Statistique descriptive

Pascal Sarda

Licence Mathématiques Informatique Appliquées aux Sciences Humaines et Sociales

Définitions et terminologie

Statistique

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.

Recueil des données : s'effectue lors d'une enquête.

- recensement : porte sur tous les sujets concernés par l'étude.

- sondage : enquête partielle portant sur une partie des sujets (cf. **théorie des sondages**).

Traitement des données : deux branches principales en statistique :

a. La statistique descriptive : 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 ; d'autre part, le résumé des données par des **caractéristiques numériques**.

b. La statistique inférentielle : a pour objectif de généraliser (inférer) à une population les résultats observés sur un échantillon. On suppose ainsi que le phénomène étudié peut être décrit par un **modèle mathématique** permettant d'approcher les propriétés de ce phénomène. Les méthodes utilisées en statistique mathématique font appel au **calcul des probabilités**.

Population

ensemble des éléments sur lesquels porte l'étude, notée Ω

Individu (ou unité statistique)

tout élément de la population.
Notation : ω pour un individu générique de la population Ω.
Mais, le plus souvent, les individus sont repérés par un identifiant $1, 2, \ldots$

Échantillon

partie de la population

NB. L'extrapolation des résultats obtenus sur l'échantillon à la population entière relève de la **statistique inférentielle**.

Taille de la population (ou de l'échantillon)

nombre d'individus de la population (ou de l'échantillon)
Notation : N.

Chaque individu de la population Ω est décrit par une ou plusieurs variables statistiques, appelées aussi caractères.

Variable statistique (ou caractère)

application, notée X, définie sur Ω et à valeurs dans un ensemble F : à tout individu ω on fait correspondre une valeur $X(\omega)$ de F.

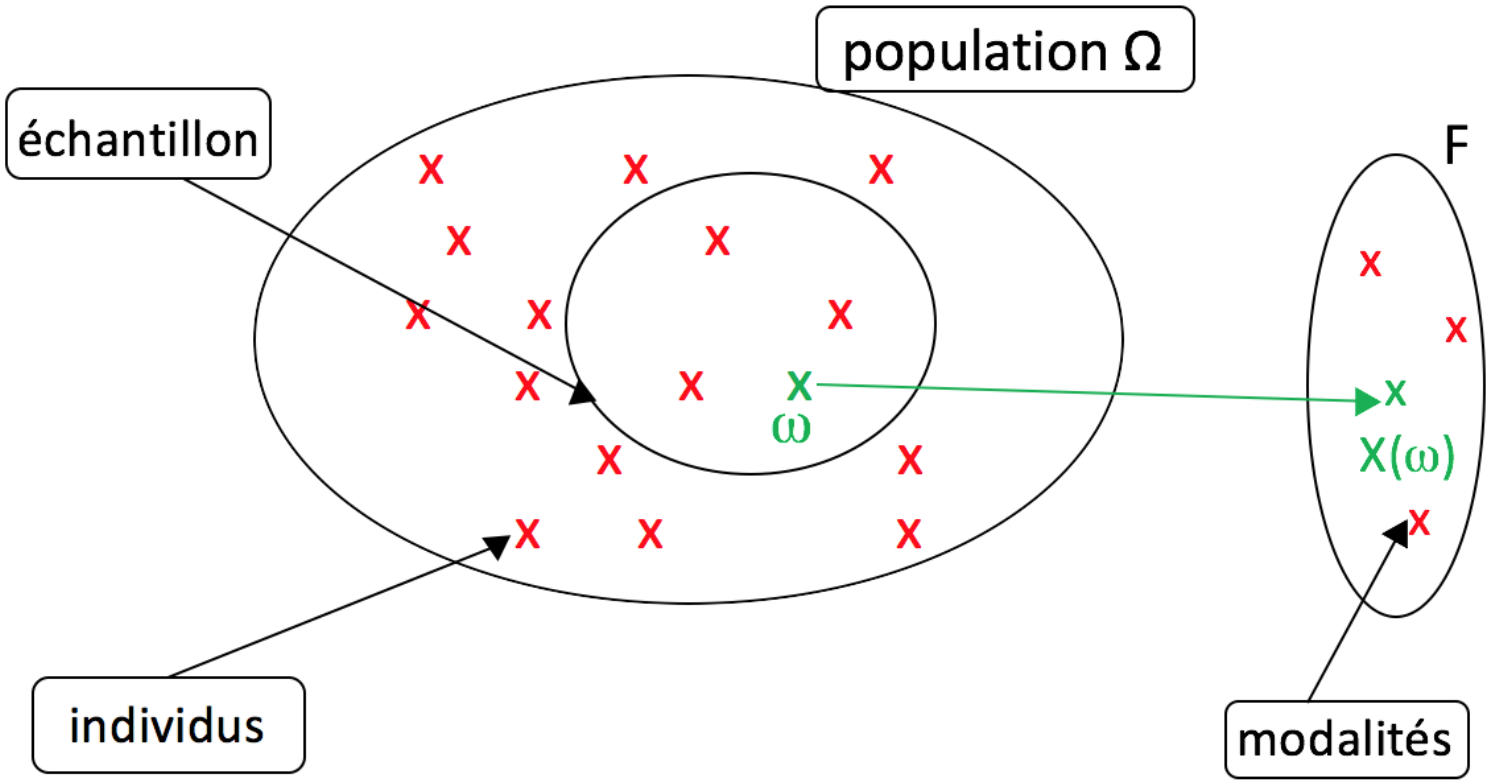
$$\begin{array}{rccc} X : & \Omega & \longrightarrow & F \\ & \omega & \longmapsto & X(\omega). \end{array}$$

