L3 Mathématiques, Informatique Appliquées aux Sciences Humaines et Sociales
Parcours Informatique - SHS

UE 503C - Statistique

2017-2018

Pascal Sarda

Présentation générale

Objectifs

Ce cours aborde des notions élémentaires permettant d'organiser, de représenter, de décrire et de synthétiser un ensemble de données recueillies dans le cadre d'une étude statistique.

Ces techniques s'appuient sur des outils issus de la statistique descriptive univariée (représentations et indicateurs usuels), bivariée (description de la relation entre deux variables, mesures d'association) et multivariées.

L'objectif est double : familiariser l'auditoire avec ces différents outils statistiques et l'initier à en faire un usage pertinent.

Bibliographie

Bressoud, E., Kahané, J.-C. *Statistique descriptive - Applications avec Excel et calculatrices* (2e édition), Pearson Education.

Bertrand, F., Maumy-Bertrand, M. *Initiation à la statistique avec R*, Dunod.

Saporta, G. *Probabilités, analyse des données et statistique (2nd édition)*, éditions Technip.

Responsable de l'UE

Pascal Sarda
Université Toulouse Jean Jaurès
Département Mathématiques-Informatique
Bât. O. de Gouges, porte GS283
Tél : 05 61 50 46 06
sarda@univ-tlse2.fr

Contrôle des connaissances (régime examen)

Épreuve sur table d'une durée de deux heures. Document autorisé : une feuille A4 recto-verso avec le résumé du cours rédigé par l'étudiant.

Table des matières

Chapitre 1

Introduction à la Statistique

1.1 Le concept de statistique

On peut trouver dans la littérature diverses définitions de la **statistique**. Par exemple, selon celle du Petit Robert, la statistique est *l'ensemble des techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité.* [1]

De manière générale, la statistique est un ensemble de méthodes (techniques) permettant de traiter ou d'analyser des ensembles d'observations ou de données afin de répondre à diverses problématiques (connaissance d'un phénomène, prise de décision,...).

Une étude statistique comporte différentes étapes faisant intervenir différents acteurs. Schématiquement, ces étapes sont les suivantes :

- définition des objectifs de l'étude
- choix des informations à recueillir (population, variables)
- éventuellement rédaction d'un questionnaire
- recueil et saisie des données (traitement informatique)
- traitement des données et analyse des résultats.

Dans ce cours, nous présentons des techniques relevant essentiellement de la statistique descriptive. Le vocabulaire de base de la statistique est donné dans la suite de ce chapitre. Au chapitre 2, sont décrites les principales méthodes relevant de la statistique descriptive unidimensionnelle (étude d'un caractère unique sur la population). Le chapitre 3 est consacré à la statistique descriptive bidimensionnelle (deux caractères) avec le but d'étudier la liaison entre les deux caractères en donnant un sens à cette notion.

1.2 Vocabulaire de la statistique

Population Ω : ensemble sur lequel porte l'étude statistique.

Individu ou **Unité statistique** $\omega \in \Omega$: tout élément de la population.

Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon N : nombre d'individus appartenant à l'échantillon.

Variable statistique ou **caractère** X : application définie sur Ω et à valeurs dans un ensemble F : à toute unité statistique ω (individu) on fait correspondre une valeur $X(\omega)$ de F.

1. Le Larousse propose une définition différente (plus obscure...) : *la statistique est la science des grands nombres regroupant l'ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions*

$$\begin{array}{rccc} X : & \Omega & \longrightarrow & F \\ & \omega & \longmapsto & X(\omega). \end{array}$$

Si la variable est à valeurs dans $\mathbb{R}$ (F est un sous-ensemble de $\mathbb{R}$), la variable est dite **quantitative** (par exemple, l'âge, la taille, le nombre d'enfants, ...). Dans le cas contraire, la variable statistique est dite **qualitative** : **qualitative ordinale** lorsque les valeurs de la variables (ou **modalités**) peuvent être ordonnées (consommation d'un produit : nulle, faible, moyenne, forte, ...) et **qualitative nominale** sinon (sexe, marque de la voiture).
Dans le cas d'une variable qualitative, il est fréquent de remplacer les modalités de la variable par des **codes** numériques facilitant le traitement (informatique) de la variable (masculin : 1, féminin : 2).

Recueil des données : il s'effectue lors d'une **enquête**. Si celle-ci est **exhaustive**, c'est-à-dire porte sur tous les sujets de la population, il s'agit dans ce cas d'un **recensement**. Dans le cas où l'enquête porte sur un échantillon, on effectue dans ce cas un **sondage**.
A l'issu de l'enquête, on dispose d'un premier **tableau statistique** se présentant généralement sous la forme de N lignes sur lesquelles sont portées les individus puis les valeurs de la ou des variables statistiques pour chaque individu (une variable par colonne).

Traitement des données : étape au cours de laquelle interviennent les techniques statistiques à proprement parler. On distingue deux branches principales en statistique :

a. **La statistique descriptive** : celle-ci a pour but la description des données et vise principalement deux objectifs : d'une part, la **représentation graphique** des données en alliant à la fois la simplicité (la "lisibilité") de la représentation et la fidélité aux données (en déformant au minimum la réalité) ; d'autre part, le résumé des données par des **caractéristiques numériques**. Les méthodes mathématiques utilisées sont relativement élémentaires.
b. **La statistique inférentielle (ou mathématique ou inductive)** : celle-ci a pour objectif de généraliser (inférer) à une population (ou un ensemble plus large) les résultats observés sur un échantillon (ensemble restreint). On suppose ainsi que le phénomène étudié peut être décrit par un **modèle mathématique** (donc théorique) permettant d'approcher les propriétés de ce phénomène. Le choix de ce modèle est bien sûr un problème important puisqu'il doit représenter au mieux la réalité. Les méthodes utilisées en statistique mathématique font appel au **calcul des probabilités** [2].

On voit dans ces définitions, que les objectifs de la statistique descriptive et de la statistique inférentielle sont assez différents. Pour autant, ceux-ci ne s'opposent pas et le plus souvent une étude descriptive précède l'utilisation de techniques de statistique inférentielle.

2. Le calcul des probabilités est une théorie mathématique dont le but est d'étudier les lois régissant des phénomènes ou des expériences aléatoires. Un phénomène (ou une expérience) est aléatoire lorsqu'il n'est pas possible de prévoir de manière certaine le résultat de ce phénomène (ou de cette expérience).

Chapitre 2

Statistique descriptive unidimensionnelle

Dans tout ce chapitre, on suppose qu'on observe une variable X sur N individus constituant la population ou l'échantillon suivant le cas.

2.1 Variables quantitatives

2.1.1 Variables quantitatives discrètes

On appelle **variable quantitative discrète** une variable quantitative dont l'ensemble des valeurs distinctes (possibles) est un ensemble fini ou dénombrable. Le plus souvent, ces valeurs sont des entiers (rarement des valeurs décimales) et le nombre de valeurs distinctes est assez faible, par exemple le *nombre d'enfants*. C'est ce dernier cas que nous considérons dans la suite : ainsi, nous notons $x_1, x_2, \ldots, x_k$ les valeurs distinctes de la variable, k étant le nombre de valeurs distinctes avec $k < N$. Enfin, les valeurs seront toujours ordonnées suivant l'ordre croissant : $x_1 < x_2 < \ldots < x_k$.

Organisation des données - Tableau statistique

Le tableau statistique est constitué d'une première colonne dans laquelle sont portées les valeurs distinctes de la variable rangées par ordre croissant. Dans les colonnes suivantes sont portés les effectifs, effectifs cumulés, fréquences et fréquences cumulées correspondant aux valeurs de la variable X. Soit x_j une de ces valeurs, $j = 1, \ldots, k$ une quelconque de ces valeurs.

Effectifs et fréquences : on appelle **effectif** de la valeur x_j le nombre d'individus observés ayant pris cette valeur de la variable. On note n_j cet effectif.
La **fréquence** de la valeur x_j est la proportion d'individus ayant pris cette valeur. Il s'agit donc de l'effectif divisé par la taille de l'échantillon, multiplié par 100 si elle est exprimée sous forme de pourcentage. On note f_j cette fréquence et on a :

$$f_j \;=\; \frac{n_j}{N} \;=\; \frac{n_j \times 100}{N}\%.$$

Effectifs et fréquences cumulés : On appelle **effectif cumulé** de la valeur x_j de la variable X le nombre d'individus ayant pris cette valeur ou une valeur inférieure . On note N_j cet effectif cumulé.
La **fréquence cumulée** de la valeur x_j est la proportion d'individus ayant pris cette valeur ou une valeur inférieure. Il s'agit donc de l'effectif cumulé divisé par la taille de l'échantillon. On note F_j cette fréquence cumulée et on a :

$$F_j \;=\; \frac{N_j}{N} \;=\; \frac{N_j \times 100}{N}\%.$$

Les valeurs de la variable X ayant été rangées par ordre croissant, l'effectif cumulé de la valeur x_j s'obtient en sommant les effectifs $n_1, \ldots, n_j$:

$$N_j \;=\; n_1 + \ldots + n_j \;=\; \sum_{l=1}^{j} n_l.$$

De même pour les fréquences cumulées :

$$F_j \;=\; f_1 + \ldots + f_j \;=\; \sum_{l=1}^{j} f_l.$$

L'effectif cumulé N_k (effectif cumulé de la plus grande valeur x_j de X) est égal au nombre d'individus N et la fréquence cumulée F_k est égale à 1 (ou 100%) :

$$N_k \;=\; N, \quad F_k \;=\; 1 \;=\; 100\%.$$

Exemple 2.1. *(Source : Bressoud, E. et Kahané, J.-C. Statistique descriptive. Applications avec Excel et calculatrices, PEARSON). La liste suivante est composée du nombre de films vus au cours du mois dernier par chaque étudiant issu d'un groupe de taille $N = 20$:*

3, 2, 2, 3, 1, 2, 0, 1, 2, 2, 0, 3, 0, 3, 2, 3, 3, 2, 1, 1

Tableau statistique :

x_j	n_j	N_j	f_j	F_j
0	*3*	*3*	*0,15*	*0,15*
1	*4*	*7*	*0,2*	*0,35*
2	*7*	*14*	*0,35*	*0,7*
3	*6*	*20*	*0,3*	*1*

```
> films1<-c(3,2,2,3,1,2,0,1,2,2,0,3,0,3,2,3,3,2,1,1)
> films1

 [1] 3 2 2 3 1 2 0 1 2 2 0 3 0 3 2 3 3 2 1 1

> table(films1)

films1
0 1 2 3
3 4 7 6
```

Représentations graphiques

Le tableau statistique complet comprend 5 colonnes : valeurs x_j de la variable, effectifs n_j, effectifs cumulés N_j, fréquences f_j, fréquences cumulées F_j (en général dans cet ordre). Il comprend k lignes (nombre de valeurs de la variable). Les représentations graphiques ont pour but de représenter par un graphique une des colonnes du tableau : diagramme en bâtons pour les effectifs ou les fréquences et **diagrammes cumulatifs** pour les effectifs cumulés ou les fréquences cumulées.

