

2.3 Variable qualitative ordinale

Considérons une variable Y répresentant le niveau de satisfaction d'un produit de vente à la clientèle basé sur l'échelle de Likert [1] ayant les modalités

- Très satisfait (Tr)
- Plutôt satisfait (Pl)
- Ni satisfait, ni insatisfait (Ni)
- Plutôt insatisfait (Pl.ins)
- Très insatisfait (Tr.ins)

dans la série statistique suivante avec $n = 40$.

Tr	Tr	Tr	Tr	Tr	Tr	Tr	Tr	Tr	Tr
Tr	Tr	Tr	Tr	Tr	Tr	Tr	Tr	Pl	Pl
Pl	Pl	Pl	Pl	Pl	Pl	Pl	Pl	Ni	Ni
Ni	Pl.ins	Pl.ins	Pl.ins	Pl.ins	Pl.ins	Tr.ins	Tr.ins	Tr.ins	Tr.ins

Table 2.3 – Série statistique sur le niveau de satisfaction

2.3.1 Tableau des données

Avec cet exemple, le tableau statistique se présente comme suit :

y_i	n_i	N_i	f_i	F_i
Tr	18	18	0.45	0.45
Pl	10	28	0.25	0.7
Ni	3	31	0.075	0.775
Pl.ins	5	36	0.125	0.9
Tr.ins	4	40	0.10	1
Total	$n = \sum_{i=1}^{4} n_i = 40$		1	

Table 2.4 – Tableau des données

2.3.2 Représentation graphique

Pour une variable qualitative ordinale, la répresentation graphique peut se faire en utilisant des *diagramme en barres* pour les effectifs et les effectifs cumulés, ainsi qu'un *diagramme en secteurs (ou camembert)* pour les fréquences relatives.

1. Parfois appelée "échelle de satisfaction", l'échelle de Likert comprend cinq ou sept options de réponse, qui couvrent le spectre d'opinions, d'un extrême à l'autre.

Diagramme en barres pour les effectifs

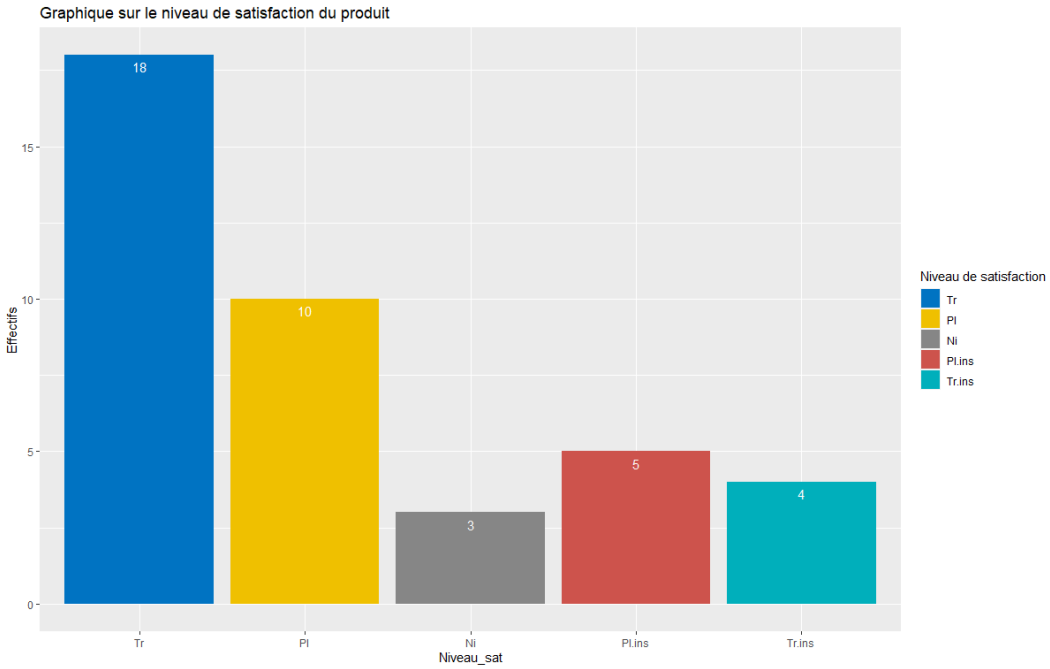

FIGURE 2.3 – Diagramme en barres pour la variable niveau de satisfaction

Diagramme en barres pour les effectifs cumulés

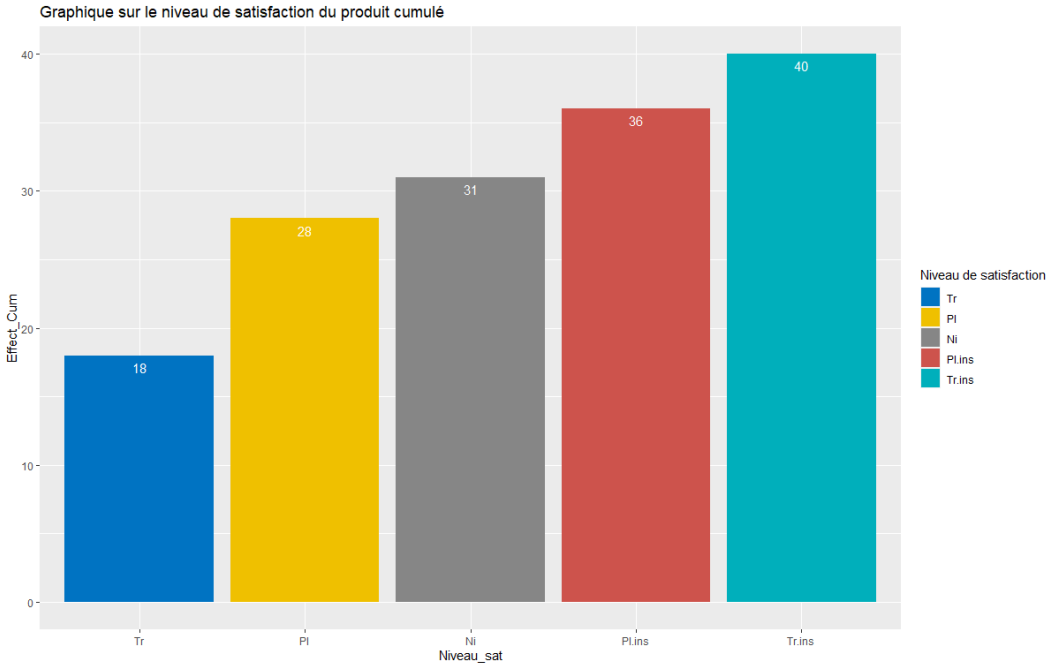

FIGURE 2.4 – Diagramme en barres pour la variable niveau de satisfaction cumulés

Diagramme en secteurs (ou camembert)

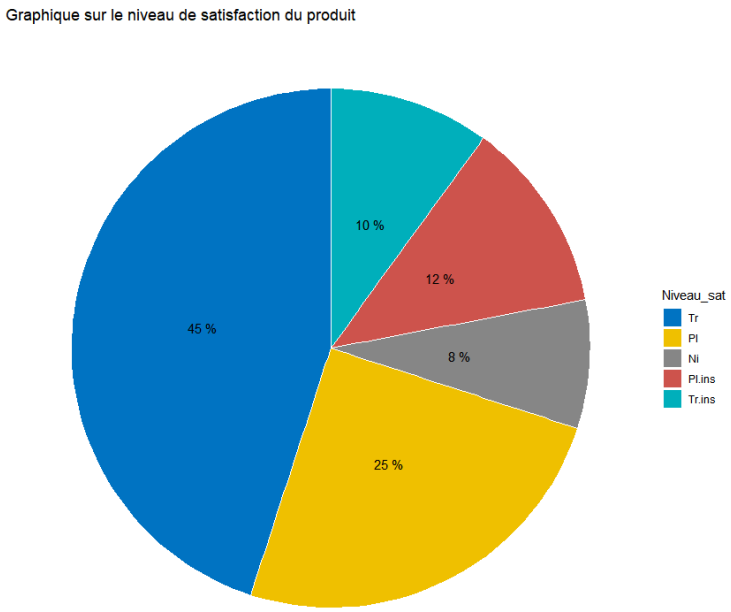

FIGURE 2.5 – Diagramme en secteurs pour la variable niveau de satisfaction du produit

2.4 Variable quantitative discrète

Considérons une variable Y répresentant le nombre d'individus par ménage, l'étude est réalisée dans un village composé de 30 ménages.

1	1	1	2	2	2	2	3	3	3
3	3	3	4	4	4	4	4	4	4
5	5	5	6	6	7	7	7	8	8

TABLE 2.5 – Série statistique sur le nombre d'individus par ménage

2.4.1 Tableau des données

Avec cet exemple, le tableau statistique se présente comme suit :

y_i	n_i	N_i	f_i	F_i
1	3	3	0.10	0.10
2	4	7	0.1333	0.2333
3	6	13	0.2	0.4333
4	7	20	0.2333	0.6667
5	3	23	0.10	0.7667
6	2	25	0.0667	0.8333
7	3	28	0.10	0.9333
8	2	30	0.0667	1
Total	$n = \sum_{i=1}^{4} n_i = 30$		1	

TABLE 2.6 – Tableau des données

On constate qu'on peut calculer les effectifs, les effectifs cumulés, les fréquences relatives et les fréquences relatives cumulées pour les variables quantitatves discrètes comme nous l'avions fait précédement pour les variables qualitatives ordinales.

2.4.2 Représentation graphique

Pour une variable quantitative discrète, la répresentation graphique peut se faire en utilisant le *diagramme en bâtonnet des effectifs*.

Diagramme en bâtonnet des effectifs

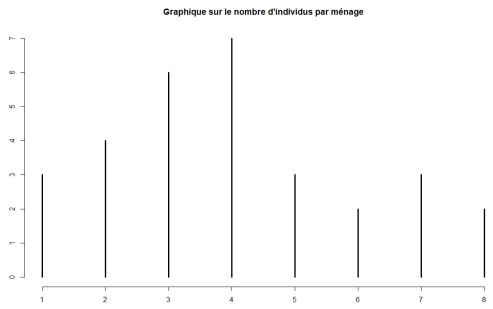

FIGURE 2.6 – Diagramme en bâtonnet pour le nombre d'individus par ménage

2.5 Variable quantitative continue

Pour une variable quantitative continuée dont les valeurs possibles peuvent être dans un intervalle de $\mathbb{R}$. Les représentations graphiques et la construction du tableau statistique se font au moyen de regroupements en classe. Le tableau statistique dans lequel les données sont regroupées en classe est appelé une *distribution de fréquence* ou *une distribution groupée*.

2.5.1 Construction des classes

La construction des classes doit être bien définies de sorte que chaque observation soit contenue dans une et une seule classe. En général le nombre de classes ne peut pas être ni trop grand ni trop petit (entre 6 à 12).

Il existe des formules qui permettent de calculer le nombre des classes et l'intervalle de classe qu'on appelle aussi l'*amplitude* pour une série statistique de n observations. Les formules suivantes permettent de calculer le nombre des classes.

- **La règle de Sturges** [2] **:**

$$K = 1 + \left(3.3 \log_{10} n\right)$$

- **La règle de Yule** [3] **:**

$$K = 2.5 \sqrt[4]{n}$$

L'intervalle de classe est calculé comme suit :

$$I = \frac{\left(y_{(n)} - y_{(1)}\right)}{K},$$

où $y_{(n)}$ (resp. $y_{(1)}$) représente la plus grande (resp. la plus petite) valeur observée.

Pour une classe i définie comme suit $\left[a_i^-, a_i^+\right[$, on note :

- a_i^- : la borne inférieure de la classe i,
- a_i^+ : la borne supérieure de la classe i,
- $x_i = \left(a_i^+ + a_i^-\right)/2$: le centre de la classe i,
- $A_i = \left(a_i^+ - a_i^-\right)$: l'amplitude de la classe i,
- n_i : l'effectif de la classe i,
- N_i : l'effectif cumulé de la classe i,
- f_i : la fréquence relative de la classe i,
- F_i : la fréquence relative cumulée de la classe i.

2. La règle de Sturges est une formule mathématique proposée par Herbert Sturges (1882-1958).
3. George Udny Yule était un statisticien écossais(1871-1951). Dans les années 1920, ses recherches sur les séries chronologiques l'amenèrent à formuler la notion de processus autorégressif.

Il est toujours récommandé d'*arrondir* le nombre de classe J à l'entier le plus proche. Il faudrait se rassurer que le début de la classe contienne la première valeurs observée.

Exemple : Un biologiste fait une étude consistant à mesurer la longueur totale du crâne (mm) pour un échantillon de 60 souris sylvestres adultes (I, II et III). La série statistique se présente comme suite :

22.28	23.18	23.47	23.72	24.09	24.56
22.56	23.23	23.48	23.48	24.13	24.63
22.57	23.29	23.48	23.48	24.32	24.83
22.60	23.30	23.49	23.49	24.35	24.94
22.69	23.34	23.51	23.51	24.36	24.95
22.73	23.35	23.56	23.56	24.37	25.00
22.78	23.35	23.57	23.57	24.41	25.07
22.91	23.37	23.60	23.60	24.43	25.16
23.05	23.39	23.61	23.61	24.43	25.48
23.14	23.47	23.71	23.71	24.52	25.74

TABLE 2.7 – Série statistique sur la longueur totale du crâne de souris sylvestres

Avec $n = 60$, calculons le nombre des classes en utilisant la règle de Sturges et l'intervalle de classes.

$$K = 1 + \left(3.3 \log_{10} n\right) = 1 + \left(3.3 \log_{10} 60\right) = 6.9 \simeq 7 \text{ classes}$$

et

$$I = \frac{\left(y_{(n)} - y_{(1)}\right)}{K} = \frac{25.74 - 22.28}{7} = 0.5$$

La première classe sera $[22.28, 22.28 + 0.5[= [22.28, 22.78[$ et ainsi de suite jusqu'à la 7^{ieme} classe.