Exemple. cf. feuille.

Modalité

toute valeur possible de la variable statistique X.
k étant le nombre de modalités distinctes, on note $x_1, \ldots, x_k$, ces modalités.

Remarquons que $k \leq N$.


population Ω
échantillon
F
ω
X(ω)
individus
modalités

Variables quantitatives

variables numériques telles que F est un ensemble de nombres (F est un sous-ensemble de $\mathbb{R}$, l'ensemble des réels)

Par exemple : l'âge, la taille ; le nombre d'enfants ; le nombre de salariés d'une entreprise, ...

Variables quantitatives discrètes : variables numériques dont l'ensemble des valeurs possibles est un ensemble fini ou dénombrable et associées à un comptage.

Par exemple le nombre d'enfants par ménage.

$\longrightarrow$ Le plus souvent, ces valeurs sont des entiers et le nombre de valeurs distinctes est assez faible.

$\longrightarrow$ On note $x_1, x_2, \ldots, x_k$ les valeurs distinctes de la variable, k étant le nombre de valeurs distinctes avec $k < N$; es valeurs seront toujours ordonnées suivant l'ordre croissant : $x_1 < x_2 < \ldots < x_k$.

Variables quantitatives discrètes - tableaux statistiques

- Les données brutes recueillies à l'issue d'une enquête sont organisées sous la forme de tableaux statistiques qui seront ensuite utilisés pour obtenir des graphiques ou encore extraire des caractéristiques numériques.
- 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

l'effectif d'une modalité x_j est nombre d'individus observés ayant pris cette modalité de la variable
Notation : $n_1, \ldots, n_k$.
La fréquence d'une modalité x_j est la proportion d'individus de l'échantillon ayant pris cette modalité, c'est-à-dire l'effectif divisé par la taille de l'échantillon.
Notation : $f_1, \ldots, f_k$.

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

effectif cumulé et fréquence cumulée

l'effectif cumulé de la modalité x_j est le nombre d'individus ayant pris cette modalité ou une modalité inférieure.
Notation : $N_1, \ldots, N_k$.
La fréquence cumulée de la modalité x_j est la proportion d'individus ayant pris cette modalité ou une modalité inférieure ou une valeur se trouvant dans cette classe ou dans une classe plus petite.
Notation : $F_1, \ldots, F_k$.

$$F_j = \frac{N_j}{N}$$

Exemple. (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

Avec R

```
> 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
```

a. Les diagrammes en bâtons

servent à représenter les effectifs ou les fréquences de l'ensemble des modalités d'une variable quantitative discrète.