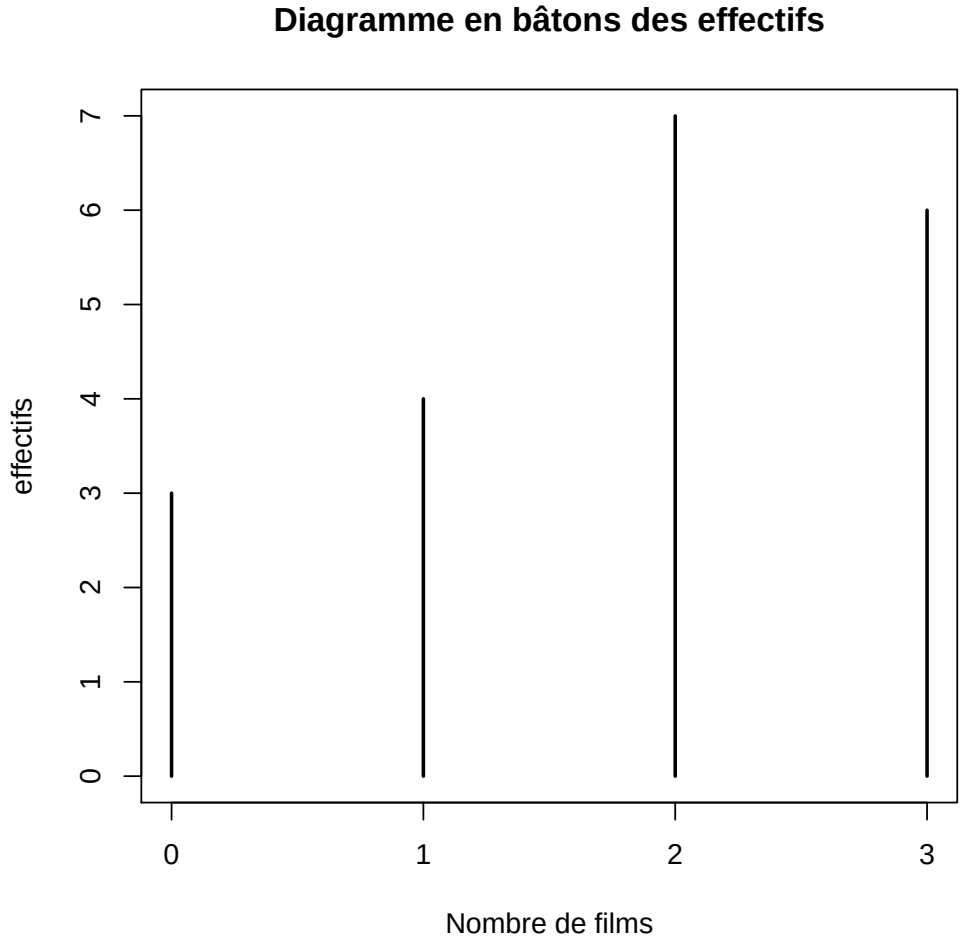
Les diagrammes en bâtons : les **diagrammes en bâtons** servent à représenter les effectifs ou les fréquences de l'ensemble des valeurs d'une variable **quantitative discrète**.

Principe : pour tracer un diagramme en bâtons, on choisit tout d'abord deux axes perpendiculaires et une échelle pour chacun de ces axes. L'axe des abscisses (ou axe horizontal) sert à porter les valeurs de la variable et l'axe des ordonnées (axe vertical) est l'axe des effectifs ou des fréquences suivant le cas. Il suffit ensuite de tracer pour chaque valeur un trait vertical (bâton) dont la hauteur correspond à la valeur de l'effectif ou de la fréquence.

Remarque 2.1. *Les diagrammes en bâtons des effectifs et des fréquences d'une même variable diffèrent simplement par l'échelle des ordonnées : on passe, par exemple, du diagramme en bâtons des effectifs au diagramme en bâtons des fréquences en divisant l'échelle des ordonnées par N (taille de l'échantillon).*

Exemple 2.2. *On reprend les données de l'exemple ci-dessus.*

```
> plot(table(films1),xlab="Nombre de films",ylab="effectifs",main="Diagramme en bâtons des eff
```



Les diagrammes cumulatifs : les **diagrammes cumulatifs** sont des graphes de fonctions en escaliers servant à représenter les effectifs ou fréquences cumulé(e)s d'une variable quantitative discrète. Le diagramme cumulatif est par définition le graphe de la fonction qui à toute valeur réelle x fait correspondre le nombre (ou la proportion) d'individus ayant pris une valeur inférieure ou égale à x.

Principe : on choisit, pour tracer un diagramme cumulatif, deux axes gradués perpendiculaires : l'axe des abscisses est identique à celui du diagramme en bâtons et est donc l'axe des valeurs de X tandis que sur l'axe des ordonnées sont portées les effectifs cumulés ou les fréquences cumulées suivant le cas. L'axe des ordonnées est ainsi gradué de 0 à N (lorsqu'on veut représenter les effectifs cumulés) ou de 0 à 1 (lorsqu'on veut représenter les fréquences cumulées). Les deux diagrammes sont identiques à l'échelle des ordonnées près.

On voit ainsi que dans le cas général le diagramme cumulatif des effectifs est le graphe d'une fonction en escalier qui vaut :

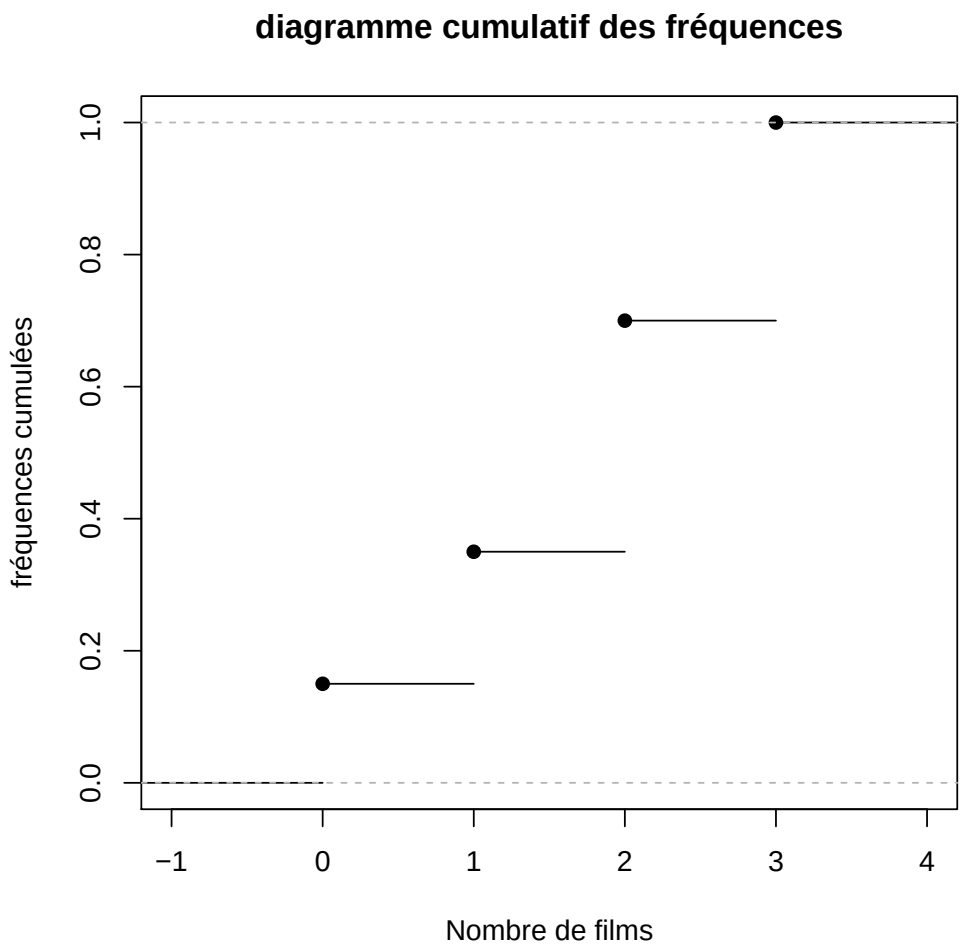
$$\begin{cases} 0, & \text{si } x < x_1, \\ N_j, & \text{si } x_j \leq x < x_{j+1},\ j = 1, \ldots, k-1, \\ N, & \text{si } x_k \leq x, \end{cases}$$

où x_j, $j = 1, \ldots, k$ sont les valeurs distinctes de la variable statistique et N_j l'effectif cumulé de x_j.

On trace suivant le même principe le diagramme cumulatif des fréquences en remplaçant les N_j par les F_j : à toute valeur x de la droite réelle on fait correspondre ainsi la proportion d'individus (de l'échantillon) ayant pris une valeur inférieure ou égale à x.

Exemple 2.3. *On reprend les données de l'exemple ci-dessus.*

```
> plot(ecdf(films1),xlab="Nombre de films",ylab="fréquences cumulées",main="diagramme cumulati
```



Résumé des données - Caractéristiques numériques

Le but est ici de définir des **caractéristiques numériques** permettant de résumer la variable X. Puisqu'il s'agit d'un résumé, il est clair que toute l'information contenue dans le tableau statistique ne sera pas représentée par ces caractéristiques. Il y a deux types de caractéristiques numériques :

— les **caractéristiques de position centrale**, qui servent à "situer" les observations de la série sur la droite réelle ;

— les **caractéristiques de dispersion** qui servent à évaluer la dispersion des observations autour de positions centrales.

Caractéristiques de tendance centrale

La moyenne arithmétique : c'est la quantité, notée $\overline{X}$, définie par :

$$\overline{X} \;=\; \frac{1}{N}\sum_{j=1}^{k} n_j x_j.$$

Remarque 2.2. *1. L'unité de mesure de la moyenne est celle des observations. Notons que la moyenne n'est pas nécessairement une valeur de la variable.*

2. Autre expression de la moyenne :

$$\overline{X} \;=\; \sum_{j=1}^{k} \frac{1}{N} n_j x_j \;=\; \sum_{j=1}^{k} f_j x_j.$$

La médiane : toute valeur, notée m, telle qu'il y ait autant d'individus ayant pris une valeur inférieure ou égale à m que d'individus ayant pris une valeur supérieure ou égale à m.

Calcul de la médiane. Prenons trois exemples simples pour illustrer les différents cas pouvant se présenter.
1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

$$1,\ 3,\ 4,\ 5,\ 5.$$

Ce cas est le plus simple puisqu'il y a 3 individus qui ont pris une valeur inférieure ou égale à 4 et 3 qui ont pris une valeur supérieure ou égale à 4 : la médiane de la série est donc $m = 4$.
2. Sur un échantillon de 6 individus on a observé les valeurs suivantes :

$$1,\ 3,\ 4,\ 5,\ 6,\ 6.$$

Toute valeur comprise entre 4 et 5 peut convenir pour la médiane : si $4 < m < 5$, il y a 3 individus ayant pris une valeur inférieure à m et 3 individus ayant pris une valeur supérieure à m. On dit alors que $]4, 5[$ est un intervalle médian. On prend alors par convention comme valeur de la médiane le centre de cet intervalle c'est-à-dire $m = 4,5$.
3. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

$$1,\ 3,\ 4,\ 4,\ 5.$$

Dans ce cas, il y a 4 observations inférieures à 4 et 3 observations supérieures à 4. On prend cependant (le premier) 4 pour valeur de la médiane en notant qu'il y a 3 individus qui ont pris une valeur inférieure ou égale à ce premier 4 et 3 individus qui ont pris une valeur supérieure ou égale.

A partir de ces trois exemples, on adopte les règles de calcul suivantes pour déterminer la médiane : on range les modalités par ordre croissant et on porte les effectifs cumulés.
- Si N est impair, la médiane est la valeur correspondant au premier effectif cumulé supérieur ou égal à $\dfrac{N+1}{2}$;
- Si N est pair, la médiane est le centre des deux valeurs correspondant aux effectifs cumulés $\dfrac{N}{2}$ et $\dfrac{N+2}{2}$ dans le cas où ces effectifs cumulés apparaissent dans le tableau et sinon on prend pour la médiane la valeur correspondant au premier effectif cumulé supérieur (strictement) à $\dfrac{N+2}{2}$.

Le mode : toute valeur de la série dont l'effectif (ou la fréquence) est supérieur aux effectifs (ou aux fréquences) de la valeur précédente et de la valeur suivante.

Remarque 2.3. *Une série statistique peut avoir plusieurs modes. Dans le cas d'une série ayant deux modes, on parle de série statistique* **bimodale**.

Caractéristiques de dispersion

Les caractéristiques de tendance centrale vues dans la section précédente ne suffisent pas à elles seules à résumer une série statistique. En particulier deux séries très différentes peuvent avoir la même moyenne : la moyenne des 4 valeurs 2, 5, 15, 18 est 10 et celle des 4 valeurs 9, 10, 10, 11

est également 10. la première série est beaucoup plus "dispersée" que la seconde. On souhaite alors définir des caractéristiques numériques permettant de rendre compte de la dispersion des observations. On considère dans la suite une variable quantitative X observée sur une population Ω (ou sur un échantillon).

L'étendue : c'est la différence entre la plus grande et la plus petite observation.