2.5.2 Tableau des données

Classe	x_i	n_i	N_i	f_i	F_i
[22.28, 22.78[	22.53	6	6	0.10	0.10
[22.78, 23.28[	23.03	6	12	0.10	0.20
[23.28, 23.78[	23.53	28	40	0.4667	0.4667
[23.78, 24.28[	24.03	2	42	0.0333	0.70
[24.28, 24.78[	24.53	10	52	0.1667	0.8667
[24.78, 25.28[	25.03	6	58	0.10	0.9667
[25.28, 25.78[	25.53	2	60	0.0333	1
		$n = \sum_{i=1}^{4} n_i = 60$	1		

TABLE 2.8 – Tableau des données

2.5.3 Représentation graphique

Pour une variable quantitative continue, la répresentation graphique se fait à l'aide d'un *histogramme* pour les effectifs et les fréquences relatives des classes. L'histogramme permet de faire une bonne idée de la localisation, de la variabilité, de l'asymétrie et de l'aplatissement de la distribution des observations. Toutes ces mesures citées seront exploitées au chapitre suivant. L'axe associé aux effectifs doit toujours commencer à 0.

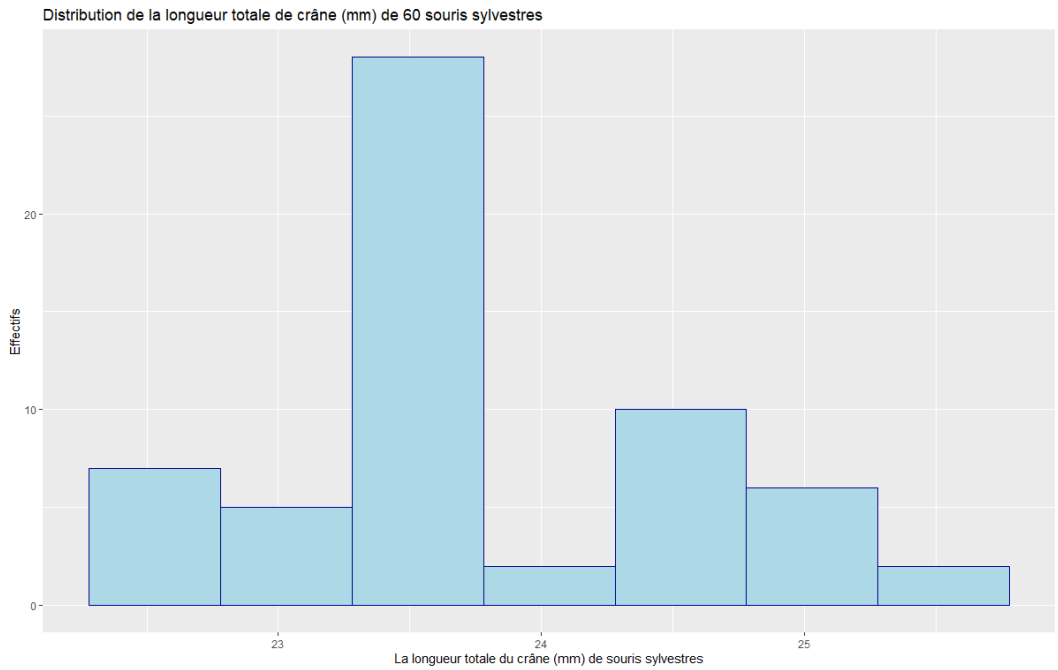

FIGURE 2.7 – Distribution de la longueur totale de crâne de 60 souris sylvestres

Il existe d'autres représentations graphiques que nous verrons dans les chapitres suivant telles que :

- le diagramme en boîte (boxplot ou diagramme de Tukey) ;
- le diagramme de dispersion.

Certains graphiques cités ci-dessus peuvent être utilisés aussi pour une variable quantitative discrète.

Exercice 2.1.

1. Après la déliberation du cours de statistique descriptive à la prémière session. Parmi les 256 étudiants ayant suivi le cours, 10 ont reçu la mention (Excellent), 50 (Très bien), 100 (Bien), 70 (Moins bien), 26 (Médiocre).

   (i) S'agit t-il de quel type de variable ?
   (ii) Représentez ces données à l'aide d'un diagramme en secteurs
   (iii) Représentez ces données à l'aide d'un diagramme en barres.

2. On a mesuré les diamètres de 30 eucalyptus. Voici les résultats obtenus, en centimètres.

64.03	54.05	53.96	50.09	61.28	70.04
67.18	57.23	51.14	63.12	69.03	55.94
52.02	65.08	68.13	56.04	57.93	69.08
70.37	53.07	69.11	61.03	70.00	63.97
58.01	54.09	55.06	52.92	65.19	69.96

   (i) S'agit t-il de quel type de variable ?
   (ii) Construire les classes pour la distribution des fréquences.
   (iii) Compléter le tableau statistique (centre des classes, effectifs cumulés, fréquences relatives, fréquences relatives cumulés).
   (iv) Représentez les classes pour la distribution des fréquences à l'aide de l'histogramme.

3. Dans un quatier de la commune de N'djili, on mène une enquête sur le nombre de pièces des habitations :

Nombre de pièces	1	2	3	4	5
Nombre d'appartements	99	150	125	70	50

   (i) S'agit t-il de quel type de variable ?
   (ii) Déterminer effectifs cumulés, fréquences relatives, fréquences relatives cumulés.
   (iii) Représentez ces données à l'aide d'un diagramme en bâtonnet.

2.6 Code R

```
#####################################################################


#: Permet de faire un commentaire

#### Lien de téléchargement R pour Windows et Mac #####

https://cran.r-project.org/bin/windows/base/

https://cran.r-project.org/bin/macosx/


#### Lien de téléchargement R Studio pour Windows et Mac #####

https://www.rstudio.com/products/rstudio/download/


##### Packages R pour les graphiques du support ####

mypackages<-c("psych","tidyverse","corrplot")

check_fct<-function(pkg){
  if(!require(pkg,character.only = TRUE)){
    install.packages(pkg,dependencies = TRUE)
    library(pkg,character.only = TRUE)
  }
}

check_mypackages<-lapply(mypackages,check_fct)


#####################################################
#####################################################

### Variable qualitative nominale ###

# Création du vecteur en chaine des caractères

Data1 <- c('D', 'C', 'M', 'M', 'C', 'A', 'D',
'C', 'D', 'M', 'A', 'M', 'D', 'C', 'A', 'M',
'D', 'A', 'M', 'C', 'A', 'M', 'A', 'M')

Data1 <- as.factor(Data1) # Création de la variable qualitative
```

```
# Rénommer les modalités

levels(Data1) <- c("Autres", "Célibataire", "Divorcé(e)", "Marié(e)")


Tab1 = table(Data1) # Création de la table
Tab = as.data.frame(Tab)

#Fréquence relative

Tab$Freq_relat <- round((Tab$Freq/sum(Tab$Freq)) *100,0)
colnames(Tab) <- c("Etat_civil", "Effectifs", "Freq_relative")


Objet = c(Tab1) # Création de l'objet table1


## Tableau statistique ##

# Création du Jeu de données

Data_fr1 <- data.frame(Effectif = Objet, Frequence = Objet/sum(Objet))

## Graphique Camembert ##

library(ggplot2)

couleurs <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF")

ggplot(Tab, aes(x = "", y = Effectifs, fill = Etat_civil)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  scale_fill_manual(values = couleurs) +
  labs(title = "Graphique sur l'état civil")+
    geom_text(aes(label = paste(Freq_relative,"%")),
              position = position_stack(vjust = 0.5))+
  theme_void()

## Graphique en Barres ##

ggplot(data = Tab, mapping = aes(x = Etat_civil,
y = Effectifs , fill =  Etat_civil )) +
  geom_col() +
  labs(title = "Graphique sur l'état civil",
  fill = "État civil")+
  scale_fill_manual(values = couleurs)+
```

```
  geom_text(aes(label = Effectifs), vjust = 1.6, color = "white")


### Variable qualitative ordinale ###

Niveau_sat <- c('Tr', 'Tr', 'Tr', 'Tr', 'Tr',
 'Tr', 'Tr', 'Tr', 'Tr', 'Tr','Tr', 'Tr',
 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Pl',
 'Pl','Pl', 'Pl', 'Pl', 'Pl', 'Pl', 'Pl',
  'Pl', 'Pl', 'Ni', 'Ni','Ni', 'Pl.ins',
'Pl.ins', 'Pl.ins', 'Pl.ins', 'Pl.ins',
 'Tr.ins', 'Tr.ins', 'Tr.ins' , 'Tr.ins')

# Création de la variable qualitative

Niveau_sat <- factor(Niveau_sat, levels = c('Tr','Pl','Ni','Pl.ins',
                                            'Tr.ins'), ordered = TRUE)


Tab2 <- table(Niveau_sat) # Création de la table
Tab12 <- as.data.frame(Tab2) # Création du jeu de données
colnames(Tab12) <- c("Niveau_sat", "Effectifs")
Tab12$Effect_Cum <- cumsum(Tab12$Effectifs)
Niveau_sat <- c('Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr',
                'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Tr', 'Pl', 'Pl',
                'Pl', 'Pl', 'Pl', 'Pl', 'Pl', 'Pl', 'Pl', 'Pl', 'Ni', 'Ni',
  'Ni', 'Pl.ins',  'Pl.ins', 'Pl.ins', 'Pl.ins', 'Pl.ins', 'Tr.ins',
  'Tr.ins', 'Tr.ins' , 'Tr.ins')

## Création de la variable qualitative ##

Niveau_sat <- factor(Niveau_sat, levels = c('Tr','Pl','Ni','Pl.ins',
                                            'Tr.ins'), ordered = TRUE)

# Création de la table

Tab2 <- table(Niveau_sat)

# Création du jeu de données

Tab12 <- as.data.frame(Tab2)
colnames(Tab12) <- c("Niveau_sat", "Effectifs")
Tab12$Effect_Cum <- cumsum(Tab12$Effectifs)
Tab12$Freq_relat <- round((Tab12$Effectifs/
sum(Tab12$Effectifs)) *100,0)
```

```
# Création de l'objet table2

Objet2 <- c(Tab2)

## Tableau statistique ##

# Création du Jeu de données

Data_fr2 <- data.frame(Effectif = Objet2, Eff_Cum = cumsum(Objet2),
                       Frequence = Objet2/sum(Objet2),
                       Freq_Cum = cumsum(Objet2/sum(Objet2)))

## Graphique Camembert ##

library(ggplot2)

couleurs1 <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF", "#00AFBB")

ggplot(Tab12, aes(x = "", y = Freq_relat, fill = Niveau_sat)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  scale_fill_manual(values = couleurs1) +
  labs(title = "Graphique sur le niveau de satisfaction du produit")+
  geom_text(aes(label = paste(Freq_relat,"%")),
            position = position_stack(vjust = 0.5))+
  theme_void()


Objet2 <- c(Tab2) # Création de l'objet table2

## Tableau statistique ##

# Création du Jeu de données

Data_fr2 <- data.frame(Effectif = Objet2, Eff_Cum = cumsum(Objet2),
                       Frequence = Objet2/sum(Objet2),
                       Freq_Cum = cumsum(Objet2/sum(Objet2)))



## Graphique Camembert ##
library(ggplot2)

couleurs1 <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF", "#00AFBB")

ggplot(Tab12, aes(x = "", y = Effectifs, fill = Niveau_sat)) +
```

```
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  scale_fill_manual(values = couleurs1) +
  labs(title = "Graphique sur le niveau de satisfaction du produit")+
  theme_void()

## Graphique en Barres ##

library(ggplot2)

ggplot(data = Tab12, mapping = aes(x = Niveau_sat,
 y = Effectifs , fill =  Niveau_sat )) +
  geom_col() +
  labs(title = "Graphique sur le niveau de satisfaction du produit",
       fill = "Niveau de satisfaction")+
  scale_fill_manual(values = couleurs1)+
  geom_text(aes(label = Effectifs), vjust = 1.6, color = "white")


ggplot(data = Tab12, mapping = aes(x = Niveau_sat,
  y = Effect_Cum , fill =  Niveau_sat )) +
  geom_col() +
  labs(title = "Graphique sur le niveau de satisfaction du produit cumulé",
       fill = "Niveau de satisfaction")+
  scale_fill_manual(values = couleurs1)+
  geom_text(aes(label = Effect_Cum), vjust = 1.6, color = "white")


### Variables quantitatives discrètes ###

Indiv <- c(1, 1, 1, 2, 2, 2, 2, 3, 3, 3,
           3, 3, 3, 4, 4, 4, 4, 4, 4, 4,
           5, 5, 5, 6, 6, 7, 7, 7,  8, 8)


Tab3 = table(Indiv ) # Création de la table
Tab13 = as.data.frame(Tab3) # Création du jeu de données

colnames(Tab13) <- c("Nbre_indiv", "Effectifs")
Tab13$Effect_Cum <- cumsum(Tab13$Effectifs)

Objet3 = c(Tab3) # Création de l'objet table2

## Tableau statistique ##

# Création du Jeu de données
```

```
Data_fr3 <- data.frame(Effectif = Objet3, Eff_Cum = cumsum(Objet3),
                       Frequence = Objet3/sum(Objet3),
                       Freq_Cum = cumsum(Objet3/sum(Objet3)))


## Graphique en batons ##

library(ggplot2)

  plot(Tab3,type="h",xlab="",ylab="",
       main="Graphique sur le nombre d'individus par ménage",frame=0,lwd=3)


## Histograme ##

  hist(Tab3)

### Variable continues ###


Poids <- c(22.28,23.18,23.47,23.72,24.09,24.56,
               22.56,23.23,23.48,23.48,24.13,24.63,
               22.57,23.29,23.48,23.48,24.32,24.83,
               22.60,23.30,23.49,23.49,24.35,24.94,
               22.69,23.34,23.51,23.51,24.36,24.95,
               22.73,23.35,23.56,23.56,24.37,25.00,
               22.78,23.35,23.57,23.57,24.41,25.07,
               22.91,23.37,23.60,23.60,24.43,25.16,
               23.05,23.39,23.61,23.61,24.43,25.48,
               23.14,23.47,23.71,23.71,24.52,25.74)

E = diff(range(Poids)) # y(n) - y(1)

K = 1+3.3* log(60, base = 10) # Règle de Sturges pour le nombre des classes

I = E/K # Intervelle des classes


## Construction de classe ##

# Couper les classes

Tab4 = table(cut(Poids,
                 breaks=c(22.2, 22.78, 23.28, 23.78, 24.28,
                  24.78, 25.28, 25.78)))
```

```
Tab5= c(Tab4)

## Tableau statistique ##

Data_fr4 <- data.frame(Eff=Tab5,EffCum=cumsum(Tab5),Freq=Tab5/sum(Tab5),
                       FreqCum=cumsum(Tab5/sum(Tab5)))

## Graphique histograme ##

hist(Poids, breaks = c(22.28, 22.78, 23.28, 23.78, 24.28, 24.78, 25.28, 25.78))

library(ggplot2)

df1 <- data.frame(Poids)

ggplot(df1, aes(x=Poids))+
  geom_histogram(color="darkblue", fill="lightblue",
  breaks = c(22.28, 22.78, 23.28, 23.78, 24.28,
  24.78, 25.28, 25.78))+
  labs(title = "Distribution de la longueur totale de crâne (mm)
   de 60 souris sylvestres",
x = "La longueur totale du crâne (mm) de souris sylvestres" ,
y = "Effectifs")
```

Chapitre 3

Statistique univariée

Dans ce chapitre, il est question de calculer les mesures statistiques d'une série statistique. Nous verrons les mesures *de localisation*, de *dispersion* et de *forme*.