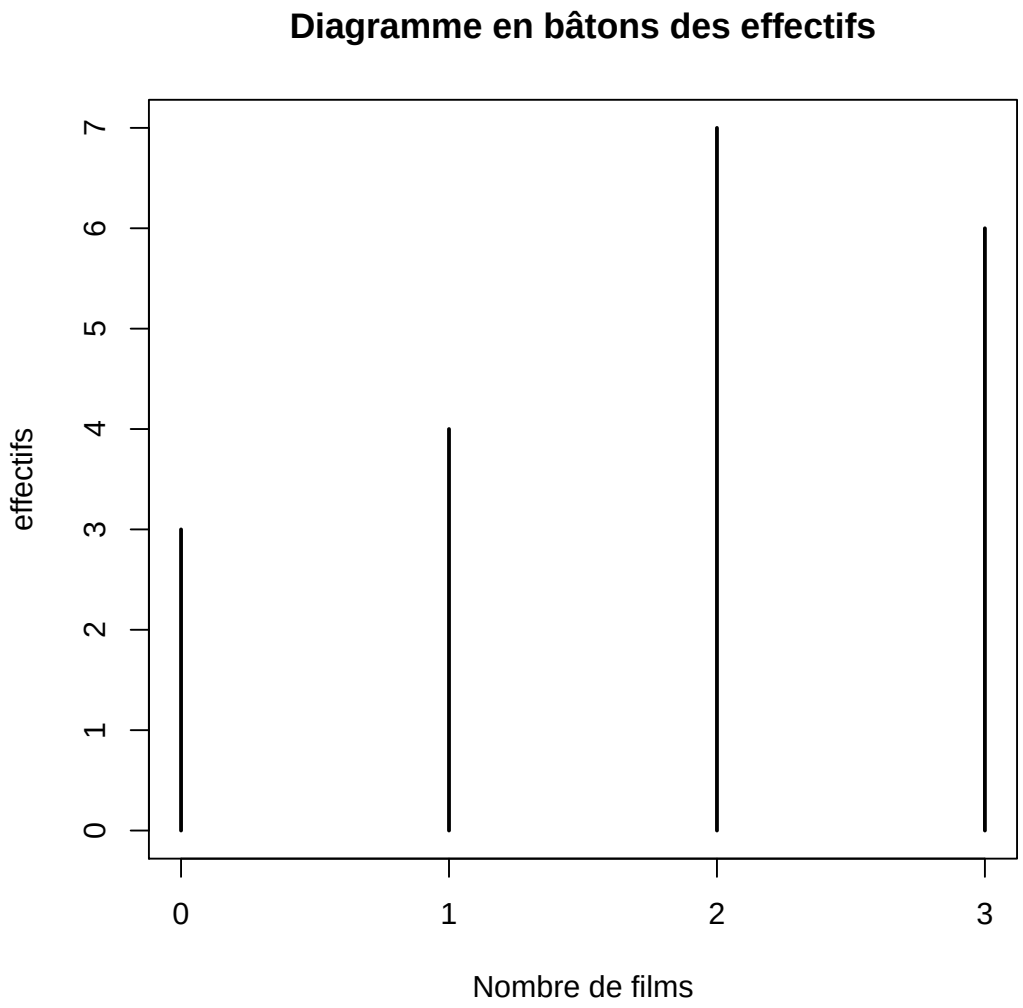
On choisit tout d'abord deux axes perpendiculaires et une échelle pour chacun de ces axes :
axe des abscisses (axe horizontal) sert à porter les modalités de la variable
axe des ordonnées (axe vertical) : axe des effectifs ou des fréquences suivant le cas
On trace en chaque modalité un trait vertical (bâton) dont la hauteur correspond à la valeur de l'effectif ou de la fréquence.

Remarque

1. Respecter l'échelle choisie pour l'axe des abscisses ($\neq$ du diagramme en colonnes).
2. Les diagrammes en bâtons des effectifs et des fréquences d'une même variable diffèrent par l'échelle des ordonnées (facteur N).

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

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



b. Les diagrammes cumulatifs

graphes de fonctions en escaliers servant à représenter les effectifs ou fréquences cumulé(e)s d'une variable quantitative discrète.
Graphe de la fonction qui à toute valeur réelle fait correspondre le nombre (ou la proportion) d'individus ayant pris une modalité inférieure ou égale à cette valeur.

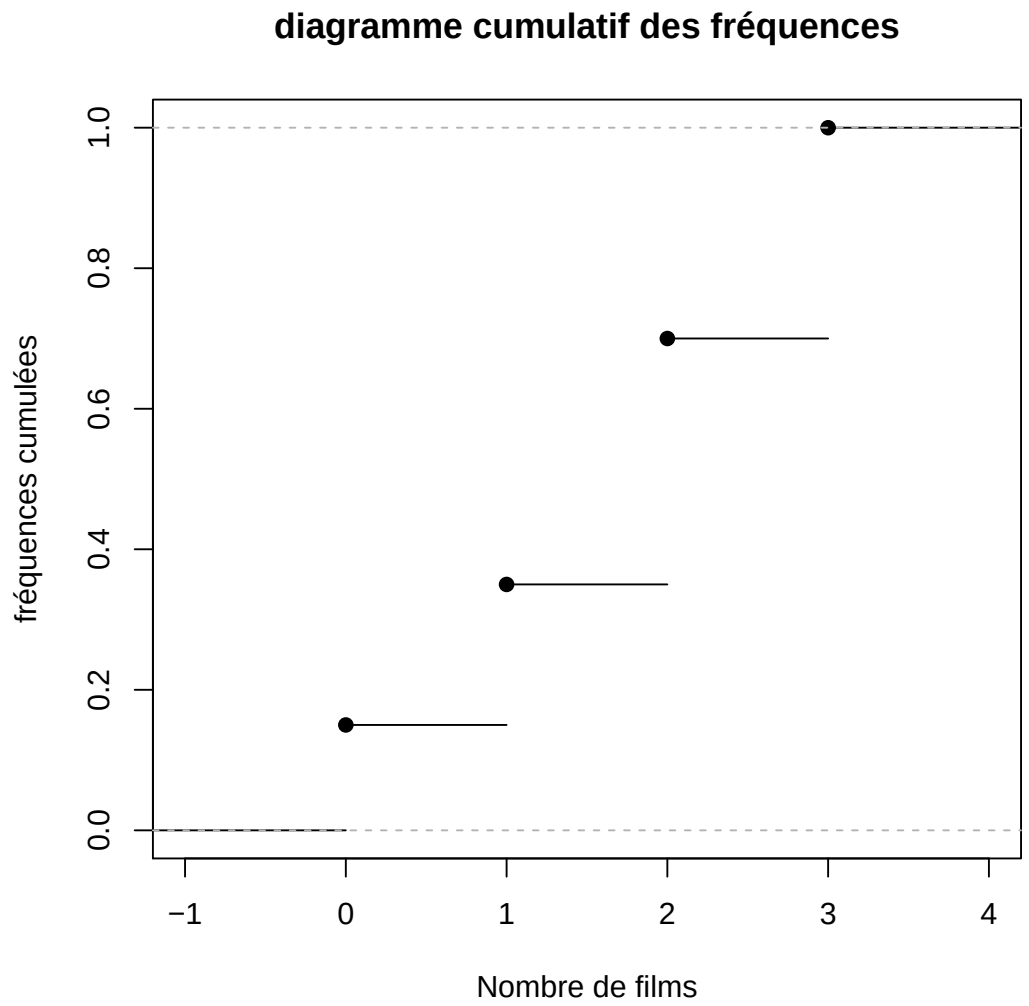
On choisit deux axes gradués perpendiculaires :