L'intervalle interquartile : on définit tout d'abord les **quartiles** d'une série statistique.

Quartiles d'une série statistique : ce sont les trois quantités q_1, q_2 et q_3 telles que :
q_1 est la valeur telle que 25% des observations lui sont inférieures ou égales et 75% lui sont supérieures ou égales ;
q_2 est la valeur telle que 50% des observations lui sont inférieures ou égales et 50% lui sont supérieures ou égales ;
q_3 est la valeur telle que 75% des observations lui sont inférieures ou égales et 25% lui sont supérieures ou égales.

Remarque 2.4. *Le calcul des quartiles se fait sur le même principe que celui de la médiane.*

Intervalle interquartile d'une série statistique : c'est l'intervalle $[q_1, q_3]$.

A partir des quartiles, on trace la boîte à moustaches ou boîtes à pattes (box-plot en anglais).
Généralisation. Quantiles. On définit de manière analogue le **quantile** d'ordre α, $\alpha \in]0,1[$: ce quantile est la valeur telle qu'une proportion α des observations lui sont inférieures ou égales et une proportion $1-\alpha$ lui sont supérieures ou égales.
Les **déciles** sont ainsi les 9 valeurs correspondant aux quantiles d'ordre 0,1 (ou 10%-, 0,2 (ou 20%), ..., 0,9 (ou 90%).
Les **centiles** sont les 99 valeurs correspondant aux quantiles d'ordre 0,01 (ou 1%-, 0,02 (ou 2%), ..., 0,99 (ou 99%).

La variance : c'est la quantité, notée σ_X^2 définie par :

$$\sigma_X^2 \;=\; \frac{1}{N}\sum_{j=1}^{k} n_j \left(x_j - \overline{X}\right)^2 .$$

Remarque 2.5. *1. La variance est la moyenne des carrés des écarts à la moyenne : on calcule tout d'abord les différents carrés des écarts $\left(x_j - \overline{X}\right)^2$ dont on calcule ensuite la moyenne. Ainsi, la variance mesure la distance moyenne entre les observations et la moyenne.*
2. On a également une formule de la variance à l'aide des fréquences :

$$\sigma_X^2 \;=\; \sum_{j=1}^{k} f_j \left(x_j - \overline{X}\right)^2 .$$

3. Dans la pratique on utilise pour calculer la variance une expression plus commode à manipuler. On montre en effet que la variance s'écrit :

$$\sigma_X^2 \;=\; \frac{1}{N}\sum_{j=1}^{k} n_j x_j^2 - \left(\overline{X}\right)^2 .$$

On voit ainsi que la variance est également la moyenne de la variable X^2 moins le carré de la moyenne de X.

L'écart-type ; c'est la quantité, notée σ_X, égale à la racine carrée de sa variance :

$$\sigma_X \;=\; \sqrt{\sigma_X^2}.$$

Remarque 2.6. *Contrairement à la variance (qui n'a pas réellement d'unité) l'écart-type s'exprime dans l'unité de la variable X.*

Exemple 2.4. *Pour l'exemple ci-dessus, on trouve les différentes caractéristiques :*
$\overline{X} = 1,8$, $\sigma_X^2 = 1,06$, $\sigma_X \simeq 1,03$;
$q_1 = 1$, $q_2 = 2$, $q_3 = 3$;
étendue : $[0, 3]$.
Avec le logiciel R on obtient un résumé des caractéristiques numériques la variable en tapant :

```
> summary(films1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     1.0     2.0     1.8     3.0     3.0
```

que l'on retrouve à l'aide de différentes commandes :

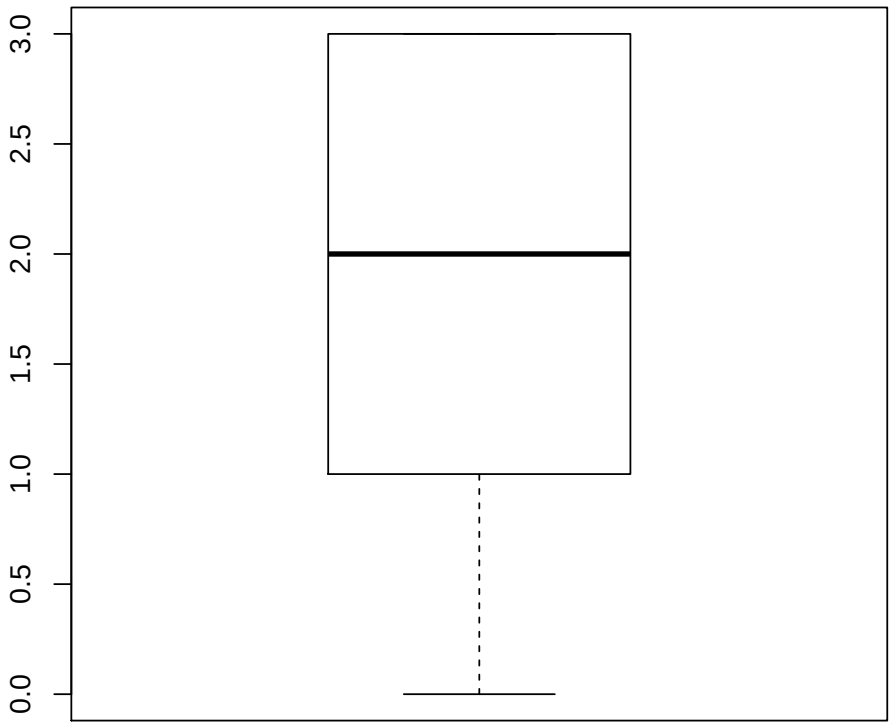
```
> mean(films1)

[1] 1.8

> min(films1)

[1] 0

> max(films1)

[1] 3

> median(films1)

[1] 2
```

La fonction var calcule la **variance de l'échantillon** *ou* **variance corrigée**, *c'est-à-dire la quantié* $\frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{X})^2$ *et sd calcule l'***écart-type corrigé** *c'est-à-dire la racine carrée de la variance corrigée. Pour retrouver la variance de la population (définie ci-dessus), on doit donc multiplier par* $\frac{N-1}{N}$ *puis calculer la racine carrée pour avoir l'écart-type :*

```
> var(films1)

[1] 1.115789

> 19*var(films1)/20

[1] 1.06

> sd(films1)

[1] 1.056309

> sqrt(19*var(films1)/20)

[1] 1.029563
```

On peut également obtenir la **boîte à moustaches** *ou* **boxplot** *qui rpermet de visualiser l'étendue de la série et ses quartiles :*

```
> boxplot(films1)
```



2.1.2 Variables quantitatives continues

On appelle **variable quantitative continue** une variable quantitative pouvant prendre les valeurs d'un intervalle. Le plus souvent, ces valeurs sont des entiers (rarement des valeurs décimales) et le nombre de valeurs distinctes est assez faible, par exemple le *nombre d'enfants*. C'est ce dernier cas que nous considérons dans la suite : ainsi, nous notons $x_1, x_2, \ldots, x_k$ les valeurs distinctes de la variable, k étant le nombre de valeurs distinctes avec $k < N$. Enfin, les valeurs seront toujours ordonnées suivant l'ordre croissant : $x_1 < x_2 < \ldots < x_k$.

Organisation des données - Tableau statistique

Regroupement en classes : le nombre de valeurs distinctes observées d'une variable quantitative continue est important (souvent égal au nombre d'individus N). On regroupe alors les valeurs en classes, c'est-à-dire en intervalles de valeurs. On note par k le nombre de classes et par $[b_j, b_{j+1}[$, $j = 1, \ldots, k$, les différentes classes.
Par convention, l'intervalle définissant une classe est fermé à gauche et ouvert à droite, de sorte qu'il n'y a pas de chevauchement des classes ; celles-ci sont disjointes.

Amplitude : c'est la largeur de la classe, notée a_j : $a_j = b_{j+1} - b_j$.
On choisit généralement des classes d'amplitude constante, bien que pour diverses raisons on peut (notamment pour les classes extrémales) choisir des amplitudes différentes.

Bornes de la classe : ce sont les valeurs minimales et maximales : b_j et b_{j+1}.

Centre : valeur équidistante des deux bornes : $x_j = \dfrac{b_j + b_{j+1}}{2}$.

Les centres de classes peuvent être utilisés pour **discrétiser** une variable quantitative continue. En effet, on peut résumer la classe $[b_j, b_{j+1}[$ par son centre x_j. Si maintenant on considère les centres ainsi définis en lieu et place des classes, cela revient à considérer une variable quantitative discrète.

Le tableau statistique est est analogue à celui présenté ci-dessus : dans la première colonne sont portées les classes rangées par ordre croissant. Dans les colonnes suivantes sont portés les effectifs, effectifs cumulés, fréquences et fréquences cumulées correspondant aux classes de la variable X. Ceux-ci sont définis de la même manière que pour une variable quantitative discrète.

Exemple 2.5. *(Source : Bertrand, F. et Maumy, M.) On a relevé les poids (en grammes) de souris soumises à une expérience de supplémentation en vitamines :*

74, 85, 95, 84, 68, 93, 84, 87, 78, 72, 81, 91, 80, 65, 76, 81, 97, 69, 70, 98

1. Regroupement en classes d'amplitudes constantes à l'aide de la fonction hist :

```
> souris<-c(74, 85, 95, 84, 68, 93, 84, 87, 78, 72, 81, 91, 80, 65, 76, 81, 97, 69, 70, 98)
> souris.hist1<-hist(souris)
> souris.hist1

£breaks
[1]  65  70  75  80  85  90  95 100

£counts
[1] 4 2 3 5 1 3 2

£density
[1] 0.04 0.02 0.03 0.05 0.01 0.03 0.02

£mids
[1] 67.5 72.5 77.5 82.5 87.5 92.5 97.5

£xname
[1] "souris"

£equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

les bornes de classes sont 65, 70, 75, 80, 85, 90, 95, 100 (7 classes d'amplitudes constantes 5g). Les effectifs respectifs sont 4, 2, 3, 5, 1, 2 et les densités de réquences 0,04, 0,02, 0,03, 0,05, 0,01, 0,03, 0,02. On en déduit le tableau statistique :

Classes	*Effectifs*	*Effectifs cumulés*	*Fréquences*	*Fréquences cumulées*	*Densités de fréquences*
$[65, 70[$	*4*	*4*	*0,2*	*0,2*	*0,04*
$[70, 75[$	*2*	*6*	*0,1*	*0,3*	*0,02*
$[75, 80[$	*3*	*9*	*0,15*	*0,45*	*0,03*
$[80, 85[$	*5*	*14*	*0,25*	*0,7*	*0,05*
$[85, 90[$	*1*	*15*	*0,05*	*0,75*	*0,01*
$[90, 95[$	*3*	*18*	*0,15*	*0,9*	*0,03*
$[95, 100[$	*2*	*20*	*0,1*	*1*	*0,02*

2. Regroupement en classes choisies par l'utilisateur : on choisit 3 classes de bornes 60, 70, 90, 100 et d'amplitudes respectives 10, 20, 10.