3.1 Mesures de localisation (ou de la tendance centrale)

Nous considérons un échantillon de taille n avec $x_1, \ldots, x_n$ comme valeurs observées. Les mesures de localisation décrivent la position de l'échantillon sur l'axe des réels. Il s'agit de la *moyenne*, la *médiane* et la *mode*.

3.1.1 Moyenne arithmétique

Elle est définie comme la somme des valeurs observées divisé par la taille de l'échantion.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n} \tag{3.1}$$

Dans le cas où les valeurs observées sont distincts avec des effectifs, on parle de la *moyenne pondérée*.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + \ldots + n_k x_k}{n} \tag{3.2}$$

La moyenne pondérée est beaucoup utile dans le cas d'une variable quantitative continue avec regroupement des classes.

Exemple : Considérons les nombres d'employés de 10 entreprises sont les suivant : 10, 20, 35, 17, 12, 16, 10, 20, 18, 20.

Calculons la moyenne arithmétique :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{10+20+35+17+12+16+10+20+18+20}{10} = \frac{178}{10} = 17.8$$

On peut aussi calculer la moyenne pondérée en considérant les valeurs observées distinctes et leurs effectifs respectifs.

x_i	n_i
10	2
12	1
16	1
17	1
18	1
20	3
35	1

TABLE 3.1 – Tableau des données

Ainsi nous avons

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + \ldots + n_k x_k}{n} = \frac{10\times 2 + 12\times 1 + \ldots + 35\times 1}{10} = \frac{178}{10} = 17.8$$

La moyenne est sensible aux valeurs extrêmes, une seule valeur observée est suffisante pour que la moyenne soit grand (ou petit). Si on ajoute une valeur observée de 350 à la serie précédente, la moyenne sera :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{10+20+35+17+12+16+10+20+18+20+350}{11} = \frac{528}{11} = 48$$

On voit que la moyenne dépasse les 10 plus petites observations à cause de cette valeur extrême.

Il existe d'autres types de moyennes qui sont utilisées en physque, le cas de la *moyenne harmonique*, en mathématiques financières le cas de la *moyenne géométrique*.

3.1.2 Moyenne harmonique

La moyenne harmonique est définie pour tout $x_i \geq 0$:

$$\bar{h} = \frac{n}{\sum_{i=1}^{n} 1/x_i} \tag{3.3}$$

Exemple : Un chaffeur de taxi (Kinshasa - Matadi) parcoure le traject par la route nationale $N°1$ en 7 étapes de 346.34 km. Les vitessent pour ces étapes sont de 50 km/h, 45 km/h, 75km/h,
80 km/h, 65 km/h, 15km/h, 16.34 km/h. Quelle est la vitesse moyenne parcourus ?

$$\bar{h} = \frac{n}{\sum_{i=1}^{n} 1/x_i} = \frac{7}{1/50 + 1/45 + \ldots + 1/16.34} = \frac{7}{0.02 + 0.022 + \ldots + 0.061} = \frac{7}{0.207} \simeq 33.82\, km/h$$

3.1.3 Moyenne géométrique

La moyenne géométrique est définie pour tout $x_i \geq 0$:

$$\bar{g} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} \tag{3.4}$$

Exemple : Un étudiant souhaite financer ses études de master. Il place 100 000 Fc dans une banque de la place. On suppose que le taux d'intérêt à la banque pour les 3 ans est respectivement 3, 5 et 10%. Quelle est la somme que l'étudiant aura après 3 ans ?

En calculant la moyenne géométrique des taux on a :

$$\bar{g} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \left(\prod_{i=1}^{3} x_i\right)^{1/3} = (1.03 \times 1.05 \times 1.1)^{1/3} = (1.18965)^{1/3} = 1.0595946$$

Donc la somme que l'étudiant aura après 3 ans est :

$$100\ 000\ Fc \times \bar{g}^3 = 100\ 000 Fc \times (1.0595946)^3 = 100\ 000 Fc \times 1.18965 = 118\ 965\ Fc$$

Si on utilise les notions de mathétiques financières avec le calcul d'intérêt simple on aura après 3 ans :

$$100\ 000\ Fc \times 1.03 \times 1.05 \times 1.1 = 118\ 965\ Fc$$

3.1.4 Mode

Le mode est la valeur d'une variable la plus observée (ou fréquente) dans une série statistique. Lorsqu'il s'agit d'un regroupement en classes, une classe modale est une classe pour laquelle l'effectif associé est le plus grand.

Exemple : Soit une série statistique ayant les valeurs observées suivantes : 0, 1, 3, 0, 0, 4, 3, 4, 5, 4, 4, 4. Le tableau des données se présente comme suit :

x_i	n_i
0	3
1	1
3	2
4	5
5	1

TABLE 3.2 – Tableau des données

On a que 4 est une valeur modale.

Dans le cas d'une varible quantitative avec classe, la valeur modale est contenue dans la classe modale. Cette valeur est calculée par la formule suivante :

$$Mo = a_i^- + A_i\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \tag{3.5}$$

où $A_i = \left(a_i^+ - a_i^-\right)$ est l'amplitude de la classe modale i, a_i^- est la borne inférieure de la classe modale i, $\Delta_1 = n_i - n_{i-1}$ est la différence entre l'effectif de la classe modale et l'effectif de la classe précédente, $\Delta_2 = n_i - n_{i+1}$ est la différence entre l'effectif de la classe modale et l'effectif de la classe suivante.

Nous calculons la valeur modale de l'exemple concernant la mesure de la longueur de 60 souris sylvestres, à partir du tableau des données (2.8).

La classe modale est la classe [23.28, 23.78] ayant l'effectif n_i le plus élevé, avec $a_i^+ = 23.78$, $a_i^- = 23.28$, $A_i = 0.5$, $\Delta_1 = n_i - n_{i-1} = 28 - 6 = 22$ et $\Delta_2 = n_i - n_{i+1} = 28 - 2 = 26$. Nous pouvons donc trouver cette valeur.

$$Mo = a_i^- + A_i\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 23.28 + 0.5 \times \left(\frac{22}{22 + 28}\right) = 23.28 + 0.5 \times 0.44 = 23.28 + 0.22 = 23.5$$

3.1.5 Médiane

La médiane est la valeur centrale d'un échantillon. Cela revient à dire qu'il y a presque 50% des observations qui sont inférieures à cette valeur et 50% sont supérieures à la valeur .

De manière générale pour calculer la médiane d'une série statistique, il faudrait ordonner d'abord la série. La médiane est définie comme suit :

$$Me = x_{1/2} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ est impair} \\ \frac{x_{\left(\frac{n}{2}\right)} + \left(\frac{n}{2}+1\right)}{2} & \text{si } n \text{ est pair} \end{cases} \tag{3.6}$$

La médiane est une mesure de localisation *robuste*, parce qu'il est insensible aux valeurs extrêmes.

Exemple : La compagnie de construction Ferlon présente les salaires de ses 9 employés choisis aléatoirement (en milliers de Fc) : 100, 800, 500, 350, 400, 200, 650, 900, 830.

Premièrement, il faut ordonner la série en $x_{(1)} = 100$, $x_{(2)} = 200$, $x_{(3)} = 350$, $x_{(4)} = 400$, $x_{(5)} = 500$, $x_{(6)} = 650$, $x_{(7)} = 800$, $x_{(8)} = 830$ et $x_{(9)} = 900$. En suite avec $n = 9$ qui est impair, on peut calculer la médiane maintenant,

$$x_{1/2} = x_{(\frac{n+1}{2})} = x_{(\frac{9+1}{2})} = x_{(5)} = 500.$$

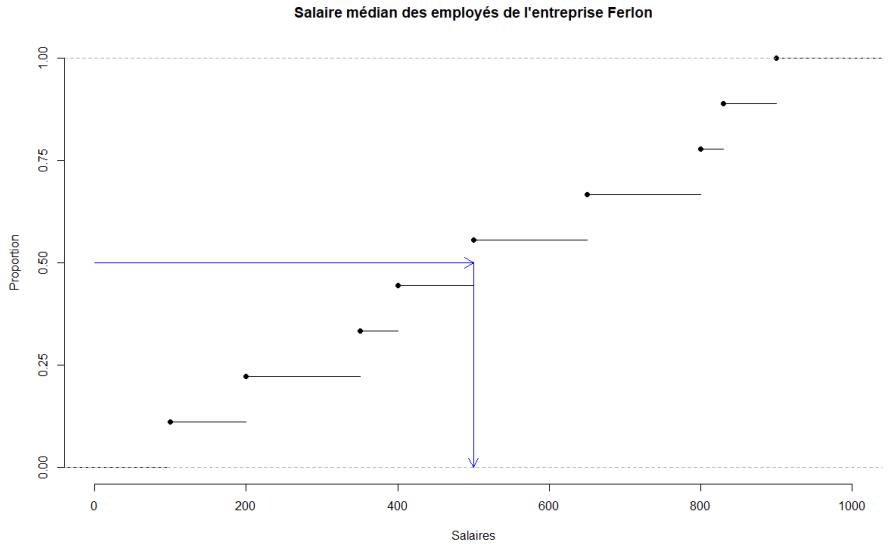

FIGURE 3.1 – Salaire médian des employés de l'entreprise Ferlon

La médiane est définie comme l'inverse de la fonction de répartition d'une série statistique dont la valeur vaut 1/2. L'inverse de la fonction de répartition est appelée une *quantile*, que nous parlérons dans la suite.

Dans le cas d'une varible quantitative avec classe, la valeur de la médiane est contenue dans la classe médiane. Cette valeur est calculée en utilisant une *interpolation linéaire* qui est une application directe du théorème de Thalès [1].

1. Thalès de Milet, est un philosophe et savant grec, né à Milet vers 625-620 av. J.-C. et mort vers 548-545 av. J.-C. On lui attribue de nombreux exploits, comme le calcul de la hauteur de la grande pyramide ou la prédiction d'une éclipse, ainsi que le théorème de Thalès. Il fut l'auteur de nombreuses recherches mathématiques, notamment en géométrie.

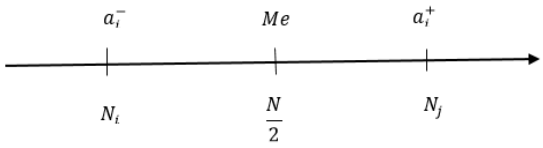

FIGURE 3.2 – Application du théorème de Thalès

$$\frac{\left(Me - a_i^+\right)}{\left(a_i^+ - a_i^-\right)} = \frac{(N/2 - N_i)}{\left(N_j - N_i\right)}$$

En faisant quelques manipulations, on a :

$$Me = a_i^- + A_i \frac{(N/2 - N_i)}{\left(N_j - N_i\right)} \tag{3.7}$$

où A_i est l'amplitude de la classe i, a_i^+ est la borne supérieure de la classe i, a_i^- est la borne inférieure de la classe i, N_i est l'effectif cumulé de la précédente et N_j est l'effectif cumulé de la classe modale.

Nous calculons la valeur médiane de l'exemple concernant la mesure de la longueur de 60 souris sylvestres, à partir du tableau des données (2.8).

La classe médiane est la classe $[23.28, 23.78]$, avec $a_i^+ = 23.78$, $a_i^- = 23.28$, $N_i = 12$, $N_j = 40$ et $N/2 = 30$. Nous pouvons donc trouver cette valeur.

$$Me = a_i^- + A_i \frac{(N/2 - N_i)}{\left(N_j - N_i\right)} = 23.28 + (23.78 - 23.28) \times \frac{(30 - 12)}{(40 - 12)} = 23.28 + 0.3214 = 23.6014 \simeq 23.6$$

3.2 Mesures de variabilité (ou de dispersion)

Nous considérons un échantillon de taille n avec $x_1, \ldots, x_n$ comme valeurs observées. Les mesures de variabilité permettent de quantifier la dispersion des valeurs d'un échantillon. Il s'agit de l' *Étendu*, la *variance*, l'*écart-type*, le *coefficient de variation*, *Écart interquartile*.

3.2.1 Étendu

C'est la différence entre la plus grande $x_{(n)}$ et la plus petite $x_{(1)}$ valeur observée.

$$E = x_{(n)} - x_{(1)} \tag{3.8}$$

Son utilité reside dans le calcul de l'intervalle de classe. Elle est très sensible aux valeurs extrêmes et à la taille d'échantillon. Car elle n'utilise que deux valeurs extrêmes $x_{(1)}$ et $x_{(n)}$.

3.2.2 Variance

C'est la somme des carrés des écarts à la moyenne divisée par $n-1$ observations.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.9}$$

La variance est une mesure positive ou nulle, elle s'annule lorsque tous les x_i sont égaux à $\bar{x}$. La variance mesure la dispersion des valeurs observées autour de la moyenne. La raison pour le dénominateur $n-1$ dans la formule (3.9) est un résultat du cours de *statistique inférentielle*, qui dépasse le cadre de ce cours. L'unité de mesure pour la variance s'exprime au carrée. Si x_i est en m, alors sa variance s^2 sera en m^2.