- l'axe des abscisses est identique à celui du diagramme en bâtons et est donc l'axe des modalités
- l'axe des ordonnées est l'axe des effectifs cumulés, gradué de 0 à N, ou des fréquences cumulées, gradué de 0 à 1.

Le diagramme cumulatif est le diagramme d'une fonction en escaliers :

- nulle pour $x < x_1$ (x_1 plus petite modalité)
- égale à l'effectif cumulé N_j (ou à la fréquence cumulée F_j) pour $x_j \leq x < x_{j+1}, j = 1, 2, \ldots, k-1$
- égale à N (ou 1) pour $x \geq x_k$ (x_k plus grande modalité)

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



1. Caractéristiques de tendance centrale

a. La moyenne arithmétique

c'est la quantité, notée $\overline{X}$, égale à la somme des produits effectifs$\times$modalités (ou effectifs$\times$centres) divisée par la taille N :

$$\overline{X} = \frac{n_1 x_1 + \ldots + n_k x_k}{N}$$

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

comme valeur de la médiane.

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

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 $\boxed{m = 4}$

comme valeur de la médiane.

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

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 :

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

comme valeur de la médiane.

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

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, par convention, comme valeur de la médiane le centre de cet intervalle : $m = 4,5$

comme valeur de la médiane.

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

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, par convention, comme valeur de la médiane le centre de cet intervalle : $m = 4,5$

3. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

comme valeur de la médiane.

b. 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.

Cas d'une variable quantitative discrète

1. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

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, par convention, comme valeur de la médiane le centre de cet intervalle : $m = 4,5$

3. Sur un échantillon de 5 individus, on a observé les valeurs suivantes :

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

Il y a 4 observations inférieures à 4 et 3 observations supérieures à 4. Par convention, on prend $m = 4$ comme valeur de la médiane.

Règle de calcul : on range les modalités par ordre croissant et on calcule les effectifs cumulés.
- Si N est impair, la médiane est la valeur correspondant au premier effectif cumulé supérieur ou égal à $\frac{N+1}{2}$;
- Si N est pair, et si l'effectif cumulé $\frac{N}{2}$ apparait dans la tableau, la médiane est le centre entre la modalité correspondant à cet effectif cumulé et la modalité suivante ; si l'effectif cumulé $\frac{N}{2}$ n'apparait pas dans le tableau, la médiane est la modalité correspondant au premier effectif cumulé supérieur ou égal à $\frac{N+2}{2}$.

c. Le mode

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

2. Caractéristiques de dispersion

Les caractéristiques de tendance centrale ne suffisent pas à elles seules à résumer une série statistique :
Par exemple, les deux séries de valeurs 2, 5, 15, 18 et 9, 10, 10, 11 ont la même moyenne égale à 10.
Mais la première série est beaucoup plus “dispersée” que la seconde.

a. L’étendue

c’est la différence entre la plus grande et la plus petite observation.

b. L'intervalle interquartile

Quartiles

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.