```
> souris.hist2<-hist(souris,breaks=c(60,70,90,100))
> souris.hist2

$breaks
[1]  60  70  90 100

$counts
[1]  4 11  5

$density
[1] 0.0200 0.0275 0.0250

$mids
[1] 65 80 95

$xname
[1] "souris"

$equidist
[1] FALSE

attr(,"class")
[1] "histogram"
```

Tableau statistique :

Classes	*Effectifs*	*Effectifs cumulés*	*Fréquences*	*Fréquences cumulées*	*Densités de fréquences*
$[60, 70[$	*4*	*4*	*0,2*	*0,2*	*0,02*
$[70, 90[$	*11*	*15*	*0,55*	*0,75*	*0,0275*
$[90, 100[$	*5*	*20*	*0,25*	*1*	*0,025*

Représentations graphiques

Les histogrammes : ils servent à représenter les effectifs ou les fréquences d'une variable **quantitative continue**.

Principe : on choisit deux axes perpendiculaires et une échelle pour chacun. Sur l'axe des abscisses (axe horizontal) sont portées les valeurs de la variable, c'est-à-dire les différentes classes de cette variable. En face de chaque classe, on trace un "rectangle" dont la hauteur est égale à la **densité d'effectif** ou à la **densité de fréquence** de cette classe. Si a_j est l'amplitude de la $j^{\text{ème}}$ classe $(1 \leq j \leq k)$ et si n_j et f_j sont respectivement son effectif et sa fréquence, la densité d'effectif pour cette classe est :

$$\frac{n_j}{a_j},$$

tandis que la densité de fréquence est :

$$\frac{f_j}{a_j}.$$

Justification. Considérons le cas de deux classes $[1, 2[$ et $[2, 4[$ d'amplitudes respectives 1 et 2 et ayant le même effectif 10. Supposons par ailleurs que les observations soient uniformément réparties dans les classes. On voit facilement que tracer un rectangle de la même hauteur pour les deux classes ne

conviendrait pas et conduirait à une interprétation fausse. En effet, si on adopte un autre découpage en partageant la classe $[2, 4[$ en deux classes $[2, 3[$ et $[3, 4[$ ayant donc chacune un effectif de 5. On aurait alors un autre histogramme. On voit ainsi que la hauteur de chaque rectangle doit être proportionnelle à l'amplitude de la classe. Ces arguments sont bien entendu valables pour l'histogramme des fréquences.

Remarque 2.7. *1. L'***aire** *de chaque rectangle est égale à l'effectif ou à la fréquence de la classe qu'il représente. Pour les effectifs par exemple, l'aire d'un rectangle est le produit de sa base a_j par sa hauteur $\frac{n_j}{a_j}$, ce qui donne n_j.*
Si on calcule l'aire totale de l'histogramme, c'est-à-dire la somme des aires de tous les rectangles, on obtient

$$n_1 + \ldots + n_k \ = \ N,$$

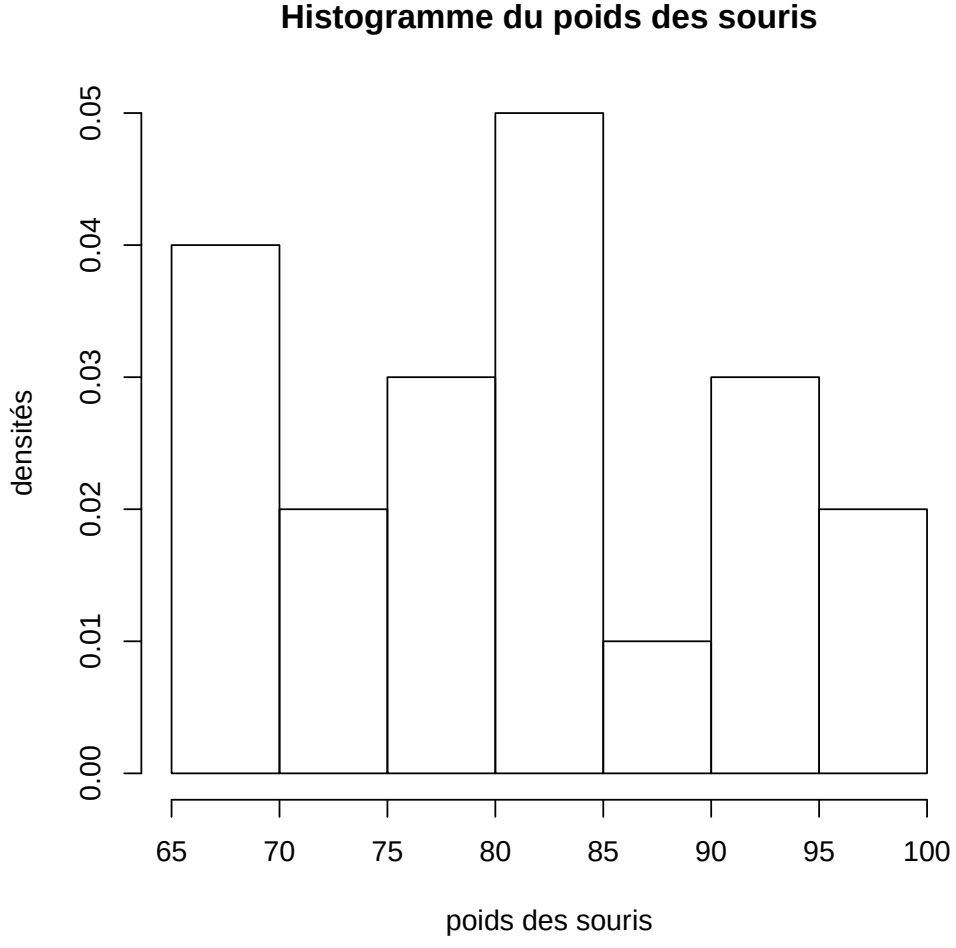
dans le cas des effectifs et :

$$f_1 + \ldots + f_k \ = \ 1,$$

dans le cas des fréquences. L'aire de l'histogramme des effectifs est égale à la taille de l'échantillon N et l'aire de l'histogramme des fréquences est égale à 1.
2. Le choix de l'amplitude et de la position des classes est un problème important en pratique : deux choix distincts de classes peuvent conduire à des histogrammes d'allures très différentes.

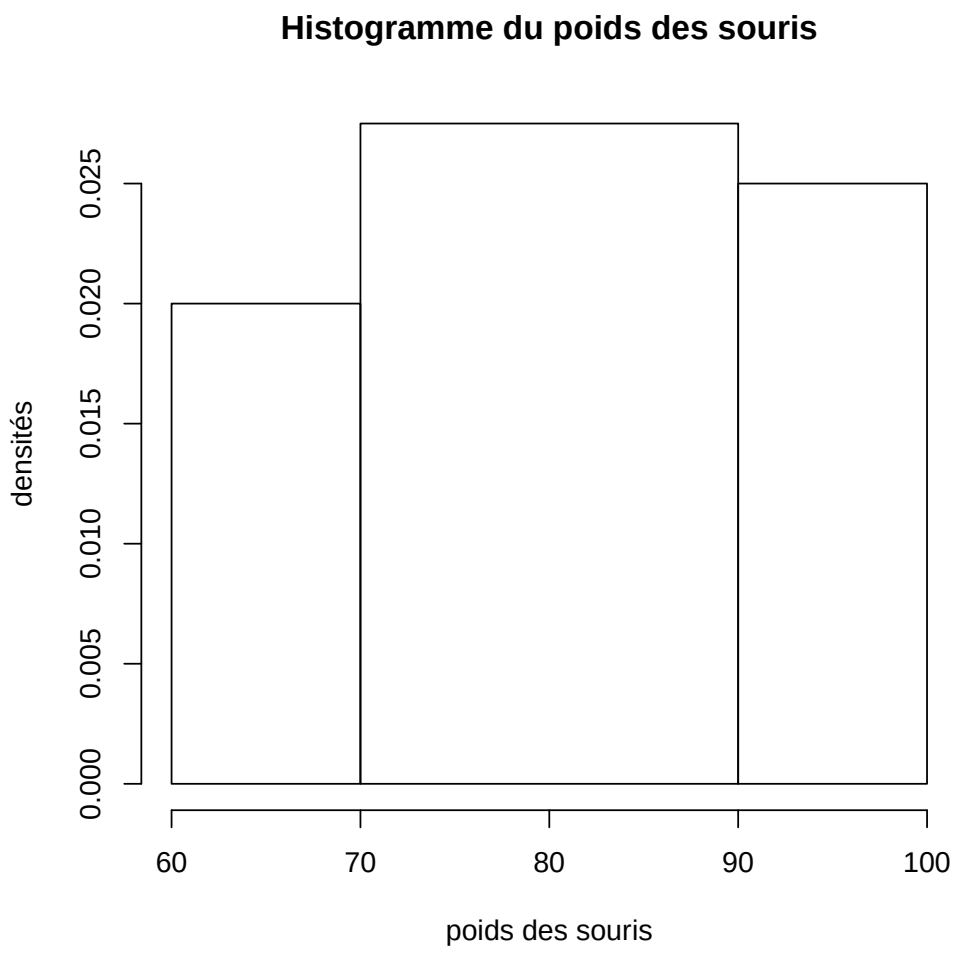
Exemple 2.6. *On reprend les données de l'exemple 2.5 et on trace l'histogramme pour le premier découpage d'amplitude constante égale à 5 :*

```
> hist(souris,freq=FALSE,xlab="poids des souris",ylab="densités",main="Histogramme du poids de
```



Pour le second découpage :

```
> hist(souris,breaks=c(60,70,90,100),freq=FALSE,xlab="poids des souris",ylab="densités",main="
```



Les polygones cumulatifs : ce sont des graphes (lignes brisées) servant à représenter les effectifs ou fréquences cumulé(e)s d'une variable quantitative continue. Le polygone cumulatif est par définition le graphe de la fonction qui à toute valeur réelle fait correspondre le nombre (ou la proportion) d'individus pour lesquels l'observation de la variable est inférieure à cette valeur.

Après avoir choisi deux axes gradués, on porte sur l'axe des abscisses les valeurs de la variable (bornes de classes) et sur l'axe des ordonnées les effectifs ou les fréquences cumulé(e)s. Le principe de construction du polygone cumulatif est le suivant :

- si x est un réel inférieur ou égal à b_1 (borne inférieure de la première classe) la valeur du polygone cumulatif en x est nulle ;
- si x est un réel supérieur ou égal à b_{k+1} (borne supérieure de la dernière classe) la valeur du polygone cumlatif des effectifs en x est égale à N et celle du polygone cumulatif des fréquences est égale à 1 ;
- si x est un réel égal à l'une des bornes supérieures de classes, c'est-à-dire $x = b_{j+1}$ où $1 \leq j \leq k$, alors la valeur du polygone cumulatif au point d'abscisse x est la valeur de l'effectif ou de la fréquence cumulé(e) de la classe $[b_j, b_{j+1}[$;
- entre deux bornes de classe, le polygone cumulatif est un segment joignant les points définis à l'étape précédente, ce qui revient à faire une interpolation linéaire.

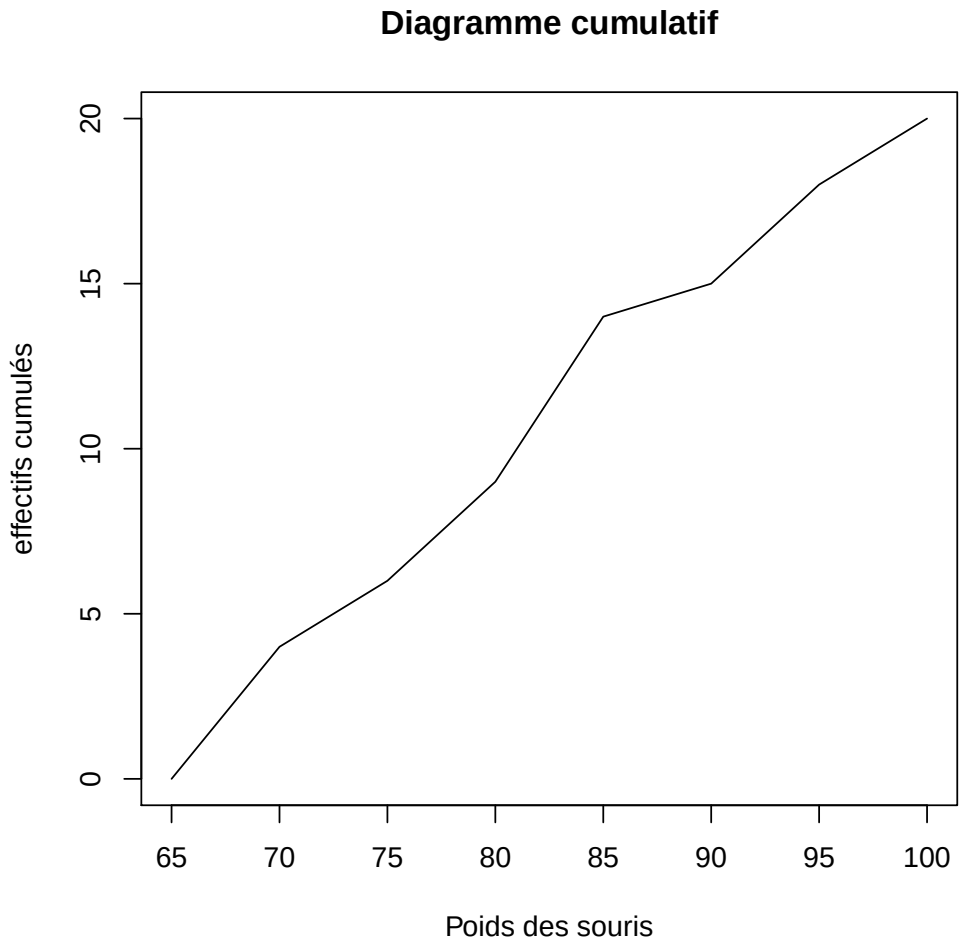
Remarque 2.8. *Le principe de construction du polygone cumulatif s'appuie sur l'hypothèse d'"uniformité" de la répartition des observations à l'intérieur de chaque classe vue plus haut. Cette hypothèse*

implique que la fonction augmente de manière constante dans chaque classe et est donc une fonction affine par morceaux.

Exemple 2.7. *Traçons le polygone cumulatif pour les données de l'exemple 2.5 regroupées en classes d'amplitude 5. On calcule d'abord les effectifs cumulés :*