En pratique, la pluspart de logiciel statistique (R, SAS, SPSS, Stata,..) utilise cette *variance*. Cette variance permet d'estimer la variance d'une variable statistique à partir d'un échantillon. Dans le cas où les valeurs observées sont distincts avec des effectifs, on parle de la *variance pondérée*.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} n_i\,(x_i - \bar{x})^2 \tag{3.10}$$

Le théorème suivant permet de facilité le calcul de la variance.

Théorème 3.1. *La variance peut s'écrire sous la forme*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right] \tag{3.11}$$

Démonstration. En effet,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

En divisant par $n-1$ de deux côtés, on a le résultat. □

3.2.3 Écart-type

C'est la racine carrée de la varaince.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.12}$$

3.2.4 Coefficient de variation

C'est le rapport entre l'écart-type la moyenne.

$$C.V = \frac{s}{\bar{x}} \tag{3.13}$$

Il mesure la variabilité des observations par rapport à la moyenne de manière relative. Il compare la variabilité de deux variables possédant des moyennes différentes ou des unités différentes. Il est souvent exprimé en pourcentage.

Exemple : On mesure les poids de 45 rats de laboratoire en gramme et ensuite on mesure les poids de 55 lapins en gramme. On calcul la moyenne, l'écart-type et CV pour chaque population.

Population	n_i	$\bar{x}$	s	CV
Les rats	45	120	16	0.13
Les lapins	55	450	32	0.075

TABLE 3.3 – Tableau des données

On remarque que l'écart-type est petit chez les rats, tandisque le coefficient de variation est plus petit chez les lapins.

3.2.5 Les quantiles d'ordre α

Le quantile d'ordre α est un nombre q_α tel qu'une proportion α (environ) des x_i sont inférieurs à q_α et une proportion $(1-\alpha)$ des x_i sont supérieurs à q_α.

Il généralise la notion de la médiane. Dans l'exemple sur les salaires de 9 employés, le quantile d'ordre 0.3 est donc la valeur, notée $q_{0.3}$, telle que 30% du salaire de nos 9 salariés lui sont inférieurs et 70% lui sont supérieurs. Il existe plusieurs méthodes pour calculer les quantiles, nous présenterons deux méthodes.

Méthode 1

Étant donné notre échantillon de taille n avec $x_{(1)}, \ldots, x_n$ comme valeurs observées. On écrit $x_1,\ x_{(2)}, \ldots, x_{(n)}$pour dénoter nos observations placées en ordre croissant.

- Si $n\alpha$ est un nombre entier, on a

$$q_\alpha = \frac{1}{2}\left\{x_{(n\alpha)} + x_{(n\alpha+1)}\right\} \tag{3.14}$$

- Si $n\alpha$ n'est pas un nombre entier, on a

$$q_\alpha = x_{(\lceil n\alpha \rceil)} \tag{3.15}$$

où $\lceil n\alpha \rceil$ est le plus petit nombre entier supérieur ou égal à $n\alpha$.

Méthode 2

Étant donné notre échantillon de taille n avec $x_1, \ldots, x_n$ comme valeurs observées. On écrit $x_{(1)}$, $x_{(2)}, \ldots, x_{(n)}$pour dénoter nos observations placées en ordre croissant.

On calcul directement le quantile d'ordre α par

$$q_\alpha = x_{(1+(n-1)\alpha)} \tag{3.16}$$

Exemple : Nous allons utiliser la méthode 1 pour l'exemple sur les salaires de 9 employés. Avec $n = 9$, calculons le quantile d'order $\alpha = 0.3$. Il faut toujours ordonner les valeurs observées en ordre croissant.

on a $n\alpha = 9\times0.3 = 2.7$ qui n'est pas un entier. On utilise la formule de l'équation (3.15).

$$q_{0.3} = x_{(\lceil 2.7 \rceil)} = x_{(3)} = 350$$

3.2.6 Cas particuliers des quantiles

Nous définirons les quartiles, déciles et percentiles.

Les quartiles

Pour les quartiles, il y a le premier quartile $q_{0.25} \equiv Q_1$, la médiane $q_{0.50} \equiv Q_2$ et le troisième quartile $q_{0.75} \equiv Q_3$ qui sont présentés dans les formules (3.14) et (3.15). Il suffit de remplacer les valeurs de α et la taille de l'échantillon n.

Les déciles

Pour les déciles, il y a le premier décile $q_{1/10} \equiv Q_{1/10}$, le cinquième décile $q_{0.50} \equiv Q_2$ et le neuvième décile $q_{9/10} \equiv Q_{9/10}$ qui sont présentés dans les formules (3.14) et (3.15).

Les percentiles

Pour les percentiles, il y a le cinquième percentile $q_{0.05} \equiv Q_{5/100}$, le cinquantième percentile $q_{0.50} \equiv Q_2$ et le nonante-cinquième percentile $q_{0.95} \equiv Q_{95/100}$ qui sont présentés dans les formules (3.14) et (3.15).

3.2.7 Écart-interquantile

C'est la distance entre le premier quatile Q_1 et le troisième quartile Q_3.

$$EIQ = Q_3 - Q_1 \tag{3.17}$$

Il est une mesure de dispersion plus robuste que la variance, car son intervalle contient pratiquement 50% des données de l'échantillon (chaque coté de la médiane vaut 25%).

Le diagramme en boîte (ou diagramme de Tukey [2])

C'est une représentation graphique résumant sept nombres (E_1, I_1, Q_1, Q_2, Q_3, I_2 et E_2). Il permet de détecter les données abérrantes et valeurs extrêmes. Il est plus facile de voir la variablité et la symétrie des données, car il donne une idée sur la dispersion des données au centre (environ 50%).

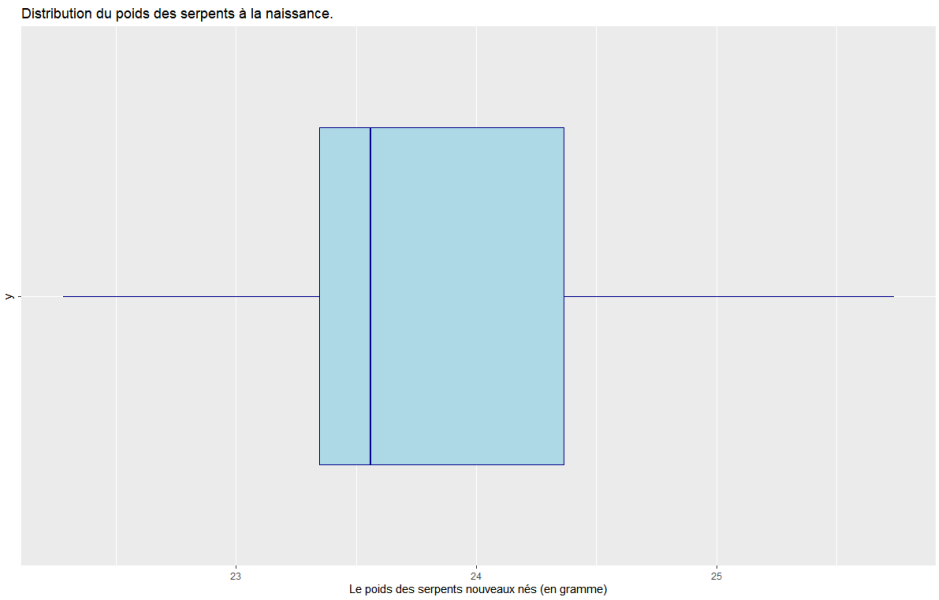

FIGURE 3.3 – Distribution du poids des serpents à la naissance avec n=60

Les I_1 et I_2 sont appelés *limites internes* de la boîte, tandis que les E_1 et E_2 sont appelées les *limites externes* de la boîte. La ligne qui se trouve à l'intérieur du

2. John Wilder Tukey (16 juin 1915 à New Bedford - 26 juillet 2000 à New Brunswick) est l'un des plus importants statisticiens américains du xxe siècle. Il créé et développe de nombreuses méthodes statistiques. Il est notamment connu pour son développement en 1965, avec James Cooley, de l'algorithme de la transformée de Fourier rapide.

rectangle est la médiane Q_2. Le bord du rectangle de gauche (respectivement de droite) est le quartile Q_1 (respectivement le quartile Q_3). Les limites internes et externes de la boîte sont calculées par les formules suivantes.

$$I_1 = Q_1 - 1.5 \times EIQ;\ I_2 = Q_3 + 1.5 \times EIQ \tag{3.18}$$

et

$$E_1 = Q_1 - 3 \times EIQ;\ I_2 = Q_3 + 3 \times EIQ \tag{3.19}$$

Il est plus facile de détecter une observation aberrante si l'on connait les limites internes I_1 et I_2. Une observation x_i est dite aberrante si $x_i > Q_3 + 1.5 \times EIQ$ ou $x_i < Q_1 - 1.5 \times EIQ$.

Le diagramme en boîte est la meilleure représentation lorsqu'il s'agit de comparer plusieurs échantillons.

Exemple : Trois vaccins (Pfizer, Moderna et Astra-Zaneca) ont été utilisés pour lutter contre la pandémie COVID-19 qui a ravagé le pays. Un échantillon pour chaque vaccin a été pris sur le nombre de vaccinés durant une période bien définie.

Pfizzer : 14, 56, 37, 93, 56, 37 , 90, 67, 87, 35, 89, 43
Moderna : 53, 67, 83, 25, 78, 66, 35, 97, 19, 26, 36, 25
Astra-Zeneca : 56, 27, 39, 68, 10, 67, 53, 36, 38,19, 25, 28.

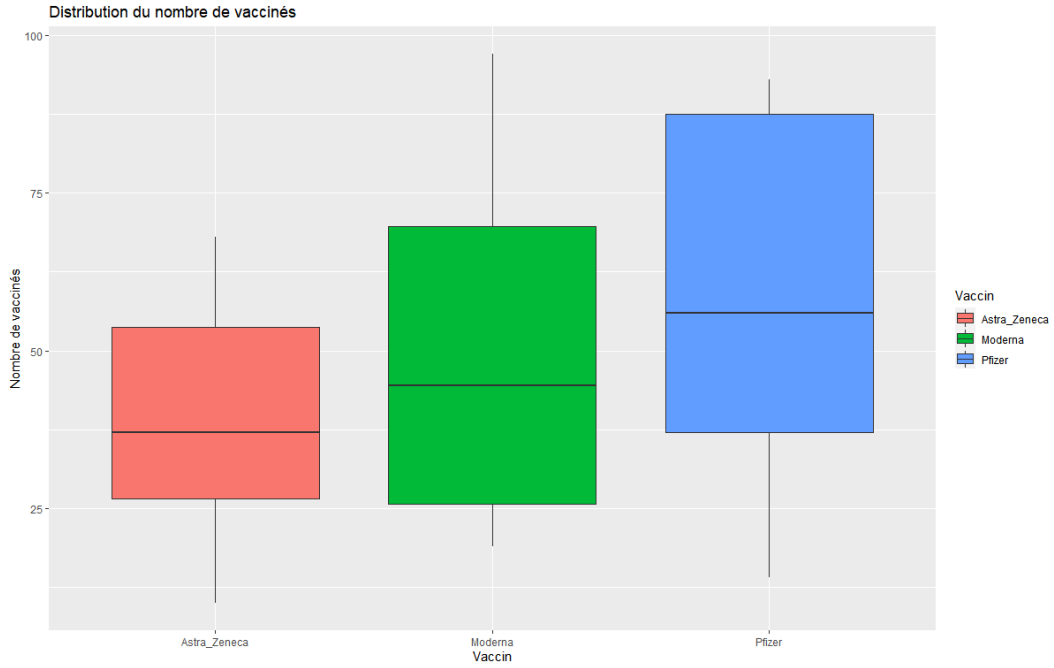

FIGURE 3.4 – Distribution du nombre de vaccinés avec n=12 par groupe

D'après le graphique, le nombre de vaccinés pour le vaccin Pfizer est plus grand en moyenne que les autres vaccins (Moderna et Astra-Zeneca).

3.3 Mesures de forme

Nous considérons un échantillon de taille n avec $x_1, \ldots, x_n$ comme valeurs observées. Les mesures de forme permettent de décrire la symétrie et le niveau d'aplatissement d'une distribution des valeurs observées d'un échantillon. Il s'agit du *coefficient d'asymétrie* et du *coefficient d'aplatissement*.

3.3.1 Coefficient d'asymétrie (ou skewness)

Le coefficient d'asymétrie permet de mesurer le niveau d'écartement de la symétrie de la distribution des observations d'un échantillon.

Le coefficient d'asymétrie β_1 est définir par

$$\beta_1 = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}, \; n > 2 \tag{3.20}$$

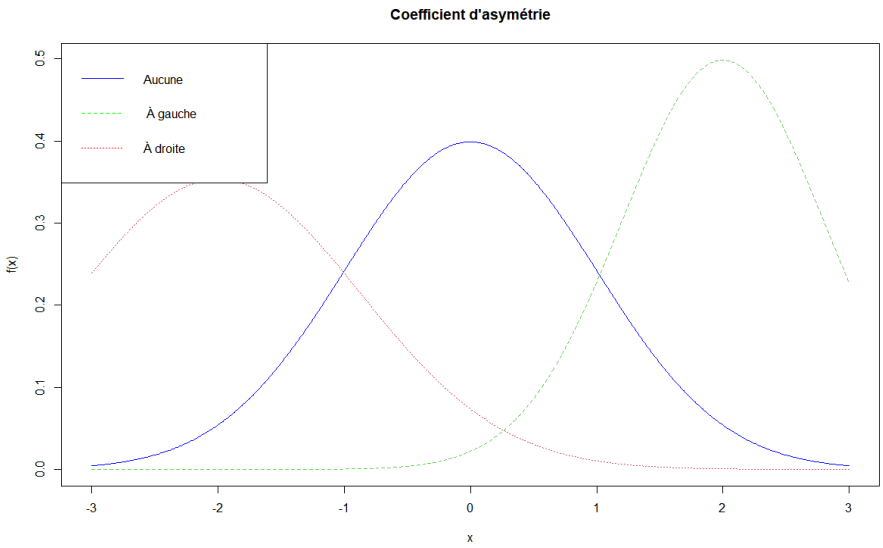

FIGURE 3.5 – Coefficient d'asymétrie

Suivant la formule (3.20), on dit qu'il y a :

- symétrie si $\beta_1 = 0$;
- asymétrie à droite si $\beta_1 > 0$;
- asymétrie à gauche si $\beta_1 < 0$.