Intervalle interquartile

c'est l'intervalle $[q_1, q_3]$.

c. La variance

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

$$\sigma_X^2 = \frac{n_1\left(x_1-\overline{X}\right)^2+\ldots+n_k\left(x_k-\overline{X}\right)^2}{N}$$

ou

$$\sigma_X^2 = \frac{n_1x_1^2+\ldots+n_kx_k^2}{N} - \overline{X}^2.$$

NB. On utilise le plus souvent la seconde formule pour calculer la variance, c'est-à-dire la formule : "moyenne des carrés - carré de la moyenne".

d. L'écart-type

c'est la quantité, notée σ_X, égale à la racine carrée de sa variance :

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

Avec le logiciel R on obtient un résumé des caractéristiques numériques la variable en tapant :

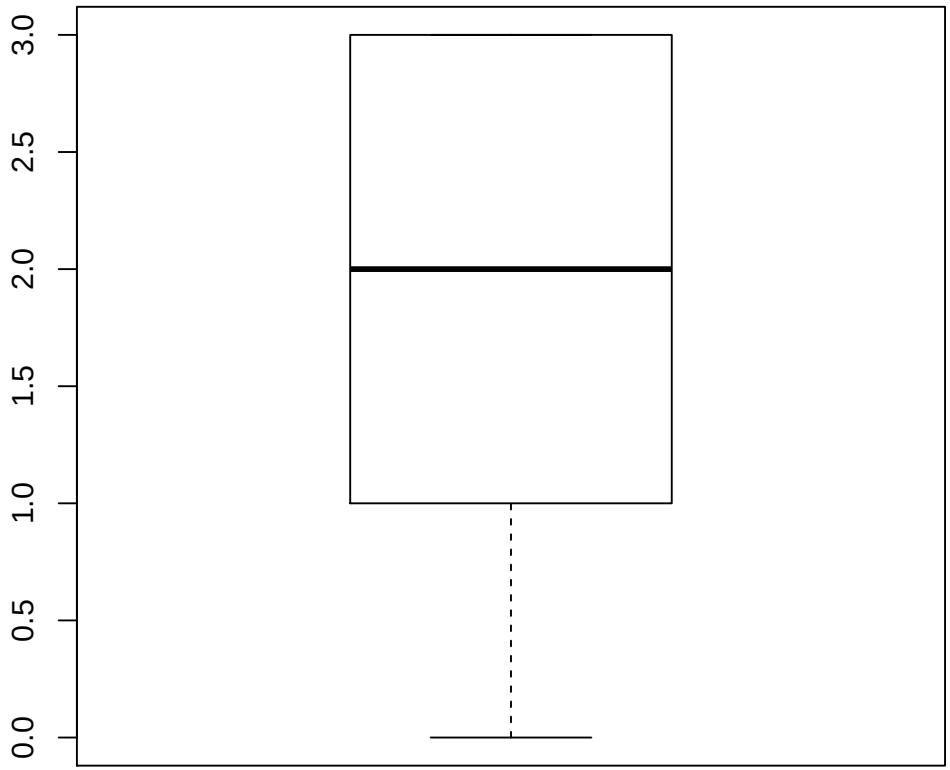
```
> summary(films1)

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
```

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

```
> boxplot(films1)
```



Variable quantitative continue - Regroupement en classes

Pour les variables quantitatives continues, le nombre de modalités distinctes sera assez important. On est amené à regrouper ces valeurs en classes, c'est-à-dire en intervalles de valeurs.
On note par k le nombre de classes et par $[b_1, b_2[, \ldots, [b_k, b_{k+1}[$, les différentes classes.
Par convention, l'intervalle définissant une classe est fermé à gauche et ouvert à droite, de sorte que les classes sont disjointes.

Bornes

valeurs minimales et maximales de cette classe : b_j et b_{j+1}.

Amplitude

largeur de la classe, notée a_j. On a ainsi : $\boxed{a_j = b_{j+1} - b_j}$

Centre

valeur équidistante des deux bornes, noté x_j : $\boxed{x_j = \dfrac{b_j + b_{j+1}}{2}}$

Le tableau statistique est est analogue à celui d'une variable quantitative discrète : dans la première colonne sont portées les classes rangées par ordre croissant puis les effectifs, effectifs cumulés, fréquences et fréquences cumulées définis de la même manière que pour une variable quantitative discrète.

Exemple (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

a. Les histogrammes

servent à représenter les effectifs ou les fréquences d'une variable quantitative continue.

On choisit deux axes perpendiculaires.

- axe des abscisses (axe horizontal) : classes de la variable.
- axe des ordonnées : densités d'effectifs (ou de fréquences)

  densité d'effectif de la $j^{\text{ème}}$ classe : effectif divisé par l'amplitude

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

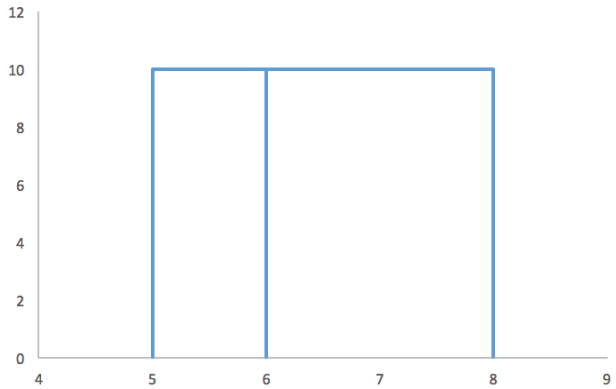
  densité de fréquence : fréquence divisée par l'amplitude

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

En chaque classe, on trace un rectangle dont la hauteur est égale à la densité d'effectif ou de fréquence.