```
> cumsum(souris.hist1$counts)

[1]  4  6  9 14 15 18 20
```

On doit ajouter la valeur 0 qui correspond à la première borne de classe (65).



Résumé des données - Caractéristiques numériques

La moyenne, la variance et l'écart-type d'une variable quantitative continue sont définis de la même manière que pour une variable quantitative discrète en prenant pour x_j les centres de classes.

La médiane. Dans le cas d'une variable quantitative continue, la médiane est la valeur de la série correspondant à la fréquence cumulée 0,5. Elle se lit donc directement sur le polygone cumulatif (des fréquences).
On peut calculer de manière précise la médiane en réalisant une interpolation linéaire. Les classes sont rangées par ordre croissant et les fréquences cumulées sont portés dans le tableau statistique. Soit $[b_j, b_{j+1}[$ la classe correspondant à la fréquence cumulée supérieure ou égale à 0,5. Il y a donc moins de 50% des individus qui ont pris une valeur inférieure à b_j et moins de 50% des individus qui ont

pris une valeur supérieure à b_{j+1}. Soit F_{j-1} la fréquence cumulée de la classe $[b_{j-1}, b_j[$ et F_j celle de la classe $[b_j, b_{j+1}[$. Pour déterminer la médiane on réalise une interpolation linéaire :

$$m \;=\; b_j \;+\; \left(\frac{0,5 - F_{j-1}}{F_j - F_{j-1}}\right)(b_{j+1} - b_j).$$

Les **quartiles** sont calculés dur le même principe.

Le mode d'une variable quantitative continue est le centre de toute classe de la série dont la densité d'effectif (ou la densité de fréquence) est supérieure aux densités d'effectifs (ou aux densités de fréquences) de la classe précédente et de la classe suivante.

Exemple 2.8. *Comme pour une variable quantitative discrète on peut calculer les caractéristiques numériques à partir des données brutes (c'est-à-dire du vecteur souris). Cependant sin on veut utiliser les centres de classes, on utilise les valeurs obtenus dans l'objet souris.hist1 ou souris.hist2 suivant le découpage. Par exemple, pour le premier découpage, les centres sont 67,5, 72,5, 77,5, 82,5, 87,5, 92,5, 97,5 et les effectifs correspondants 4, 2, 3, 5, 1, 3, 2. On calcule alors la moyenne :*

```
> weighted.mean(souris.hist1$mids,souris.hist1$counts)

[1] 81
```

On peut calculer de la même manière la variance et l'écart-type. Le calcul de la médiane et des quartiles nécessitent de revenir aux définitions.

2.2 Variables qualitatives

Organisation des données - Tableau statistique

Le tableau statistique est constitué d'une première colonne dans laquelle sont portées les modalités de la variable. Viennent ensuite les effectifs puis les fréquences de chaque modalité sur les colonnes suivantes. Les effectifs et les fréquences sont définis de la même manière que précédemment. Dans le cas d'une variable qualitative nominale, les notions d'effectifs cumulés et de fréquences cumulées n'ont pas de sens. Pour une variable qualitative ordinale on définit les effectifs cumulés et de fréquences cumulées de la même manière que pour une variable quantitative (les modalités de la variable sont rangées par ordre croissant dans ce cas).

Représentations graphiques

Les diagrammes en colonnes : servent à représenter les effectifs ou les fréquences d'une variable **qualitative**.

Pour tracer un diagramme en colonnes, on porte les modalités de la variable (qualitative) le long d'une ligne horizontale (il ne s'agit pas d'un axe muni d'une échelle puisque les modalités n'ont pas ici de valeurs numériques). Les modalités sont régulièrement espacées sur cet axe. L'axe des ordonnées (axe vertical) est l'axe des effectifs ou des fréquences suivant le cas. En face de chaque modalité figure une colonne (un rectangle) dont la hauteur correspond à la valeur de l'effectif ou de la fréquence.

Les diagrammes en secteurs et les diagrammes en barre : les diagrammes en secteurs (**camemberts**) et les diagrammes en barre servent (comme les diagrammes en colonnes) à représenter les effectifs ou les fréquences d'une variable **qualitative**.

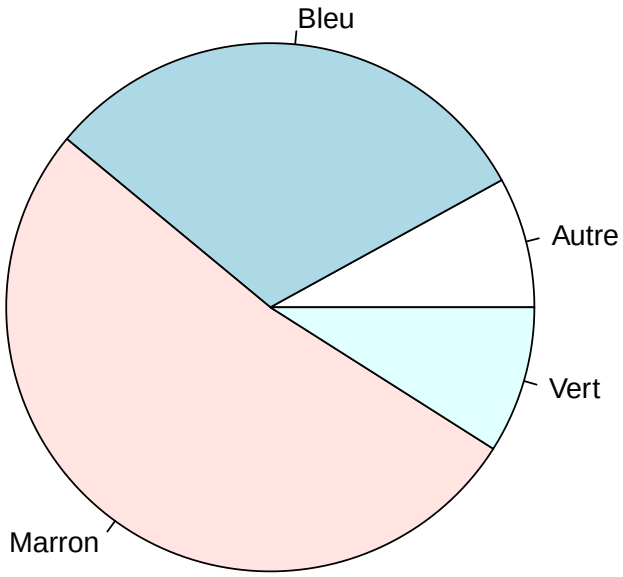
Principe : les diagrammes en secteurs se présentent sous la forme d'un disque divisé en k secteurs (k étant le nombre de modalités de la variable) : l'angle (ou l'aire ce qui revient au même) de chaque secteur est proportionnel à l'effectif ou à la fréquence de la modalité qu'il représente. Les diagrammes en barre sont construits sur le même principe mais sous la forme d'un rectangle divisé en k sous-rectangles dont les aires sont proportionnelles aux effectifs ou fréquences des modalités qu'ils représentent.

Exemple 2.9. *Le tableau ci-dessous donne la répartition de la couleur des yeux dans une population.*

Couleur des yeux	*effectifs*
Marron	*52*
Bleus	*31*
Verst	*9*
Autre	*8*

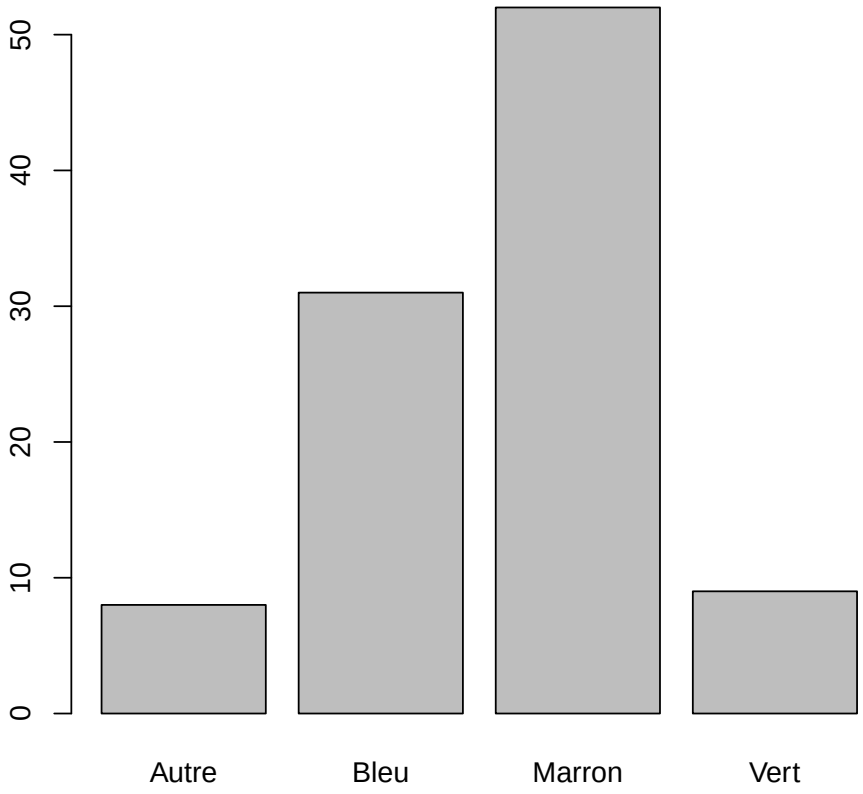
Les données sont contenues dans le vecteur couleur_yeux. On obtient le diagramme en secteurs :

```
> pie(table(couleur_yeux))
```



et le diagramme en colonnes :