Il existe dans la littérature d'autres types de coefficient d'asymétrie, nous présentons seulement le coefficient d'asymétrie de Yule qui est défini à partir de quartiles (Q_1, Q_2 et Q_3)

$$\beta_Y = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} \tag{3.21}$$

Suivant la formule (3.21), on dit qu'il y a :

- symétrie si $\beta_Y = 0$;
- asymétrie à droite si $\beta_Y > 0$;
- asymétrie à gauche si $\beta_Y < 0$.

3.3.2 Coefficient d'aplatissement (ou kurtosis)

Le coefficient d'aplatissement mesure l'importance de valeurs observées près du centre de symétrie ayant comme référence la loi normale.

Le coefficient d'aplatissement β_2 est définir par

$$\beta_2 = \frac{n\,(n+1)}{(n-1)\,(n-2)\,(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4},\ n > 3 \tag{3.22}$$

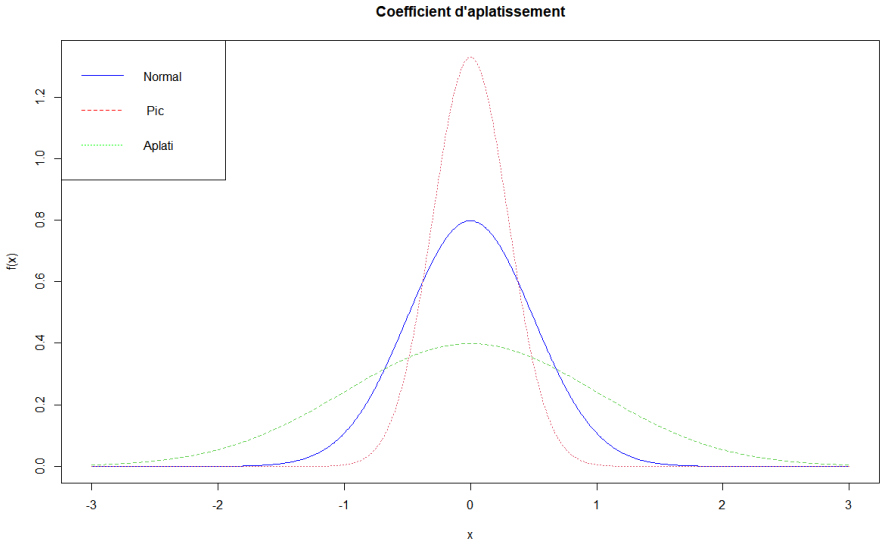

FIGURE 3.6 – Coefficient d'asymétrie

Suivant la formule (3.22), on dit qu'il y a :

- aplatissement normal (mesokurtique) si $\beta_2 = 0$;
- aplatissement leptokurtique (pic) si $\beta_2 > 0$;
- aplatissement platykurtique (aplati) si $\beta_2 < 0$.

Exercice 3.1.

1. On considère l'exemple sur le poids des serpents à la naisssance avec une taille d'échantillon $n = 70$.

33.90	34.87	34.49	34.16	35.14	34.10	34.41
33.10	35.59	35.20	34.35	33.87	35.18	34.54
35.71	34.74	36.42	34.68	34.05	34.76	35.49
35.44	36.03	34.19	35.49	35.30	34.26	35.36
35.83	33.64	34.83	36.11	35.32	37.37	34.78
36.66	33.91	33.55	34.91	34.17	34.96	34.71
34.31	35.78	34.86	34.19	35.50	32.62	33.20
35.52	34.59	35.32	36.21	35.61	36.14	34.31
34.55	37.61	35.86	33.75	34.77	34.37	33.68
35.70	35.65	35.67	34.20	34.11	35.18	35.07

TABLE 3.4 – Données brutes sur la mesure du poids des serpents à la naissance

(i) Déterminer les mesures de position (Moyenne, médiane et mode).
(ii) Déterminer les mesures de variabilité (Variance, écart-type, coefficient de variation et l'écart-interquartile).
(iii) Déssiner le diagramme en boîte en précisant les valeurs obtenus dans (i) et (ii).
(iv) Calculer le $40^{\text{ième}}$ centile.
(v) Calculer les mesures de forme (Skewness et kurtosis).

2. On considère les données de l'énoncé 1., repondez aux questions suivantes :

(i) Construire les classes pour la distribution des fréquences.
(ii) Représentez les classes pour la distribution des fréquences à l'aide de l'histogramme.
(iii) Déterminer les mesures de position (Moyenne, médiane et mode)
(iv) Déterminer les mesures de variabilité (Variance, écart-type, coefficient de variation et l'écart-interquartile).
(v) Calculer les mesures de forme (Skewness et kurtosis).

3. Après une enquête concernant les loyers annuels des habitations dans une commune de la ville de Kinshasa. On a le résultat suivant.

Montant(×100)	n_i	x_i	N_i	f_i	F_i
$[1, 3[$	50				
$[3, 6[$	150				
$[6, 9[$	100				
$[9, 12[$	25				
$[12, 15[$	5				

TABLE 3.5 – Tableau des données

(i) Compléter le tableau statistique (centre des classes, effectifs cumulés, fréquences relatives, fréquences relatives cumulés).

(ii) Déterminer les mesures de position (Moyenne, médiane et mode).
(iii) Déterminer les mesures de variabilité (Variance, écart-type, coefficient de variation et l'écart-interquartile)
(iii) Tracez l'histogramme et le diagramme en boîte de cette distribution.

3.4 Code R

```
### Mediane quand n est impair ###

x= c(100, 800, 500, 350, 400, 200, 650, 900, 830)
x1 = sort(x)

median(x1)
plot(ecdf(x),xlab="Salaires",ylab="Proportion",
     main="Salaire médian des employés de l'entreprise Ferlon",frame=FALSE,
     yaxt = "n")
axis(2, c(0.0,0.25,0.50,0.75,1.00))
arrows(0, 0.50, 500, 0.50, length=0.14,col="blue")
arrows(500,0.50,500,0,length=0.14,col="blue")


## Graphique Box-plot pour le poids ##


df1 <- data.frame(Poids)

ggplot(df1, aes(x=Poids, y = ""))+
  geom_boxplot(color="darkblue", fill="lightblue")+
  labs(title = "Distribution du poids des serpents à la naissance.  ",
       x = "Le poids des serpents nouveaux nés (en gramme)")

## Graphique Box-plot pour different vacci ##

Pfizer <- c(14, 56, 37, 93, 56, 37 , 90, 67, 87, 35, 89, 43)
Moderna <- c(53, 67, 83, 25, 78, 66, 35, 97, 19, 26, 36, 25)
Astra_Zeneca <- c(56, 27, 39, 68, 10, 67, 53, 36, 38,19, 25, 28)

# Création du Jeu de données

Vaccin2 <- data.frame(Pfizer, Moderna , Astra_Zeneca)

# Transformation de données large en longue

Vaccin1 <- pivot_longer(Vaccin2, cols = c('Pfizer', 'Moderna', 'Astra_Zeneca'),
```

```
                        names_to = 'Vaccin',
                        values_to = 'Nombre',
                        values_drop_na = TRUE)

head(Vaccin1) ## Visualisation
Vaccin1$Vaccin <- as.factor(Vaccin1$Vaccin)

boxplots <- ggplot(data = Vaccin1) +
  geom_boxplot(mapping = aes(x = Vaccin, y = Nombre, fill = Vaccin)) +
  labs(
    x = "Vaccin",
    y = "Nombre de vaccinés",
    title = "Distribution du nombre de vaccinés"
  )



## Graphique du coefficient d'asymétrie ##

x1 <- seq(-3,3,.01)
y <- dnorm(x1, 0, 1)
y2 <- dnorm(x1, 2, 0.8)
y3 <- dnorm(x1, -2, 2/sqrt(pi))
leg.txt <- c("Aucune"," À gauche","À droite")
plot(x1, y, xlab="x", ylab="f(x)", main="Coefficient d'asymétrie",
       ylim=range(y, y2, y3), col="blue", type="l")
lines(x1,y2, lty=2, col = 3)
lines(x1,y3, lty=3, col = 2)

legend("topleft",leg=leg.txt, col= c("blue", "green", "red"), lty=1:3)


## Graphique du coefficient d'applatissement ##

x1 <- seq(-3,3,.01)
y <- dnorm(x1, 0, 0.5)
y2 <- dnorm(x1, 0, 1)
y3 <- dnorm(x1, 0, 0.3)
leg.txt <- c("Normal"," Pic","Aplati")
plot(x1, y, xlab="x", ylab="f(x)", main="Coefficient d'aplatissement",
     ylim=range(y, y2, y3), col="blue", type="l")
lines(x1,y2, lty=2, col = 3)
lines(x1,y3, lty=3, col = 2)

legend("topleft",leg=leg.txt, col= c("blue", "red", "green"), lty=1:3)
```

Chapitre 4

Statistique bivariée

Dans ce chapitre, nous nous intéressons à s'informer s'il existe une relation (ou association) entre deux variables X et Y issue d'une population. Un échantillon aléatoire de taille n est obtenu pour les deux variables. Les observations sont notées sous la forme de n couples qu'on note : $(x_1, y_2),\ (x_2, y_2), \ldots, (x_n, y_n)$. Nous verrons comment décrire le lien lorsqu'il s'agit de *deux variables quantitatives*, *deux variables qualitatives* et *une variable qualitative et quantitative*.

4.1 Deux variables quantitatives

Pour caractériser le lien entre deux variables numériques, on utilise le coefficient de corrélation. Nous verrons deux types de coefficients corrélations dans cette section à savoir :

(i) le coefficient de corrélation de Pearson ;
(ii) le coefficient de corrélation de Spearman.

4.1.1 Représentation graphique

Une représentation graphique du couplet $(x_1, y_2),\ (x_2, y_2), \ldots, (x_n, y_n)$, appelée *nuage de point*(ou *diagramme de dispersion*) permet visualiser l'association entre les deux variables numériques.

Exemple : Une étude est faite dans la commune de Maluku pour évaluer le niveau de la mal nutrition dans la commune. On mesure la taille (X) et le poids (Y) de 33 individus.

x_i	y_i	x_i	y_i	x_i	y_i
150	55	165	63	180	75
171	69	166	61	157	62
158	60	169	68	170	65
169	71	152	66	165	63
160	63	157	59	155	58
175	72	167	58	151	56
165	70	166	64	165	61
162	63	168	66	175	71
171	61	153	64	156	60
173	72	157	59	179	74
160	61	167	63	169	72

TABLE 4.1 – Tableau des données

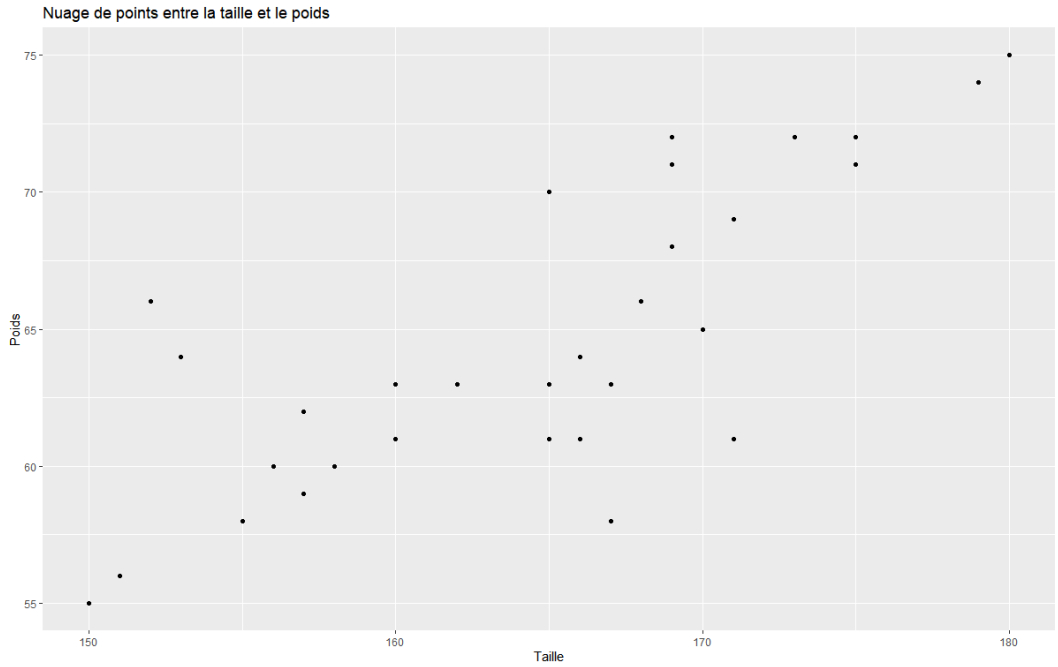

FIGURE 4.1 – Diagramme de dispersion

4.1.2 Coefficient de corrélation

Le coefficient de corrélation mesure le dégré d'association linéaire entre deux variables quantitatives.

Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson est défini comme suit :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{cov(x,y)}{(n-1)s_x s_y} \tag{4.1}$$

où x_i et y_i sont les valeurs observées des variables X et Y avec $i = 1, \ldots, n$ et $\bar{x}$, $\bar{y}$, s_x et s_y sont les moyennes et écartstypes des x_i et y_i respectivements.