1. Utilisation des effectifs (bruts).

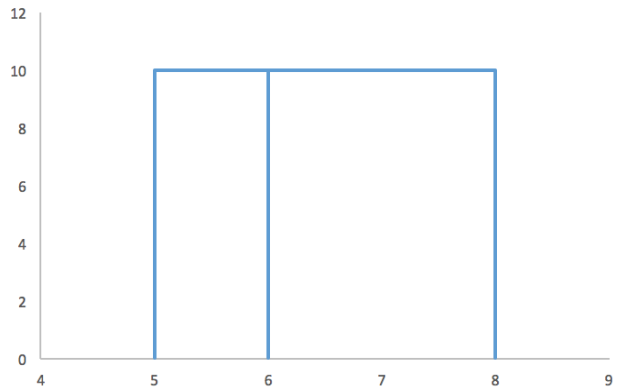
Classe	Amplitude	Effectif
$[5, 6[$	1	10
$[6, 8[$	2	10



Pourquoi utiliser les densités d'effectifs (ou de fréquences)

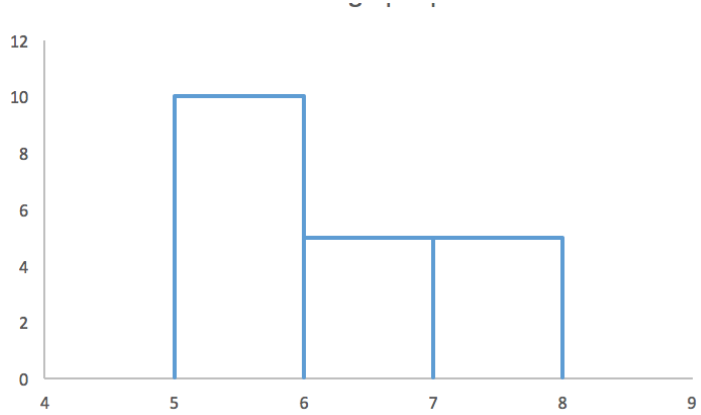
1. Utilisation des effectifs (bruts).

Classe	Amplitude	Effectif
$[5, 6[$	1	10
$[6, 8[$	2	10



2. Autre découpage en classes

Classe	Amplitude	Effectif
$[5, 6[$	1	10
$[6, 7[$	1	5
$[7, 8[$	1	5

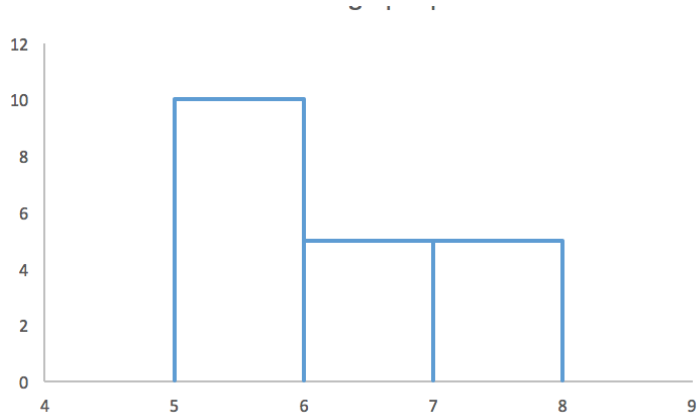


1. Utilisation des effectifs (bruts).

Classe	Amplitude	Effectif
$[5, 6[$	1	10
$[6, 8[$	2	10

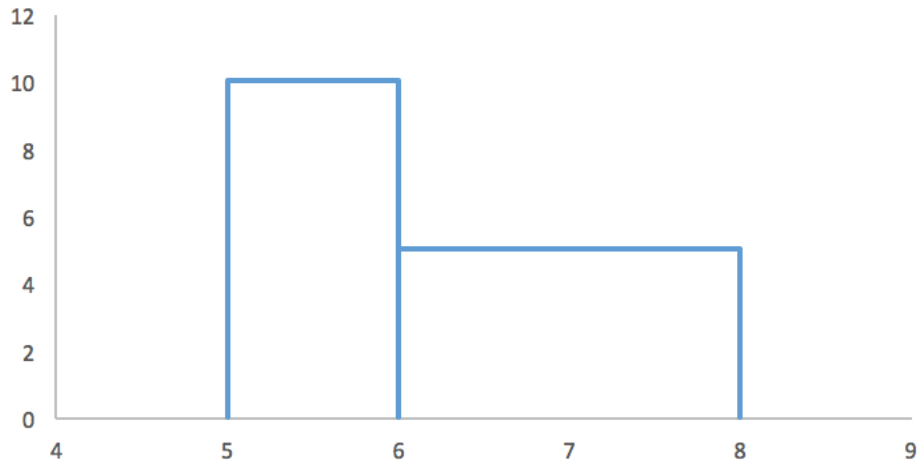
2. Autre découpage en classes

Classe	Amplitude	Effectif
$[5, 6[$	1	10
$[6, 7[$	1	5
$[7, 8[$	1	5



Il faut diviser par l'amplitude !! Pour le premier découpage, 2 pour la seconde classe