```
> barplot(table(couleur_yeux))
```


0
10
20
30
40
50
Autre
Bleu
Marron
Vert

Chapitre 3

Statistique descriptive bidimensionnelle

Dans ce chapitre nous considérons une population Ω (ou un échantillon issu de cette population) de taille N. Sur cette population nous étudions deux variables statistiques X et Y. Ces variables peuvent être quantitatives ou qualitatives (ordinales ou nominales). Le but est ici d'analyser la **liaison** (dépendance) entre ces deux variables en donnant tout d'abord un sens à cette notion.

3.1 Le khi-deux

En principe, le khi-deux est défini pour deux variables qualitatives X et Y mais de manière générale il peut être défini pour des variables quelconques prenant cependant un nombre réduit de valeurs distinctes. Le khi-deux est une mesure de liaison pour lequel les variables X et Y jouent un rôle symétrique. Il n'y a donc pas dans cas de notion de causalité entre X et Y.

Dans tous les cas, les variables sont observées sur N individus et nous notons $x_1, \ldots, x_k$ les modalités de la variable X et $y_1, \ldots, y_m$ celles de la variable Y (X a donc k modalités et Y a m modalités).

3.1.1 Tableau statistique : la table de contingence

Effectifs conjoints : l'effectif conjoint du couple (x_i, y_j) le nombre d'individus, noté n_{ij}, est pris la modalité x_i de la variable X et la modalité y_j de la variable Y. On a :

$$\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \;=\; N$$

(la sommen de tous les effectifs conjoints est égale à N).

On définit également la **fréquence conjointe** du couple (x_i, y_j) qui est la proportion d'individus, noté f_{ij}, ayant pris la modalité x_i de la variable X et la modalité y_j de la variable Y. On a donc : et :

$$\sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij} \;=\; 1.$$

Effectifs marginaux : l'effectif marginal de la modalité x_i de la variable X est le nombre d'individus, noté $n_{i.}$, ayant pris la modalité x_i. L'effectif marginal $n_{i.}$ s'obtient en additionnant les effectifs conjoints n_{ij} pour $j = 1, \ldots, m$:

$$n_{i.} \;=\; \sum_{j=1}^{m} n_{ij}.$$

De même, on appelle effectif marginal de la modalité y_j de la variable Y, le nombre d'individus, noté $n_{.j}$, ayant pris la modalité y_j. On a :

$$n_{.j} \;=\; \sum_{i=1}^{k} n_{ij}.$$

On définit également les **fréquences marginales** :

$$f_{i.} = \sum_{j=1}^{m} f_{ij} = f_{i.} = \frac{n_{i.}}{N}$$

et

$$f_{.j} = \sum_{i=1}^{k} f_{ij} = f_{.j} = \frac{n_{.j}}{N}.$$

La table de contingence : ce tableau est constitué de k lignes et m colonnes : sur les lignes sont portées les modalités de la variable X et sur les colonnes les modalités de la variable Y ; à l'intersection de la ligne i et de la colonne j est porté l'effectif (ou la fréquence suivant le cas) conjoint du couple de modalités (x_i, y_j). Enfin, on ajoute une ligne et une colonne supplémentaires dans lesquelles sont portés respectivement les effectifs (ou les fréquences) marginaux de la variable Y et de la variable X.

Exemple 3.1. *Trois échantillons de plantes sont traités avec trois engrais différents. Les résultats obtenus sont les suivants.*

Y / X	*Engrais A*	*Engrais B*	*Engrais C*	*Total*
N'ont pas fleuri	*16*	*12*	*12*	*40*
Ont fleuri	*34*	*73*	*63*	*170*
Total	*50*	*85*	*75*	

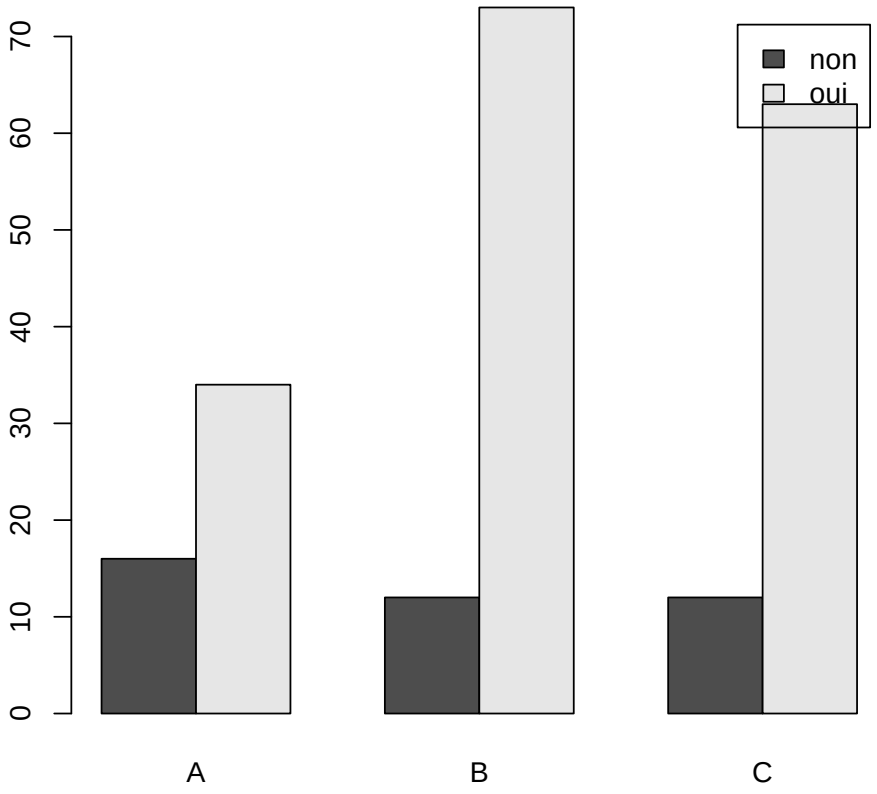
Les données ont été recensées dans le data.frame plante dans le logiciel R. Celui-ci contient deux colonnes "floraison" et "engrais" de longueur $N = 210$.

```
> table(plante$floraison,plante$engrais)
```

```
     A  B  C
non 16 12 12
oui 34 73 63
```

On obtient une représentation graphique de ce tableau :

```
> barplot(table(plante),beside=TRUE,legend=rownames(table(plante)))
```



Fréquences conditionnelles : la fréquence de la modalité y_j de Y conditionnellement à la modalité x_i de X est définie par :

$$\frac{n_{ij}}{n_{i.}}$$

C'est donc la proportion d'individus ayant pris la modalité y_j parmi ceux qui ont pris la modalité x_i. On définit ainsi le $i^{\text{ème}}$ **profil-ligne** qui est constitué des fréquences conditionnelles à x_i :

$$\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \ldots, \frac{n_{im}}{n_{i.}}.$$

De même, la fréquence de la modalité x_i conditionnellement à la modalité y_j la quantité définie par :

$$\frac{n_{ij}}{n_{.j}}.$$

et le $j^{\text{ème}}$ **profil-colonne** par :

$$\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \ldots, \frac{n_{kj}}{n_{.j}}.$$

Remarque 3.1. *La somme des fréquences conditionnelles à une modalité est égale à 1 :*

$$\sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} \;=\; 1.$$

3.1.2 Représentation graphique

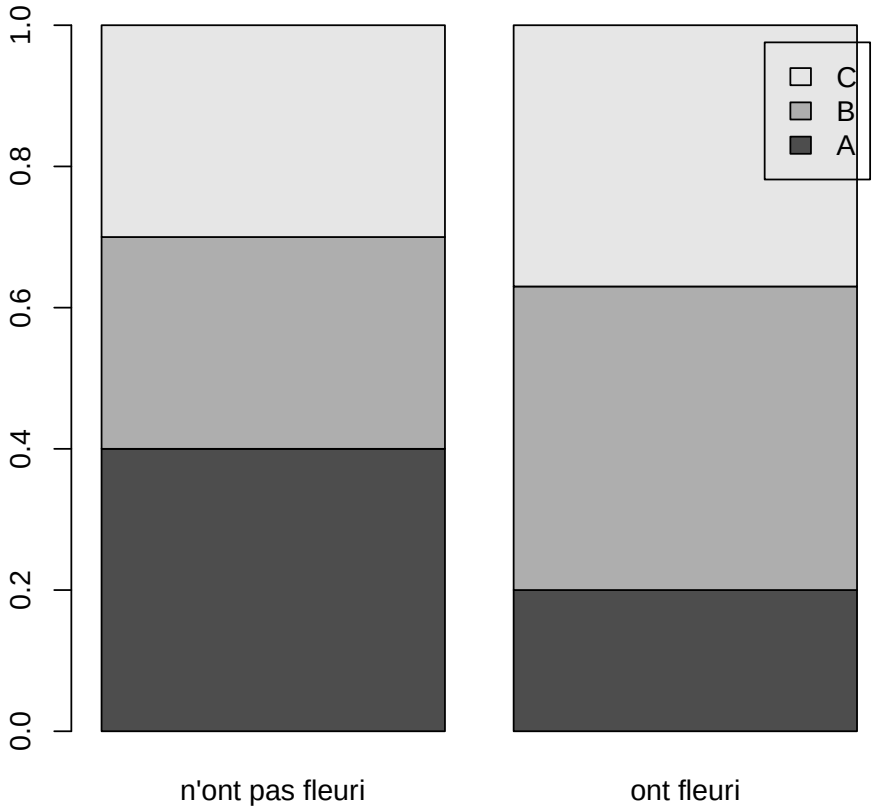
On représente les profils-lignes (ou les profils-colonnes) par un diagramme en barres parallèles.

Ce graphique est un premier outil permettant d'analyser une possible liaison entre les variables X et Y. En effet, si les deux variables ne sont pas liées alors les répartitions doivent être similaires pour chaque modalité x_j (barres identiques). Dans le cas contraire, on devrait avoir des différences plus ou moins importantes.

Exemple 3.2. *Pour l'exemple 3.1, on obtient le tableau des profils-lignes :*

Y / X	*Engrais A*	*Engrais B*	*Engrais C*	*Total*
N'ont pas fleuri	*0,4*	*0,3*	*0,3*	*1*
Ont fleuri	*0,2*	*0,43*	*0,37*	*1*

et sa représentation graphiques :



On voit ici qu'il y a des différences notables entre les répartitions des engrais suivant la floraison, ce qui laisse supposer qu'il y a une liaison entre les deux variables.

3.1.3 Mesure de liaison entre deux variables statistiques : le χ^2

État de non liaison. Suivant a remarque faite à la section précédente, on dit que les variables statistiques X et Y sont **non liées** (pour l'échantillon considéré) si et seulement si :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{lj}}{n_{l.}},\ j = 1, \ldots, m,\ i = 1, \ldots, k,\ l = 1, \ldots, k$$

Autrement dit les différentes **profils-lignes** sont égaux. De manière équivalente, on dit que les variables statistiques X et Y sont **non liées** si et seulement si :

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{il}}{n_{.l}},\ i = 1, \ldots, k,\ j = 1, \ldots, m,\ l = 1, \ldots, m.$$

Autrement dit les différentes **profils-colonnes** sont égaux.

Définition du χ^2. Revenons à l'état de non liaison entre deux variables X et Y. On montre que dans ce cas que les effectifs conjoints vérifient la relation :

$$n_{ij} = \frac{n_{i.}n_{.j}}{N},\ i = 1, \ldots, k,\ j = 1, \ldots, m, \tag{3.1}$$

Dans la pratique les effectifs conjoints vérifieront rarement la relation (3.1) mais peuvent être cependant tels que :

$$n_{ij} \simeq \frac{n_{i.}n_{.j}}{N},\ i = 1, \ldots, p,\ j = 1, \ldots, m,$$

ou de manière équivalente :

$$n_{ij} - \frac{n_{i.}n_{.j}}{N} \simeq 0,\ i = 1, \ldots, p,\ j = 1, \ldots, m. \tag{3.2}$$

Le χ^2 de contingence est un indice permettant de "mesurer" la liaison des variables X et Y, c'est-à-dire permettant d'évaluer "globalement" les différences dans (3.2). Définissons tout d'abord les effectifs conjoints théoriques t_{ij} lorsque les variables ne sont pas liées.

Effectifs conjoints théoriques : on appelle effectif conjoint théorique des modalités x_i et y_i l'effectif conjoint, noté t_{ij}, que l'on obtiendrait si les variables X et Y n'étaient pas liées :

$$t_{ij} = \frac{n_{i.}n_{.j}}{N}.$$

A partir des effectifs marginaux on construit une table de contingence théorique dans laquelle sont portés les effectifs théoriques.

Le χ^2 de contingence. On appelle **χ^2 de contingence** la quantité définie par :

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{N}\right)^2}{\frac{n_{i.}n_{.j}}{N}}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}.$$

Remarque 3.2. *1. Le χ^2 est une quantité positive ou nulle et si $\chi^2 = 0$ alors cela signifie que les variables X et Y ne sont pas liées (tous les termes de la somme sont nuls). En pratique on considèrera que X et Y ne sont pas liées lorsque la valeur du χ^2 est "proche de 0".*
2. La valeur du χ^2 dépend de la taille de l'échantillon N, du nombre de lignes k et du nombre de colonnes m. Pour N, k et m fixés, plus la valeur du χ^2 est grande et plus la liaison entre X et Y est forte et dans tous les cas on montre que la valeur maximale du χ^2 est $n(min(k, m) - 1)$.

Contribution de chaque case au χ^2 : quelle que soit la valeur du χ^2, il est souvent utile de savoir quelles sont les cases (les associations de modalités de X et de Y) qui ont le plus contribué au χ^2. On peut ainsi préciser, en cas de liaison entre X et Y, la nature de cette liaison. Si on considère la modalité i de X et la modalité j de Y (qui ont donc un effectif conjoint égal à n_{ij}), la **contribution relative** de cette case au χ^2 est par définition la quantité :

$$c_{ij} = \frac{100}{\chi^2}\frac{(n_{ij} - t_{ij})^2}{t_{ij}}.$$

Le C de Cramer : c'est la quantité définie par :

$$C \;=\; \sqrt{\frac{\chi^2}{N(\min(k,m)-1)}}.$$

Remarque 3.3. *Le C de Cramer vérifie la propriété :*

$$0 \;\leq\; C \;\leq\; 1$$

et la liaison entre les deux variables est d'autant plus forte que φ est proche de 1.

Exemple 3.3. *Reprenons l'exemple 3.1 pour lequel on obtient le tableau des effectifs théoriques ci-dessous :*

Y / *X*	*Engrais A*	*Engrais B*	*Engrais C*	*Totaux*
Ont fleuri	*40,476*	*68,810*	*60,714*	*170*
N'ont pas fleuri	*9,524*	*16,190*	*14,286*	*40*
Totaux	*50*	*85*	*75*	*210*

et les valeurs $\chi^2 = 7,232$ et $C = 0,19$. On voit qu'il y a un lien (modéré) entre l'engrais et la floraison. Dans ce cas, il faudrait pousser d'autres investigations pour se prononcer.
On peut obtenir le tableau et les valeurs ci-dessus à l'aide du logiciel R. On peut également calculer directement la valeur du χ^2 avec la fonction chisq.text :