Le coefficient de corrélation de Pearson r varie entre -1 et 1. Voici quelque cas qu'on rencotre :

- si $r = 1$, alors il y a une liaison linéaire parfaite entre les deux varaibles ;
- si $r > 75$, alors il y a une forte liaison linéaire parfaite entre les deux varaibles ;
- si $r > 0$, alors il y a une liaison linéaire positive entre les deux variables ;
- si $r = 0$, alors il n'y a pas une liaison linéaire entre les deux variables ;
- si $r < 0$, alors il y a une liaison linéaire négative entre les deux variables ;

En utilisant l'exemple précédent on a : $\sum_{i=1}^{n} x_i y_i = 350282$, $n = 33$, $\bar{x} = 164.33$, $\bar{y} = 64.393$,
$s_x = 8.002604$, $s_y = 5.407977$. On remplace les valeurs dans la formule (4.1).

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{350282 - 33 \times 164.33 \times 64.393}{32 \times 8.002604 \times 5.407977} = 0.7752707 \simeq 0.78 \tag{4.2}$$

Puisque $r = 0.78 > 0$, on conclut qu'il existe une forte liaison linéaire entre la taille et le poids. Une visualisation du résultat pourrait se faire aussi avec le graphique appelé *corrplot*.

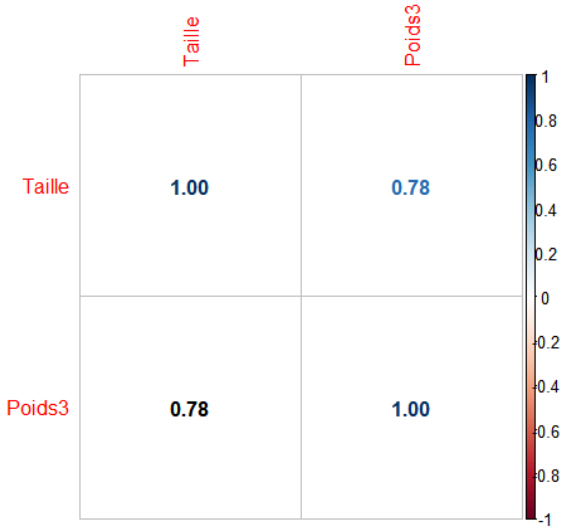

FIGURE 4.2 – Corrplot entre la taille et le poids

Coefficient de corrélation de Spearman

Le coefficient de corrélation de Spearman est beaucoup plus générale que celui de Pearson, mais utilise la formule de Pearson pour mesurer le degré d'association entre deux variables quantitatives en remplaçant les observations par leur *rang*. Dans le cas où observations ont la même valeur pour une même variable, on prendra le rang moyen de ces observations.

Soient $t_1, \ldots, t_n$ qui correspondent aux rangs des observations $x_1, \ldots, x_n$ et $s_1, \ldots, s_n$ les rangs des observations $y_1, \ldots, y_n$. Le coefficient de spearman est définit comme suit :

$$r_s = \frac{\sum_{i=1}^{n} (t_i - \bar{t})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n} (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^{n} (s_i - \bar{s})^2}} = \frac{\sum_{i=1}^{n} t_i s_i - n\bar{t}\bar{s}}{(n-1)\, s_t s_s} \tag{4.3}$$

Exemple : En utilisant l'exemple précédent avec $n = 8$ on a :

x_i	y_i	t_i	s_i	$t_i \times s_i$
150	55	1	1	1
171	69	7	5	35
158	60	2	2	4
169	71	6	7	42
160	63	3	3.5	10.5
175	72	8	8	64
165	70	5	6	30
162	63	4	3.5	14

TABLE 4.2 – Tableau des données

En remplaçant les valeurs dans la formule (4.3) on a :

$$r_s = \frac{\sum_{i=1}^{n} t_i s_i - n\bar{t}\bar{s}}{(n-1)\, s_t s_s} = \frac{200.5 - 8 \times 4.5 \times 4.5}{7 \times 2.44949 \times 2.434866} = 0.9221722 \simeq 0.92$$

4.1.3 Variable dépendant du temps

Il arrive qu'une variable quantitative dépende d'une autre variable qui est le temps (jours, semaines, mois, trimestres, années, etc..). Dans ce cas, on parle d'une *série temporelle* ou d'une *série chronologique*. L'analyse d'une série temporelle dépasse le cadre de ce cours, mais nous pourrons regarder une série chronologique de manière descriptive. Par exemple le PIB par habitant de la R.D.Congo de 1960 à 2020. On peut le représenter par un *un diagramme temporel*.

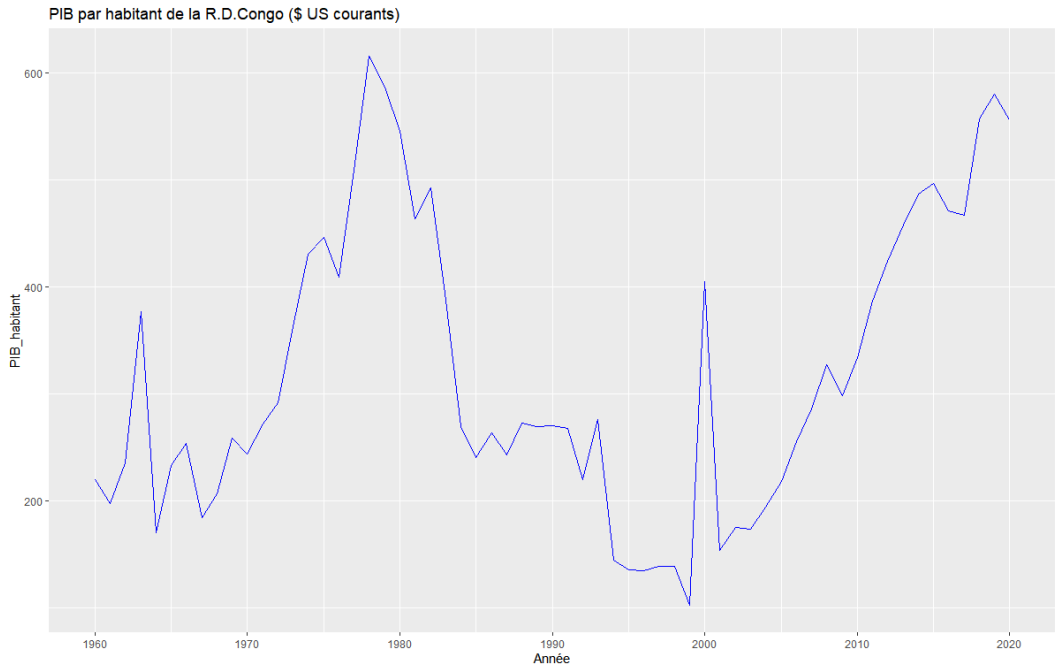

FIGURE 4.3 – Diagramme temporel du PIB par habitant pour la R.D.Congo (source : Banque mondiale)

4.2 Deux variables qualitatives

Lorsque deux variables X et Y sont qualitatives et possèdent les modalités $x_1, \ldots, x_i, \ldots x_I$ pour la variable X et $y_1, \ldots, y_j, \ldots y_J$ pour la variable Y. Les modalités des variables X et Y sont présentées sous la forme d'un tableau qu'on appelle le *tableau de fréquences* (ou *tableau de contingence*) par les *fréquences croisées* comme suit.

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_J	Total
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

TABLE 4.3 – Tableau de fréquences

où

- n_{ij} avec $i = 1, \ldots, I$ et $j = 1, \ldots, J$ est la fréquence (ou *fréquence croisée*) qui représente le nombre d'observations pour lesquelles la variable X appartient à la modalité x_i et la variable Y à la modalité y_j simultanément.

- $n = \sum_{i=1}^{I} \sum_{i=j}^{J} n_{ij}$ est le nombre total d'observations dans l'échantillon.

Exemple : Considérons l'exemple du nombre de vaccinés selon le sexe. On a constitué un echantillon de 1400 personnes vaccinées avec les 3 vaccins(Pfizer, Moderna et Astra-Zeneca).

Vaccin / *Sexe*	Astra-Zeneca	Moderna	Pfizer
Femme	$n_{11} = 69$	$n_{12} = 278$	$n_{13} = 355$
Homme	$n_{21} = 46$	$n_{22} = 232$	$n_{23} = 420$

On remarque dans l'exemple que, le nombre des hommes dans l'échantillon ayant reçu le vaccin Pfizer (n_{23}) vaut 420.

La représentation graphique pour deux variables qualitatives se fait avec le *diagramme en bâtons empilés* ou le *diagramme en bâtons groupés*.

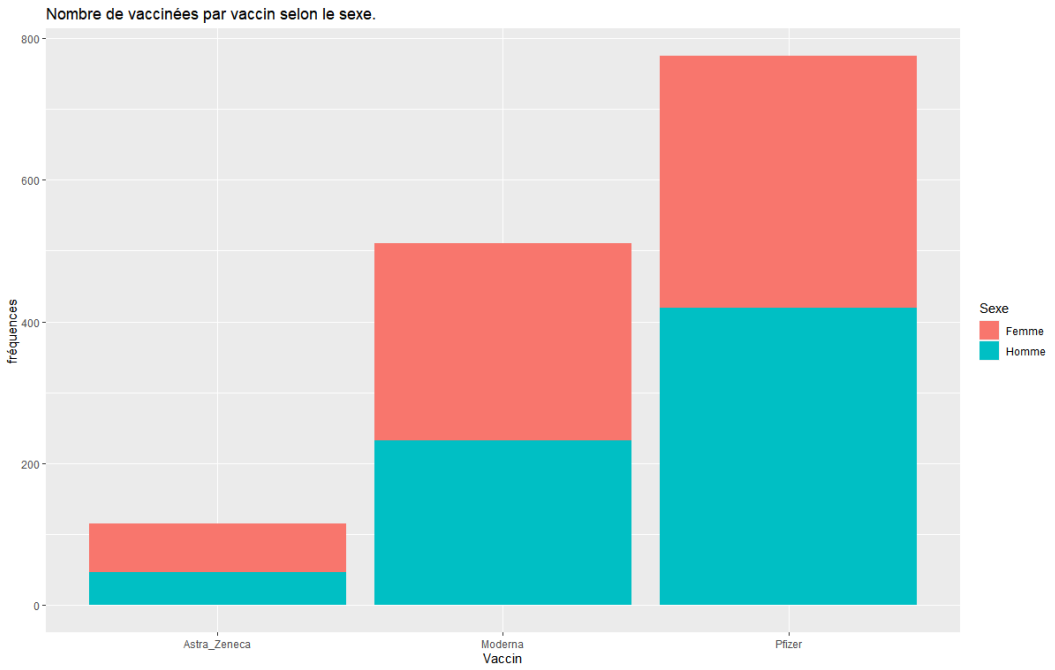

FIGURE 4.4 – Diagramme en bâtons empilés

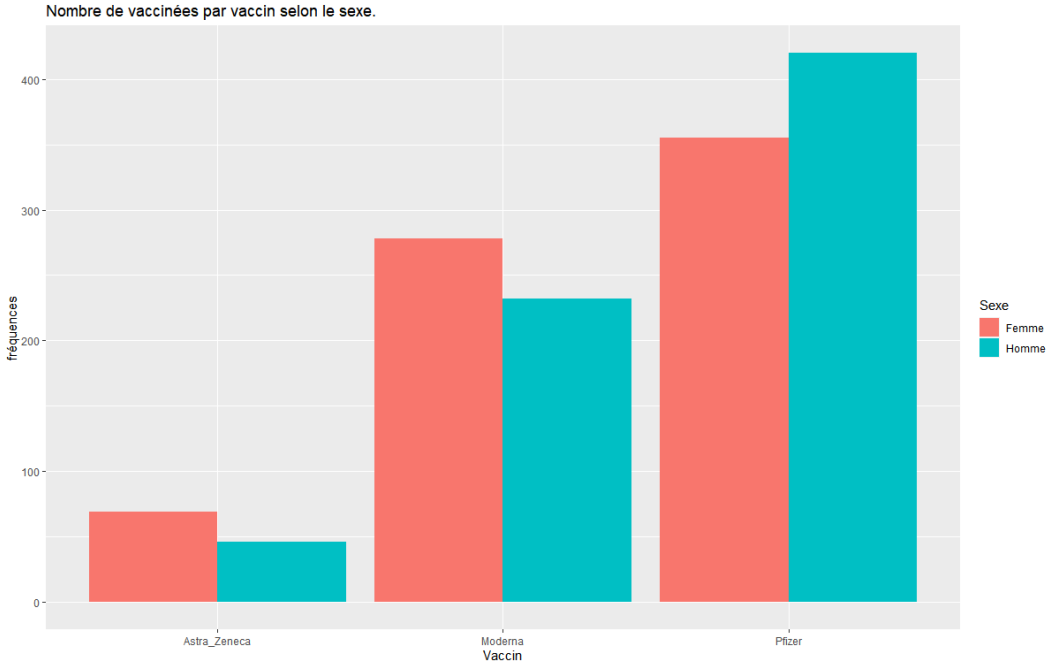

FIGURE 4.5 – Diagramme en bâtons groupés

Il existe plusieurs types des fréquences que nous pouvons dégager à partir du tableau de fréquences.

4.2.1 Fréquences marginales

Les fréquences marginales se retrouvent plus précisement sur la ligne et la colonne du total.

$$n_{i\bullet} = \sum_{i=1}^{I} n_{ij} \; et \; n_{\bullet j} = \sum_{j=1}^{J} n_{ij}$$

Exemple : En utilisant l'exemple précédent, nous allons calculer les fréquences marginales.

Vaccin / *Sexe*	Astra-Zeneca	Moderna	Pfizer	Total
Femme	$n_{11} = 69$	$n_{12} = 278$	$n_{13} = 355$	702
Homme	$n_{21} = 46$	$n_{22} = 232$	$n_{23} = 420$	698
Total	115	510	775	1400

Les fréquences marginales du sexe sont ($n_{1\bullet} = 702$; $n_{2\bullet} = 698$) et les fréquences marginales des vaccins sont ($n_{\bullet 1} = 115$; $n_{\bullet 2} = 510$; $n_{\bullet 3} = 775$). Ainsi l'échantillon comprend 702 femmes et 698 hommes. Parmi ces individus, 115 ont reçu le vaccin Astra-Zeneca, 510 ont reçu le vaccin Moderna et 775 ont reçu le vaccin Pfizer.