Classe	Amplitude	Effectif	Densité
$[5, 6[$	1	10	10
$[6, 8[$	2	10	5



Remarque

1. L'**aire** de chaque rectangle est égale à l'effectif ou à la fréquence de la classe qu'il représente.
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 :

$$n_1 + \ldots + n_k = N \qquad f_1 + \ldots + f_k = 1$$

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 (*cf.* ci-dessous).

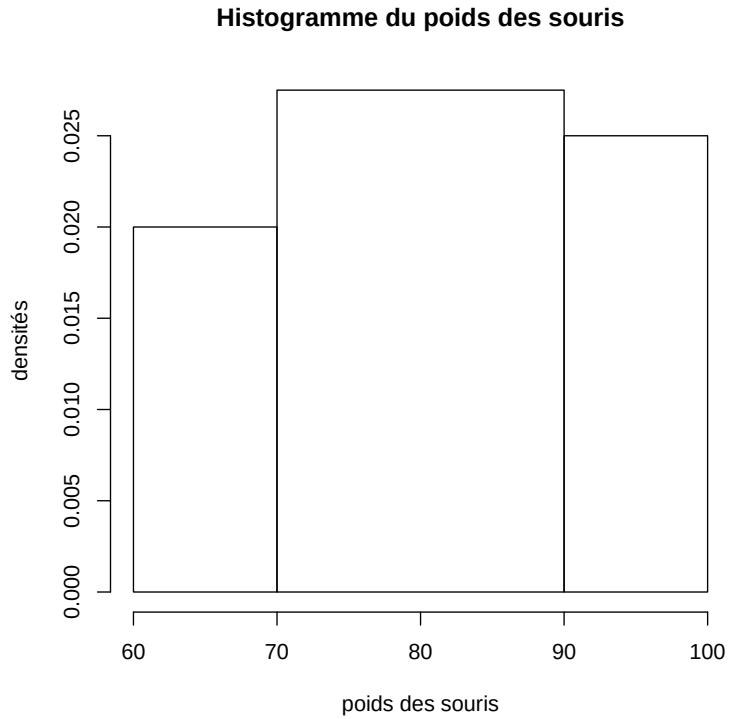
deux découpages pour les mêmes données

On reprend les données précédentes 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 des souris")
```

Pour le second découpage :

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



b. Les polygones cumulatifs

graphes 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.

Principe

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 (les observations sont uniformément réparties). Cette hypothèse implique que la fonction augmente de manière constante dans chaque classe et est donc une fonction affine par morceaux (ligne brisée).

Variable taille (classes d'amplitude constante 10cm) :
axe des abscisses : classe ; axe des ordonnées : effectifs ou fréquences cumulé(e)s.

- 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 ;

Variable taille (classes d'amplitude constante 10cm) :
axe des abscisses : classe ; axe des ordonnées : effectifs ou fréquences cumulé(e)s.

- 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 ;

Variable taille (classes d'amplitude constante 10cm) :
axe des abscisses : classe ; axe des ordonnées : effectifs ou fréquences cumulé(e)s.

- 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 $[b_j, b_{j+1}[$;

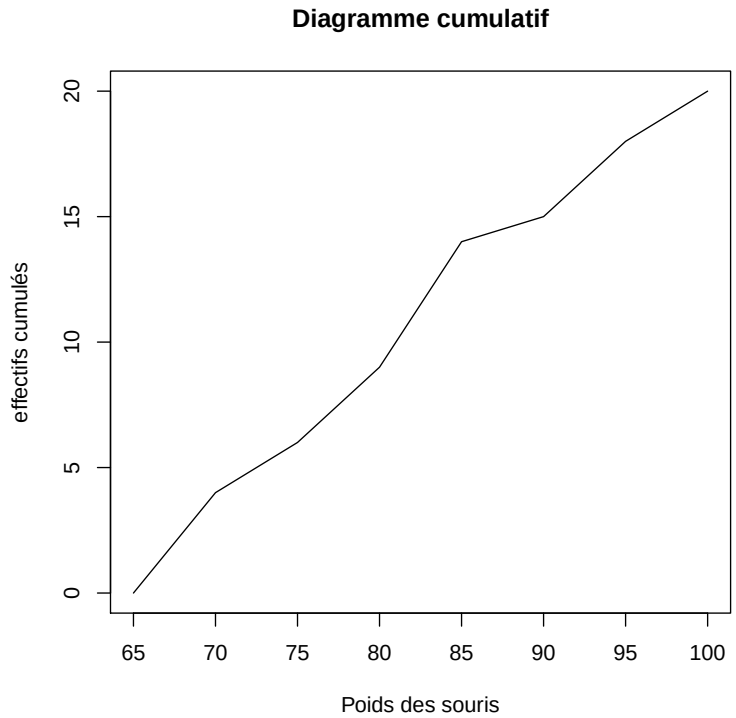
Variable taille (classes d'amplitude constante 10cm) :
axe des abscisses : classe ; axe des ordonnées : effectifs ou fréquences cumulé(e)s.

- 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.

Traçons le polygone cumulatif pour les données de l'exemple ci-dessus 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).