```
> chisq.test(plante$floraison,plante$engrais)

        Pearson's Chi-squared test

data:  plante$floraison and plante$engrais
X-squared = 7.2316, df = 2, p-value = 0.0269
```

On obtient ainsi la valeur du χ^2 ainsi que d'autres valeurs relatives à un test statistique.

3.2 Deux variables quantitatives : la corrélation linéaire

Dans cette section on considère deux variables **quantitatives** X et Y définies sur une population Ω (ou un échantillon) de taille N. On dispose pour chaque individu ω_i de deux mesures x_i et y_i :

$$(x_1, y_1), \ldots, (x_N, y_N).$$

3.2.1 La covariance et le coefficient de corrélation linéaire

On note dans la suite $\overline{X}$ la moyenne de la variable X et $\overline{Y}$ celle de la variable Y :

$$\overline{X} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \overline{Y} \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Les variances de X et de Y sont données par :

$$\sigma_X^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2, \qquad \sigma_Y^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2.$$

Enfin, les écarts-type de X et Y sont notés σ_X et σ_Y.

La ovariance des variables X et Y est la quantité notée c_{XY} définie par :

$$c_{XY} \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right).$$

Remarque 3.4. *1. La covariance est un indice de dispersion conjointe entre les variables X et Y. Il s'agit d'une quantité qui peut être positive ou négative et on voit facilement que la covariance est* **symétrique** *c'est-à-dire telle que :*

$$c_{XY} \;=\; c_{YX}.$$

2. On a de manière évidente :

$$c_{XX} \;=\; \sigma_X^2,$$

la covariance entre X et elle-même est égale à la variance de X.

Autre expression de la covariance. On utilise le plus souvent une expression qui s'avère souvent plus commode à calculer. On montre en effet que la covariance entre X et Y s'écrit :

$$c_{XY} \;=\; \frac{1}{N}\sum_{i=1}^{N}x_iy_i-\overline{X}\,\overline{Y}.$$

Inégalité de Cauchy-Schwartz. On a l'inégalité suivante liant la covariance et les variances de X et Y (démonstration admise) :

$$c_{XY}^2 \;\leq\; \sigma_X^2\sigma_Y^2,$$

ou encore :

$$-\sigma_X\sigma_Y \;\leq\; c_{XY} \;\leq\; \sigma_X\sigma_Y.$$

Le coefficient de corrélation linéaire de X et Y est la quantité notée r_{XY} définie par :

$$r_{XY} \;=\; \frac{c_{XY}}{\sigma_X\sigma_Y}.$$

Remarque 3.5. *1. Le coefficient de corrélation linéaire est symétrique :*

$$r_{XY} \;=\; r_{YX}.$$

2. L'inégalité de Cauchy-Schwartz donne :

$$\boxed{-1 \;\leq\; r_{XY} \;\leq 1}$$

Signification. Examinons ce qui se passe si $r_{XY}=\pm 1$. On a alors :

$$c_{XY}^2 \;=\; \sigma_X^2\sigma_Y^2.$$

On montre alors que cette égalité n'est possible que si et seulement si les variables X et Y sont liés par une relation du type :

$$Y \;=\; aX+b,$$

où a et b sont deux réels quelconques. Cela signifie qu'il existe une **liaison linéaire** parfaite entre X et Y. De plus si $r_{XY}=1$ alors a est strictement positif et si $r_{XY}=-1$ alors a est strictement négatif. Par ailleurs, si $r_{XY}=0$ (ce qui signifie que $c_{XY}=0$) il n'existe aucune forme de liaison linéaire entre X et Y.
Enfin, en dehors de ces valeurs, $|r_{XY}|$ est d'autant plus proche de 1 que la liaison linéaire entre X et Y est grande.

Remarque 3.6. *Il est important de noter que r_{XY} mesure la liaison* **linéaire** *entre X et Y : on peut avoir $r_{XY} = 0$ et pourtant avoir une forte liaison entre X et Y (ces variables peuvent être liées par un autre type de liaison, par exemple quadratique,...).*

Exemple 3.4. *(Source : Bressoud, E. et Kahané, J.-C. Statistique descriptive. Applications avec Excel et calculatrices, PEARSON). Les données suivantes indiquent les indices du pouvoir d'achat du salaire minimum (variable sal_min : X) et du salaire moyen (variable sal_moy : Y) pour les salariés des secteurs privé et semi-public*

```
> salaire
```

```
   sal_min sal_moy
1      293     329
2      296     336
3      296     334
4      302     337
5      311     340
6      314     346
7      315     347
8      322     349
9      326     352
10     331     351
```

On obtient la covariance et le coefficient de corrélation linéaire :

```
> 9*cov(salaire$sal_min,salaire$sal_moy)/10
```

```
[1] 92.84
```

```
> cor(salaire$sal_min,salaire$sal_moy)
```

```
[1] 0.9657632
```

3.2.2 Régression linéaire entre deux variables

On suppose qu'il existe une "causalité" entre X et Y : X est la "cause" de Y. On se propose d'examiner le lien linéaire entre X et Y :

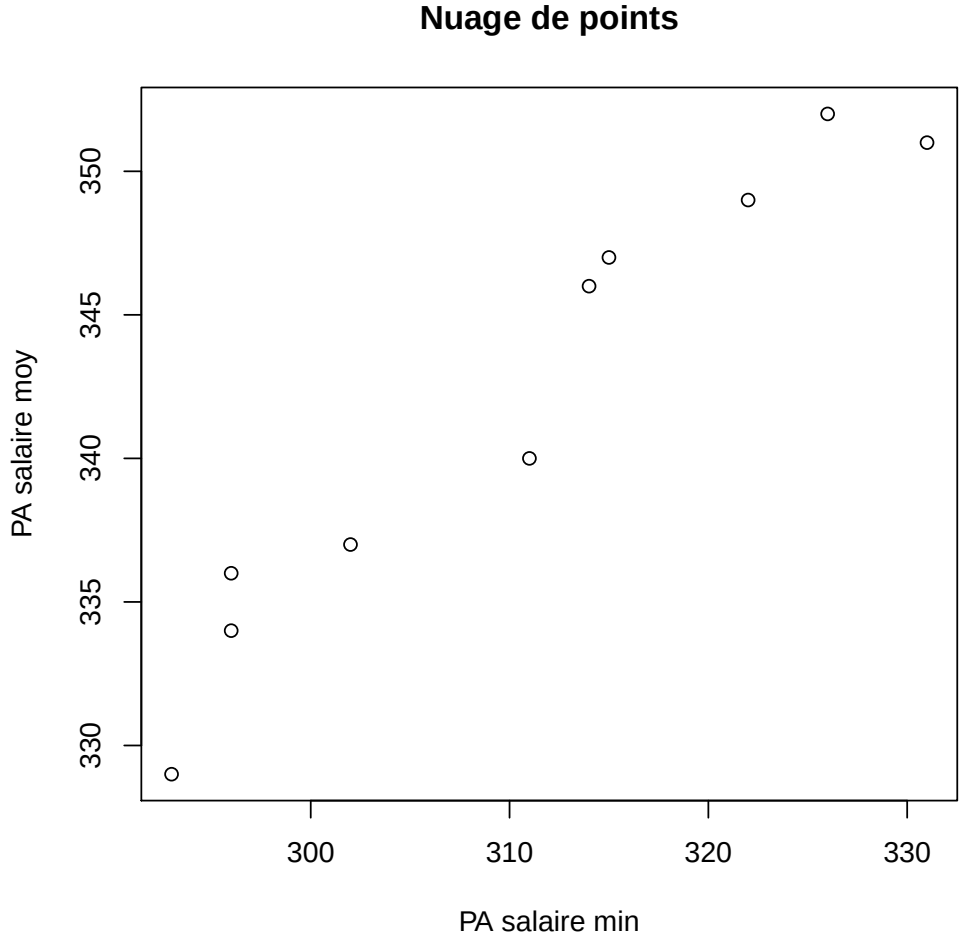
— Y a-t-il une liaison linéaire entre X et Y du type $Y \simeq aX + b$?

— Si oui, quelles sont les valeurs de a et de b ?

Le nuage de points : le nuage de points est la représentation dans un repère orthogonal des N points de coordonnées (x_i, y_i), $i = 1, \ldots, N$. L'axe des abscisses est l'axe des valeurs de la variable X et l'axe des ordonnées est l'axe des valeurs de la variable Y. Dans le cas d'une liaison linéaire forte entre X et Y, les points doivent être relativement alignés ou le nuage de points doit être "étiré". On voit ainsi l'importance de l'observation du nuage de points : on supposera l'existence d'une liaison linéaire entre X et Y lorsque celui-ci sera étiré (proche d'une droite) et que le coefficient de corrélation linéaire sera proche de 1.

Exemple 3.5. *le nuage de points relatif à l'exemple 3.4 est représenté dans la figure ci-dessous.*

```
> plot(salaire$sal_min,salaire$sal_moy,xlab="PA salaire min",ylab="PA salaire moy",main="Nuage
```



Lorsqu'une liaison linéaire forte entre deux variables X et Y a été mise à jour, on a alors une relation du type :

$$Y \simeq aX + b,$$

où les coefficients a et b sont inconnus. Le problème est donc de proposer des valeurs pour ces coefficients. Une méthode pour estimer les valeurs de a et b est la méthode des **moindres carrés** décrite ci-dessous.

La droite de régression - La méthode des moindres carrés : les coefficients $\widehat{a}$ et $\widehat{b}$ de la droite de régression obtenus par la méthode des moindres carrés sont les valeurs minimisant l'expression

$$\sum_{i=1}^{N} (y_i - ax_i - b)^2 , \tag{3.3}$$

expression appelée **critère des moindres carrés**.

On montre alors que les valeurs $\widehat{a}$ et $\widehat{b}$ minimisant le critère des moindres carrés (3.3) sont :

$$\widehat{a} = \frac{c_{XY}}{\sigma_X^2},$$

et :

$$\widehat{b} = \overline{Y} - \frac{c_{XY}}{\sigma_X^2}\overline{X} = \overline{Y} - \widehat{a}\overline{X}.$$

Remarque 3.7. *1. Les nombres $\widehat{a}$ et $\widehat{b}$ sont appelés* **coefficients de régression linéaire**.
2. La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée **droite de régression**.

3. Les nombres $e_i = y_i - (\widehat{a}x_i + \widehat{b})$, $i = 1, \ldots, N$ *sont appelés les* **résidus**. *On montre facilement que les résidus sont de moyenne nulle :*

$$\frac{1}{N}\sum_{i=1}^{N} e_i \;=\; 0.$$

Exemple 3.6. *Reprenons les données de l'exemple 3.4. On calcule les coefficients de régression à l'aide de la fonction lm :*

```
> lm(salaire$sal_moy~salaire$sal_min)

Call:
lm(formula = salaire$sal_moy ~ salaire$sal_min)