4.2.2 Fréquences relatives

- Les fréquences relatives croisées

$$f_{ij} = \frac{n_{ij}}{n} \text{ avec } i = 1, \ldots, I \text{ et } j = 1, \ldots, J$$

- Les fréquences relatives marginales pour les lignes

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} \text{ avec } i = 1, \ldots, I$$

- Les fréquences relatives marginales pour les colonnes

$$f_{\bullet j} = \frac{n_{\bullet j}}{n} \text{ avec } j = 1, \ldots, J$$

Exemple : En utilisant le même exemple sur les vaccins, nous allons calculer les fréquences relatives.

Vaccin / *Sexe*	Astra-Zeneca	Moderna	Pfizer	Total
Femme	$f_{11} = 0.0493$	$f_{12}0.1986$	$f_{13} = 0.2536$	$f_{1\bullet} = 0.5014$
Homme	$f_{21} = 0.0329$	$f_{22} = 0.1657$	$f_{23} = 0.3$	$f_{2\bullet} = 0.4986$
Total	$f_{\bullet 1} = 0.0821$	$f_{\bullet 2} = 0.3643$	$f_{\bullet 3} = 0.5536$	1

On remaque que, la somme de toutes les fréquences relatives croisées vaut 1 et la somme des fréquences relatives marginales vaut aussi 1.

4.2.3 Fréquences conditionnelles et fréquences relatives conditionnelles

On utilise la fréquence conditionnelle lorsqu'on cherche à trouver la fréquence de la variable X tout en fixant la variable Y à une certaine modalité. Cela revient à dire que chaque ligne ou colonne du table du tableau de fréquences est une classe de fréquences conditionnelles.

Par exemple, les fréquences des modalités des vaccins en conditionnant par rapport au fait d'être une femme sont ($n_{11} = 69$; $n_{12} = 278$; $n_{13} = 225$).

Fréquences relatives conditionnelles

- Les fréquences relatives conditionnelles de X par rapport à Y sont définies par

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}} \text{ avec } i = 1, \ldots, I \text{ pour } j \text{ fixé}$$

- Les fréquences relatives conditionnelles de Y par rapport à X sont définies par

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}} \text{ avec } j = 1, \ldots, J \text{ pour } i \text{ fixé}$$

Exemple : En utilisant le même exemple sur les vaccins, nous allons calculer les fréquences relatives conditionnelles.

Commençons par les fréquences relatives conditionnelles de X (Sexe) par rapport à Y (Vaccin), on a :

Sexe \ *Vaccin*	Astra-Zeneca	Moderna	Pfizer
Femme	$f_{1\|i=1} = \frac{69}{702} = 0.0983$	$f_{1\|i=2} = \frac{278}{702} = 0.3960$	$f_{1\|i=3} = \frac{355}{702} = 0.5057$
Homme	$f_{1\|i=2} = \frac{46}{698} = 0.0659$	$f_{2\|i=2} = \frac{232}{698} = 0.3324$	$f_{3\|i=2} = \frac{420}{698} = 0.6017$

D'après le tableau, il ressort qu'environ 50% des femmes ont reçu le vaccin Pfizer, 40% des femmes ont reçu le vaccin Moderna et 10% des femmes ont reçu le vaccin Astra-Zeneca.

Pour les fréquences relatives conditionnelles de Y (Vaccin) par rapport à X (Sexe), on a :

Sexe \ *Vaccin*	Astra-Zeneca	Moderna	Pfizer
Femme	$f_{1\|j=1} = \frac{69}{115} = 0.6$	$f_{2\|j=2} = \frac{278}{510} = 0.5451$	$f_{3\|i=3} = \frac{355}{775} = 0.4581$
Homme	$f_{1\|j=1} = \frac{46}{115} = 0.4$	$f_{2\|j=2} = \frac{232}{510} = 0.4549$	$f_{3\|j=3} = \frac{420}{775} = 0.5419$

D'après le tableau, il ressort qu'environ 60% des individus ayant reçu le vaccin Astra-Zeneca sont des femmes et 40% des hommes, 55% des individus ayant reçu le vaccin Moderna sont des femmes et 45% des hommes, enfin 46% des individus ayant reçu le vaccin Pfizer sont des femmes et 54% des hommes.

Pour approfondir la connaissance sur l'analyse des données catégorielles, le livre d'Agresti [2] vous serait très utile.

4.3 Une variable qualitative et quantitative

Pour décrire le lien entre une variable qualitative et une variable quantitative. On utilise les statistiques descriptives de la variable quantitative en fonction de la variable qualitative.

La représentation graphique pour ce type des variables se fait à l'aide d'un *diagramme en boîte* ou un histogramme.

Exemple : On s'intéresse à mesurer la taille des enfants issus d'une famille dont les parents sont des grandes tailles selon le sexe. Un échantillon de taille n=16 est choisi pour l'étude.

Taille	*Sexe*	*Taille*	*Sexe*
170	F	171	M
191	M	187	M
188	F	190	F
179	F	176	F
170	M	196	M
175	M	180	F
195	F	175	M
182	M	184	M

TABLE 4.4 – Tableau des données

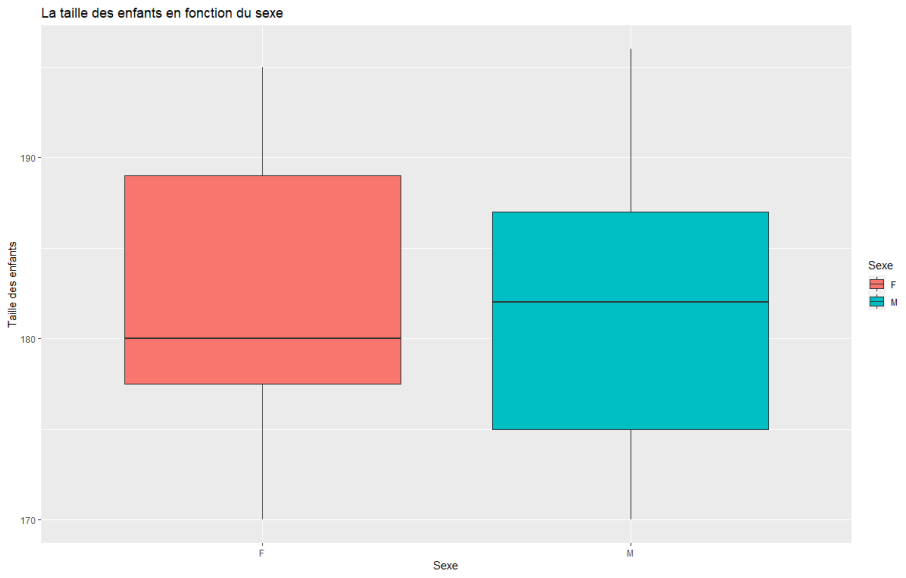

FIGURE 4.6 – Diagramme en boîte de la taille selon le sexe de

Nous pourrons en suite calculer les statistisques de la taille en fonction du sexe.

Sexe	$\bar{x}$	s	$C.V$
Fminin	182.57	8.75	0.0479
Masculin	181.22	9.11	0.0503

TABLE 4.5 – Tableau des données

On observe en moyenne que, la taille des enfants de sexe féminin est plus grande que celui de sexe masculin.

Exercice 4.1.

1. On considère la série statistique double suivante :

x_i	50	68	65	78	66	87	75	79	56	72
y_i	182	145	156	154	167	186	156	145	177	160
x_i	60	61	85	67	76	77	69	79	53	82
y_i	172	165	146	174	157	176	166	148	187	170

TABLE 4.6 – Tableau des données

   (i) Représenter graphiquement la série statistique double.
   (ii) Calculer la $cov(x, y)$.
   (iii) Calculer le coefficient de corrélation de Pearson et Spearman. Interpréter le résultat obtenu.

2. On souhaite avoir les intentions de vote à éléction présidentielle selon le sexe. On constitue un échantillon de 25000 répondants. Les fréquences croisées se présentes comme suit :

	Démocrate	Écologiste	Indépandant
Femme	$n_{11} = 5402$	$n_{12} = 3680$	$n_{13} = 1279$
Homme	$n_{21} = 9569$	$n_{22} = 3500$	$n_{23} = 1570$

TABLE 4.7 – Tableau des données

   (i) Représenter graphiquement le tableau des données.
   (ii) Calculer les fréquences marginales $n_{i\bullet}$ et $n_{\bullet j}$.
   (iii) Calculer les fréquences relatives.
   (iv) Calculer les fréquences relatives conditionnelles du sexe par rapport aux partis politiques et calculer les fréquences relatives conditionnelles des partis politiques par rapport au sexe.

3. On s'intéresse à mesurer le poids des personnes ayant contracté la tuberculose (Oui) et le poids des personnes saines(Non). Un échantillon de taille $n = 30$ est choisi pour l'étude.

Poids	*Malade*	*Poids*	*Malade*	*Poids*	*Malade*
50	Oui	71	Non	71	Non
61	Non	87	Non	53	Oui
88	Non	60	Oui	73	Non
59	Oui	76	Non	72	Non
70	Non	56	Oui	47	Oui
45	Oui	80	Non	54	Non
75	Non	75	Non	69	Non
52	Oui	48	Oui	85	Non
55	Oui	75	Non	51	Oui
72	Non	49	Oui	50	Oui

TABLE 4.8 – Tableau des données

(i) Représenter graphiquement la série statistique double.
(ii) Calculer les mesures de position (Moyenne et médiane) par rapport au malade.
(iii) Calculer les mesures de dispersion (Ecart-type, coefficient de variation) par rapport au malade.

4.4 Code R

```
## Graphique nuage des points ##

Taille <- c(150, 171, 158, 169, 160, 175, 165, 162, 171, 173, 160,
            165, 166, 169, 152, 157, 167, 166, 168, 153, 157, 167,
            180, 157, 170, 165, 155, 151, 165, 175, 156, 179, 169)

Poids3 <- c(55, 69, 60, 71, 63, 72, 70, 63, 61, 72, 61,
            63, 61, 68, 66, 59, 58, 64, 66, 64, 59, 63,
            75, 62, 65, 63, 58, 56, 61, 71, 60, 74, 72)

n <- 33

Etude <- data.frame(Taille, Poids3)


Nuage <- ggplot(data = Etude) +
  geom_point(mapping = aes(x = Taille, y = Poids3)) +
  labs(
    x = "Taille",
    y = "Poids",
    title = "Nuage de points entre la taille et le poids"
  )
```

```
### Calcul du coefficient de Pearson ###

Produit <- Taille * Poids3

## Formule directe

R_pearson = (sum(Produit) - n *mean(Taille) * mean(Poids3))/

((n-1)*sd(Taille)*sd(Poids3))

cor(Taille, Poids3) ## Fonction du package stat

## Graphique Correlation de Pearson ##

library(corrplot)

correlation2 <- cor(Etude)

corr2 <- corrplot(correlation2, method = "number")
corrplot(corr2,add=T,type="lower", method="number",
         col="black", diag=FALSE,tl.pos="n", cl.pos="n")

### Calcul du Coefficient de spearman ###

Taille2 <- c(150, 171, 158, 169, 160, 175, 165, 162)
Poids4 <- c(55, 69, 60, 71, 63, 72, 70, 63)

cor(Taille2, Poids4, method = "spearman") ## Fonction du package stat

t_i <- rank(Taille2)
s_i <- rank(Poids4)

n<-8
produit1 <- t_i * s_i

## Formule directe ##

R_spearman = (sum(produit1) - n *mean(t_i) * mean(s_i))/((n-1)*sd(t_i)*sd(s_i))

## Graphique PIB de la R.D.Congo ##

# Lecture du jeu des données PIB (Source : Banque mondiale)

df <- read_xlsx("PIB_RDC.xlsx")

View(df) ## Visualisation
```

```
# Renommer la variable PIB par habitant ($ US courants)

names(df)[names(df) == 'PIB par habitant ($ US courants)'] <- 'PIB_habitant'

## Lecture du jeu des données en time serie ##


Base <- ts(df$PIB_habitant, start = 1960 , frequency = 1)

### Tracé du graphique ###

autoplot(Base, main="PIB par habitant de la R.D.Congo ($ US courants) ",
         xlab="Année", ylab = "PIB_habitant", col = "blue")


# Vaccins selon le sexe

Vaccin <- c(rep("Astra_Zeneca", times = 69), rep("Moderna", times = 278),
            rep("Pfizer", times = 355),
            rep("Astra_Zeneca", times = 46), rep("Moderna", times = 232),
            rep("Pfizer", times = 420))
Sexe <- c(rep("Femme",times=702),rep("Homme",times=698))

Vaccin <- factor(Vaccin,levels=c("Astra_Zeneca","Moderna","Pfizer"))
Sexe <- factor(Sexe,levels=c("Femme","Homme"))
Tab_vac <- table(Sexe,Vaccin)

Freq_vac <- data.frame(Vaccin, Sexe)

## Graphique avec le diagramme en bâton empilés ##

Freq_diagram1 <- ggplot(data = Freq_vac) +
  geom_bar(mapping = aes(x = Vaccin, fill = Sexe)) +  # aesthetic fill ajoutée
  labs(
    x = "Vaccin",
    y = "fréquences",
    fill = "Sexe",
    title = "Nombre de vaccinées par vaccin selon le sexe."
  )


## Graphique avec le diagramme en bâton groupés ##

Freq_diagram2 <- ggplot(data = Freq_vac) +
  geom_bar(mapping = aes(x = Vaccin, fill = Sexe),
```

```
          position =  "dodge") +  # aesthetic fill ajoutée
  labs(
    x = "Vaccin",
    y = "fréquences",
    fill = "Sexe",
    title = "Nombre de vaccinées par vaccin selon le sexe."
  )


## Graphique variable qualit vs quantit ##

Taille <- c(170, 191, 188, 179, 170, 175, 195, 182,
            171, 187, 190, 176, 196, 180, 175, 184)

Sexe <- c("F", "M", "F", "F", "M", "M", "F", "M",
          "M", "M", "F", "F", "M", "F", "M", "M")

Sexe <- factor(Sexe, levels = c("F", "M"))

df1 <- data.frame(Taille, Sexe)


ggplot(data = df1) +
  geom_boxplot(mapping = aes(x = Sexe, y = Taille,
                             fill = Sexe)) +  # aesthetic fill ajoutée
  labs(
    x = "Sexe",
    y = "Taille des enfants",
    fill = "Sexe",
    title = "La taille des enfants en fonction du sexe ")


### Statistique de la taille en fonction du sexe

library(psych)

df2_Femme <- df1[df1$Sexe == "F",] ## Taille selon le sexe F

df2_Homme <- df1[df1$Sexe == "M",]  ## Taille selon le sexe M

describe(df2_Femme$Taille) ## Statistiques descriptives de
la taille selon le sexe F

describe(df2_Homme$Taille) ## Statistiques descriptives de
la taille selon le sexe M
```

Chapitre 5

Introduction au modèle probabiliste (Régression linéaire simple)

Dans ce chapitre nous présentons une technique permettant de modéliser une relation linéaire entre une variable explicative (ou exogène) X et une variable à expliquer (endogène)Y. Il s'agit du modèle de *régression linéaire simple*.

Il est évident que pour avoir le résultat pertinent (Intervalle de cofiance, Test d'hypothèse, sélection et validation du modèle) du modèle, il faudrait faire des cours de *statistique inférentielle* [1, 6, 8] et *théorie de régresssion* [3, 8]. Raison pour laquelle dans ce chapitre nous nous limitérons à la présentation, à l'estimation et à la prévision du modèle.

5.1 Introduction

Le mot *régression* a été utilisé pour la première fois par l'anglais Sir Francis Galton cousin de Charles Darwin. Travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères en 1885.

En 1886, il publia l'article *"Regression towards mediocrity in hereditary stature"*. Dans son article, Sir Galton montre que, lorsque le père était plus grand que la moyenne, son fils avait tendance à être plus petit que lui et, a contrario, que lorsque le père était plus petit que la moyenne, son fils avait tendance à être plus grand que lui.

Le but d'une analyse de régression est de :

- ajuster un modèle pour expliquer une variable endogène Y en fonction d'une variable exogène X.
- prédire les valeurs de la variable endogène Y pour de nouvelles observations de la variable exogène.

5.2 Modèle statistique

Le modèle de régression linéaire simple est un modèle qui s'écrit comme suit :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ avec } i = 1, \ldots, n \tag{5.1}$$

où

- $x_1, \ldots, x_n$ sont les valeurs observées de la variable exogène connues et constantes ;
- $Y_1, \ldots, Y_n$ sont les valeurs observées de la variable endogène qui est aléatoire ;
- β_0 et β_1 sont les paramètres inconnues du modèle ;
- $\varepsilon_1, \ldots, \varepsilon_n$ sont les réalisations inconnues d'une variable aléatoire.

Le paramètre β_0 est l'ordonnée à l'origine de la droite et s'interprète comme la valeur moyenne de Y lorsque x vaut zéro . Tandis que le paramètre β_1 est la pente de la droite et s'interprète comme l'accroissement moyen de Y lorsque x augmente d'une unité.

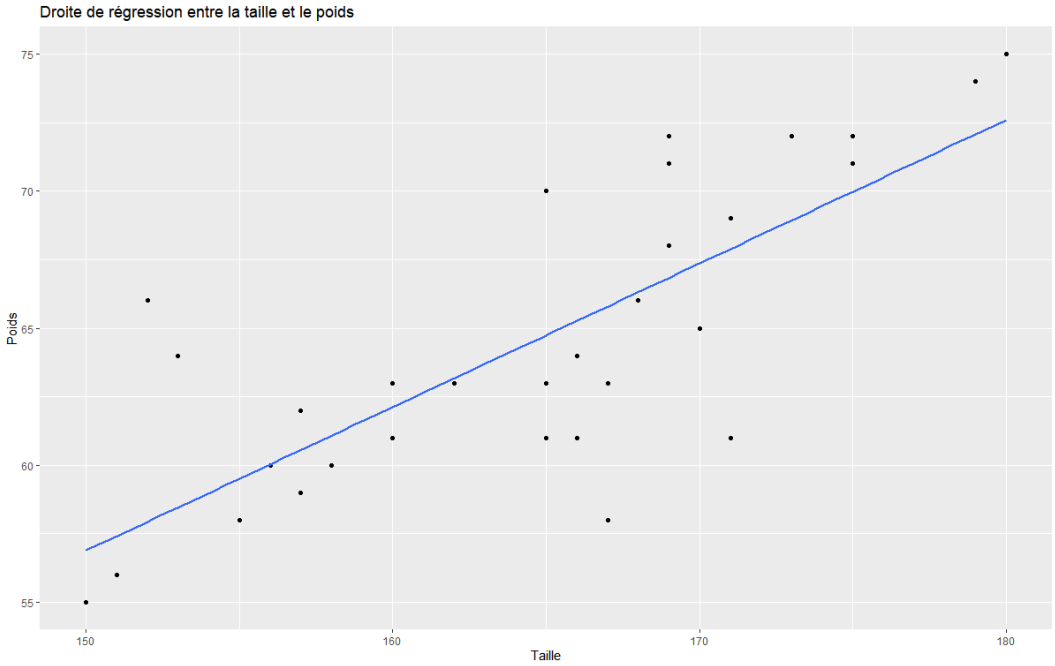

FIGURE 5.1 – Droite de régression

Le graphique de la figure (5.1) est tiré dans l'exemple sur le niveau de la mal nutrition au chapitre 4. La ligne en bleu désigne la valeur moyenne de la variable endogène Y en fonction de la valeur de x. Autour de cette droite sont distribués de façon aléatoire les observations de Y. Les $\varepsilon_i = Y_i - \hat{Y}_i$ avec $i = 1, \ldots, n$ sont les termes d'erreurs qui repésentent les différences entre les observations de Y et la droite (les valeurs ajustées ou les points sur la droite).

Ainsi les paramètres β_0 et β_1 permettent de décrire cette droite de régression. Comme ils sont inconnus, il faudrait les estimer.

5.2.1 Estimation

L'estimations des paramètres du modèle se fera par la méthode des moindres carrés. On doit cette méthode au mathématicien Adrien-Marie Legendre [1].

La méthode des moindres carrés consiste à choisir les paramètres β_0 et $\beta_1$de façon à minimiser la somme des carrés des erreurs. On cherche à minmiser la fonction suivante :

$$m(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_i)\}^2 \tag{5.2}$$

La fonction $m(\beta_0, \beta_1)$ vérifie certaines propriétés mathématiques (lisse et convexe), cela revient à dire qu'on peut trouver le minimum en dérivant la fonction $m(\beta_0, \beta_1)$ par rapport à β_0 et β_1 puis en posant ces dérivées partielles égales à zéro. On obtient un système d'équations suivantes :

$$\left\{ \begin{array}{rcl} \frac{\partial m(\beta_0,\beta_1)}{\partial \beta_0} & = & \frac{\partial}{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \\ \frac{\partial m(\beta_0,\beta_1)}{\partial \beta_1} & = & \frac{\partial}{\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \end{array} \right. \tag{5.3}$$

En dérivant par rapport à β_0, on a :

$$\begin{array}{rcl} \frac{\partial}{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 & = & -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \\ & = & -2 \left(n\bar{Y} - n\beta_0 - n\beta_1 \bar{x}\right) \end{array}$$

Trouvons maintenant le point critique de la fonction $m(\beta_0, \beta_1)$ en égalisant l'équation précédente à zéro.

$$\begin{array}{rcl} -2 \left(\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x}\right) & = & 0 \\ \bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} & = & 0 \end{array}$$

D'où

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{5.4}$$

On dérive maintenant par rapport à β_0 :

$$\begin{array}{rcl} \frac{\partial}{\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 & = & -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \\ & = & -2 \left(\sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 \right) \end{array}$$

1. Adrien-Marie Legendre, né en 1752 à Paris et mort en 1833 à Paris, est un mathématicien français.

Trouvons une fois deplus le point critique de la fonction $m(\beta_0, \beta_1)$ en égalisant l'équation précédente à zéro et l'équation (5.4).

$$\begin{aligned}
-2\left(\sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2\right) &= 0 \\
\sum_{i=1}^{n} Y_i x_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} Y_i x_i &= \left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} Y_i x_i &= \bar{Y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i\right)
\end{aligned}$$

D'où

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{Y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \tag{5.5}$$

Il est possible de réécrire l'équattion (5.5) avec certaine manipulation aux équations équivalentes suivantes :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{Y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{cov(x, Y)}{s_x^2} \tag{5.6}$$

Enfin, les estimateurs des moindres carrés pour le modèle de régression linéaire simple sont :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \text{ et } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{Y}\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5.7}$$

Si l'on veut mesurer le dégré d'association entre deux variables. Il suffit de calculer le *coefficient de corrélation de Pearson* à la formule (4.3) du chapitre 4.

5.2.2 Prévision

Soit x_{n+1} une nouvelle observation de la variable exogène X, nous cherchons à prédire la valeur y_{n+1} de la variable endogène Y. Il suffit d'utiliser le modèle estimé de la régression :

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \tag{5.8}$$

Exemple : En utilisant l'exemple sur le niveau de la mal nutrition du chapitre 4, on a : $n = 33$, $\sum_{i=1}^{n} x_i Y_i = 350282$, $\bar{Y} = 64.39394$, $\bar{x} = 164.3333$ et $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 2049.333$.

Nous pouvons calculer $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{Y}\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{350282 - 33 \times 64.39394 \times 164.3333}{2049.333} = 0.5239$$

et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 64.39394 - 0.5239 \times 164.3333 = -21.7020$$

D'où l'équation de la droite estimée est :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = -21.7020 + 0.5239 x_i, \text{ avec } i = 1, \ldots, 33.$$

Pour mesurer le dégré d'association entre la taille et le poids, nous l'avons calculé au chapitre 4. Nous avons trouvé que $r = 0.78 > 0$. On avait conclut qu'il existait une forte liaison linéaire entre la taille et le poids.

Suppons qu'il y a une nouvelle observation de la variable taille $x = 175$, alors nous pouvons prédire la valeur ajustée du poids en utilisant l'équation de la droite de régresion estimée.

$$\hat{Y}_i = -21.7020 + 0.5239 \times 170 = 67.361 \simeq 67.4\ kg$$

Exercice 5.1.

1. On mesure la longueur (X) (cm) et le poids (Y) des poissons (g) de type capitaine pechés dans le fleuve congo. On prend un échantillon de 20 capitaines dont les mesures sont :

Longueur	*Poids*	*Longueur*	*Poids*
31.5	614	36.8	560
34.8	606	33.4	645
31.1	596	36.4	700
37.5	743	29.5	620
30.4	653	31.4	690
32.6	542	30	598
36.0	680	33.8	632
27.6	694	38.1	650

TABLE 5.1 – Tableau des données

   (i) Représenter graphiquement les données à l'aide du diagramme de dispersion.
   (ii) Calculer les moyennes $\bar{x}$,$\bar{Y}$ et les écarts-types s_x, s_Y.
   (iii) Calculer la $cov(x, Y)$.
   (iv) Déterminer l'équation de la droite de régression et interpréter les paramètres estimés.
   (v) Calculer le coefficient de corrélation et interpréter le résultat.

(vi) Si une nouvelle observation de la longueur du poisson vaut 36.5, quel sera son poids estimé ?

2. À partir de la formule (5.6), montrer que :

(i)
$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

(ii)
$$\sum_{i=1}^{n} (x_i - \bar{x})\,(Y_i - \bar{Y}) = \sum_{i=1}^{n} (x_i - \bar{x})\,Y_i$$

(iii)
$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i$$

3. Considérons la droite de régresion $Y = \beta_0 + \beta_1 x_i$, avec $i = 1, \ldots, n$. Montrer qu'il existe une relation entre $\bar{x}$ et $\bar{Y}$, et entre s_x et s_y.

4. Soit la quantité $\sum (x_i - \alpha)^2$. Pour quelle valeur de α, la quantité précédente est minimisée.

5. On cherche à forcer le modéle de régression linéaire à passer par l'origine. On souhaite ajuster le modèle suivant :

$$Y_i = 2\beta x_i, \text{ avec } i = 1, \ldots, n$$

Trouver l'estimateur des moindres carrés du paramètre β correspondant au modèle.

5.3 Code R

```
### Droite de regression ###

# Création du heu de données Taille vs Poids

df3 <- data.frame(Taille2, Poids4)


# Ajouter la droite de regression dans le diagramme de dispersion

ggplot(Etude, aes(x=Taille, y=Poids3)) +
  geom_point()+
  geom_smooth(method=lm, se = FALSE)+
  #geom_jitter(colour = "blue", alpha = 0.3) + ## Ajout du bruit blanc
  labs(x ="Taille",
       y = "Poids",
       title = "Droite de régression entre la taille et le poids")
```

```
### Calcul de \beta_{0} et \bet_{1} ###

produit2 <- Taille * Poids3
n <- 33
ecart <- (Taille - mean(Taille)) **2

beta_1 <- (sum(produit2) - n* mean(Poids3) * mean(Taille))/sum(ecart)

beta_0 <- mean(Poids3) - beta_1*mean(Taille)


lm(Poids3 ~ Taille, Etude) ## Fonction du package stats
```

Bibliographie

[1] Luc Adjengue. *Méthodes statistiques : concepts, applications et exercices*. Presses internationales Polytechnique, 2014.

[2] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[3] Pierre-André Cornillon, Nicolas Hengartner, Eric Matzner-Løber, and Laurent Rouvière. *Régression avec R-2e édition*. EDP sciences, 2019.

[4] Pierre Lafaye De Micheaux, Rémy Drouilhet, and Benoît Liquet. *Le logiciel R*. Springer, 2011.

[5] Yadolah Dodge. Premiers pas en statistique. 2006.

[6] William W Hines, Douglas C Montgomery, and David M Goldman Connie M Borror. *Probability and statistics in engineering*. John Wiley & Sons, 2008.

[7] Sharon L Lohr. *Sampling : design and analysis*. Chapman and Hall/CRC, 2019.

[8] Gilbert Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

[9] Yves Tillé. *Théorie des sondages*, volume 13. 2001.