Coefficients:
    (Intercept)  salaire$sal_min
       164.5815           0.5715
```

On obtient ainsi : $\widehat{a} = 0,5715$ *et* $\widehat{b} = 164,5815$. *La droite de régression est donc la droite d'équation :* $y = 0,5715x + 164,5815$.

Prévision. Une des utilisations de la droite de régression est la **prévision**. Ayant observé pour un individu $N+1$ de la population P (n'appartenant pas à l'échantillon E) la valeur x_{N+1} de la variable X, peut-on donner une estimation $\widehat{y}_{N+1}$ de Y pour cet individu ? Cette prévision d'une valeur de Y est facile en utilisant la droite de régression : la vraie valeur (inconnue) y_{N+1} est, lorsque le modèle de régression est "bon", proche de $\widehat{a}x_{N+1} + \widehat{b}$. Une estimation de la valeur y_{N+1} est ainsi donnée par :

$$\widehat{y}_{N+1} \;=\; \widehat{a}x_{N+1} + \widehat{b}.$$

3.3 Une variable quantitative et une variable qualitative : le rapport de corrélation

3.3.1 Les données

Nous considérons maintenant une variable qualitative X et une variable quantitative Y observées sur Ω (ou un échantillon) de taille N.
Soit k le nombre de modalités de X et $x_1, \ldots, x_k$ ses modalités et soit n_j l'effectif de la modalité x_j, $j = 1, \ldots, k$.
Chaque modalité X_j de la variable X définit une classe C_j d'individus, c'est-à-dire les individus ayant pris la modalité x_j.

La notion de liaison entre les variables X et Y est basée sur la comparaison des valeurs de la variable Y sur chaque classe C_j. En effet, l'idée est que si la variable X n'a pas d'influence sur la variable Y, alors les valeurs de Y ne doivent pas différer de manière trop sensible d'une classe à l'autre. Prenons par exemple 3 établissements scolaires (variable X : établissement d'origine à 3 modalités), et les résultats obtenus au baccalauréat ES par les élèves de ces établissements (variable Y : note globale obtenue au baccalauréat ES). L'analyse de l'influence de l'établissement sur les résultats au baccalauréat ES repose ainsi sur la comparaison des notes globales des 3 établissements.
Nous détaillons ci-dessous les 3 étapes de l'analyse de la liaison entre X et Y :
- la comparaison des moyennes de Y sur chaque classe (première indication) ;
- la représentation en parallèles des boîtes à moustaches de Y sur chaque classe ;
- le calcul d'un indice de liaison, le **rapport de corrélation**, basée sur la décomposition de la variance de Y.

Exemple 3.7. *On étudie l'influence de 4 types de régime alimentaire (variable X) sur le temps de coagulation du sang. Pour chaque type de régime, on a mesuré sur plusieurs individus le temps de coagulation (variable Y) :*

Modalités de X			
Régime 1	*Régime 2*	*Régime 3*	*Régime 4*
62	*63*	*68*	*56*
60	*67*	*66*	*62*
63	*71*	*71*	*60*
59	*64*	*67*	*61*
	65	*68*	*63*
	66	*69*	*64*
			63
			59

Dans le tableau ci-dessus, les valeurs de Y sont "empilées" sur 4 colonnes correspondant à chaque classe. La variable qualitative X, "type de régime" a $k = 4$ modalités. On a : $n_1 = 4,\ n_2 = n_3 = 6, n_4 = 8$ et $N = 24$.

Sur chaque classe C_j, $j = 1, \ldots, k$, on peut calculer la moyenne $\overline{Y}_j$ et la variance partielle σ_j^2 de la variable quantitative Y données par :

$$\overline{Y}_j \ = \ \frac{1}{n_j} \sum_{\omega_i \in C_j} Y_i,$$

$$\sigma_j^2 \ = \ \frac{1}{n_j} \sum_{\omega_i \in C_j} (Y_i - \overline{Y}_j)^2.$$

Dans les formules ci-dessus, les sommes portent sur les valeurs Y_i obtenues pour les individus ω_i appartenant à la classe C_j. Ainsi pour la classe C_1 dans l'exemple 5.12, la moyenne vaut :

$$\overline{Y}_1 \ = \ \frac{62 + 60 + 63 + 59}{4} \ = \ 61.$$

On note par ailleurs $\overline{Y}$ et σ_Y^2 la moyenne et respectivement la variance de la variable Y sur les N individus :

$$\overline{Y} \ = \ \frac{1}{N} \sum_{i=1}^{N} Y_i, \qquad \sigma_Y^2 \ = \ \frac{1}{N} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Pour l'exemple 3.7, on obtient ainsi :

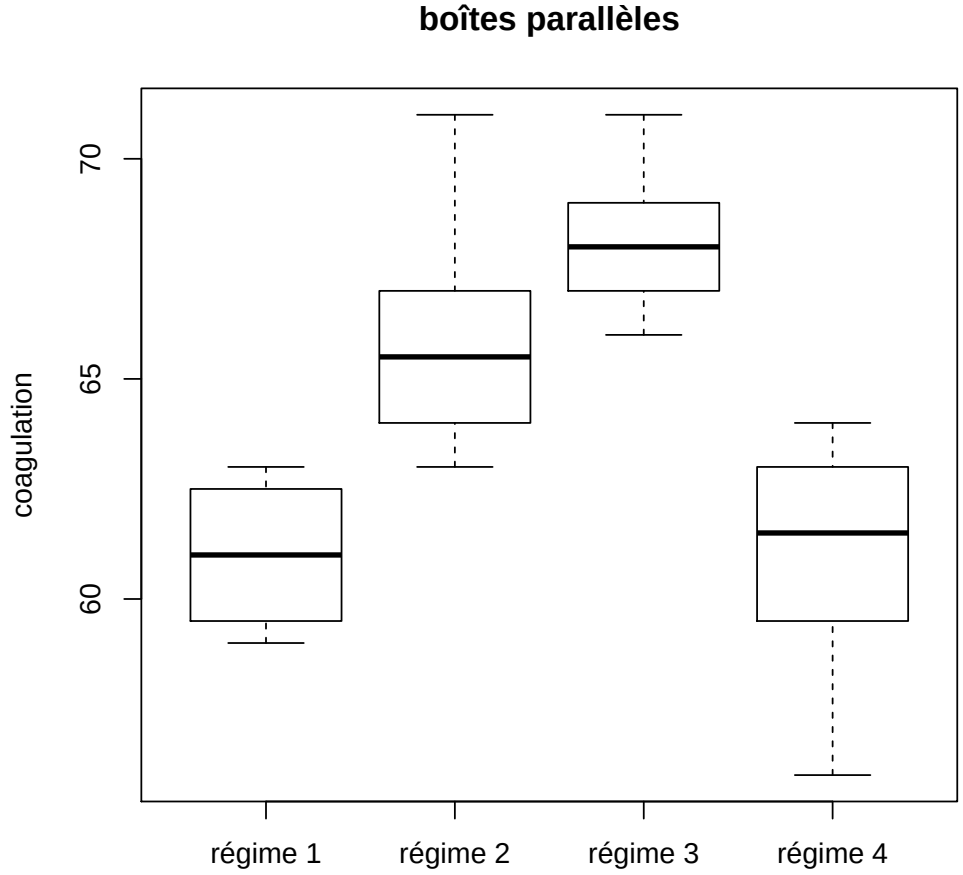
	C_1	C_2	C_3	C_4	Popu$^{\text{on}}$ totale
Moyennes	61	66	68,17	61	64
Variances	2,5	6,67	2,47	6	14,54

Les différences entre les moyennes partielles donnent une première indication sur la liaison entre les variables X et Y : si celles-ci ne sont pas liées, en toute logique les moyennes de la variable Y sur les classes définies par les modalités de la variable X devraient être proches. *A contrario*, des différences sensibles suggèrent une possible liaison entre X et Y.

3.3.2 Représentation graphique

On représente les données à l'aide de boîtes à moustaches parallèles. Pour chaque modalité x_j de la variable X (sur l'axe horizontal), on trace une boîte à moustaches obtenue en calculant les quartiles de la variable Y sur la classe C_j. Ce graphique apporte une précision sur la liaison entre les variables X et Y. Comme pour la comparaison des moyennes vue à la section précédente, si les variables X et Y ne sont pas liées les boîtes parallèles devraient être sensiblement identiques.

Exemple 3.8. *Reprenons l'exemple 3.7 pour lequel les boîtes parallèles ont été tracées dans la figure ci-dessous. Le tableau des moyennes partielles (section précédente) indique des différences assez sensibles : moyennes plus élevées pour les classes C_2 et C_3, ce qui suggèrerait une liaison entre les variables X et Y. Le graphique des boîtes parallèles confirme cette analyse puisque on observe également des différences notables : différences entre les quartiles des 4 classes, la variable Y est plus dispersée sur les classes C_2 et C_3.*



3.3.3 Le rapport de corrélation

Outre la comparaison des moyennes et des boîtes parallèles qui fournissent une première analyse de la liaison entre X et Y, on définit un indice de liaison : le **rapport de corrélation**. Celui-ci repose sur une **décomposition** de la variance σ_Y^2. Notons tout d'abord que moyenne totale et moyennes partielles sont liées par la relation :

$$\overline{Y} \;=\; \frac{1}{N}\sum_{j=1}^{k} n_j \overline{Y_j}.$$

Remarque 3.8. *1. Si on "résume" les données obtenues sur chaque classe C_j par leur moyenne $\overline{Y}_j$, on obtient ainsi une nouvelle variable ayant k modalités $\overline{Y}_1, \ldots, \overline{Y}_k$ avec des effectifs respectifs égaux à $n_1, \ldots, n_j$. La formule ci-dessus montre que la moyenne de cette nouvelle variable est égale à la moyenne totale de la variable Y.*
2. Dans le cas général où les effectifs des différentes classes C_j ne sont pas identiques, la moyenne totale **n'est pas égale** *à la moyenne des moyennes partielles : chaque moyenne partielle doit être multipliée par l'effectif n_j de la classe.*

Dans le cas d'effectifs identiques sur chaque classe, il est facile de montrer que la moyenne totale est égale à la moyenne des moyennes partielles. En effet, dans ce cas on a $n_j = \frac{N}{k}$. Lorsque les effectifs n_j sont égaux, on dit qu'on a réalisé un **plan équilibré**.

Décomposition de la variance. On montre que la variance totale σ_Y^2 (calculée sur l'ensemble des N individus) se décompose de la manière suivante :

$$\sigma_Y^2 \;=\; \frac{1}{N}\sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2 + \frac{1}{N}\sum_{j=1}^{k} n_j\sigma_j^2.$$

Interprétation

Variance expliquée par la partition ou **variance inter-classes**. C'est le premier terme de la décomposition de la variance ci-dessus. On le note σ_E^2 $(= \frac{1}{N}\sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2)$. Analysons sa signification. Il correspond à la variance d'une variable quantitative ayant k modalités $\overline{Y}_1, \ldots, \overline{Y}_k$ avec des effectifs respectifs égaux à $n_1, \ldots, n_j$ (voir le point 1. de la remarque 5.9). Dans le cas où les variables X et Y ne sont pas liées (pas d'influence des modalités de X sur la variable Y), la variance expliquée par les modalités de X devrait être très proche de 0.

Variance résiduelle ou **variance intra-classes**. C'est le second terme de la décomposition de la variance ci-dessus. On le note σ_R^2 $(= \frac{1}{N}\sum_{j=1}^{k} n_j\sigma_j^2)$. Ce terme mesure la variation à l'intérieur des classes.

En résumé, la variance totale est la somme d'une variance entre les classes et de la variance à l'intérieur des classes. Selon le principe énoncé plus haut, plus le premier terme σ_E^2 est grand comparativement au second terme σ_R^2 plus les variables X et Y sont liées.

Le rapport de corrélation

Le rapport de corrélation, noté $C_{Y|X}$, est défini par :

$$C_{Y|X} \;=\; \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}}.$$

Propriétés et interprétation

- Le coefficient $C_{Y|X}$ n'est pas symétrique.
- $0 \leq C_{Y|X} \leq 1$. Cette propriété est évidente à partir de la décomposition de la variance.
- Si $C_{Y|X} = 0$, cela signifie que $\sigma_E^2 = 0$: la variance entre les différentes classes étant nulle cela signifie que les moyennes $\overline{Y}_j$ sont égales et donc qu'il n'y a pas de lien entre X et Y.
- Si $C_{Y|X} = 1$, cela signifie que $\sigma_R^2 = 0$: d'après la définition de σ_R^2, la variable Y est constante sur chaque classe C_j (sa variance est nulle sur chacune de ces classes). Dans ce cas, la connaissance de X (c'est-à-dire de la classe à laquelle appartient chaque individu) est suffisante pour connaître Y. Il y a liaison totale entre X et Y.

Pour l'exemple 3.7, on trouve $\sigma_E^2 \simeq 9,85$ et $C_{Y|X} \;=\; \sqrt{\frac{9,85}{14,54}} \simeq 0,82$, ce qui indique une liaison forte entre X et Y.