Variables quantitatives continues - Caractéristiques numériques

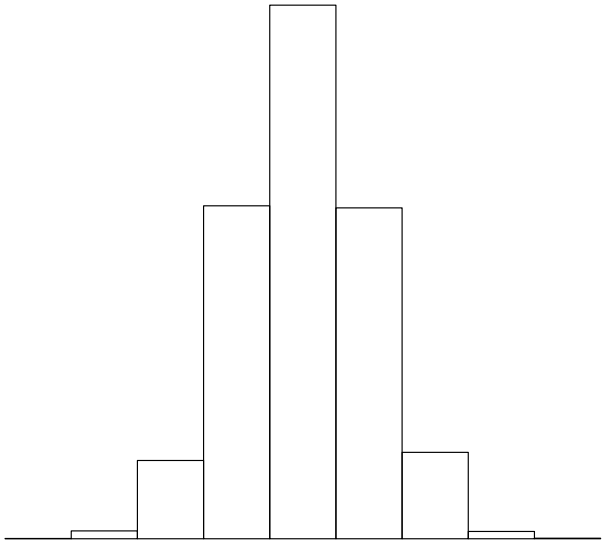
- La moyenne, la variance et l'écart-type d'une variable quantitative continue sont définis de la même manière que pour une variable quantitative discrète en prenant pour x_j les centres de classes.
- La médiane. Dans le cas d'une variable quantitative continue, la médiane est la valeur de la série correspondant à la fréquence cumulée 0,5. Elle se lit donc directement sur le polygone cumulatif (des fréquences).
  Calcul : 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 :

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

  Les **quartiles** sont calculés sur 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.

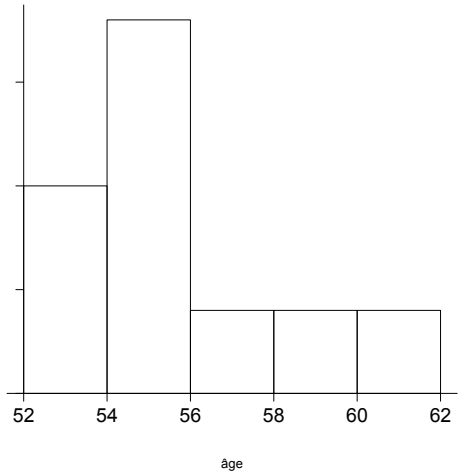
Relations entre mode, moyenne et médiane (1)

- Si la distribution de la variable est symétrique alors le mode, la moyenne et la médiane sont égaux.



Relations entre mode, moyenne et médiane (2)

- Dans le cas de distributions unimodales, asymétriques à droite, la moyenne est supérieure à la médiane. Pour une asymétrie à gauche, la moyenne est inférieure à la médiane.
  Pour la variable âge, l'histogramme montre une asymétrie à droite et on a vu que la moyenne 55,7 est supérieure à la médiane 55,11.



- La médiane est moins sensible que la moyenne à des observations "exceptionnelles" (observations **aberrantes**).

Variables qualitatives - 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).

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

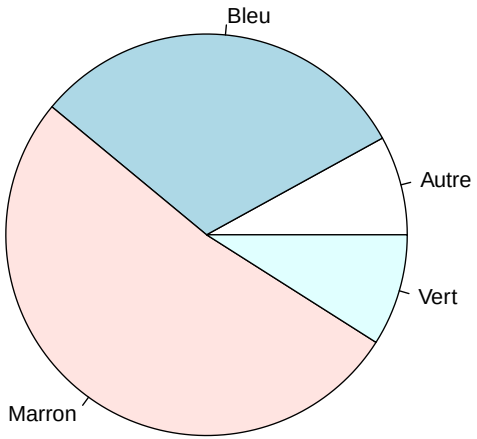
Couleur des yeux	effectifs
Marron	52
Bleus	31
Verst	9
Autre	8

a. Les diagrammes en secteurs

les diagrammes en secteurs (**camemberts**) servent à représenter les effectifs ou les fréquences d'une variable **qualitative**.

$\longrightarrow$ Principe : 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) de chaque secteur est proportionnel à l'effectif ou à la fréquence de la modalité qu'il représente.

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

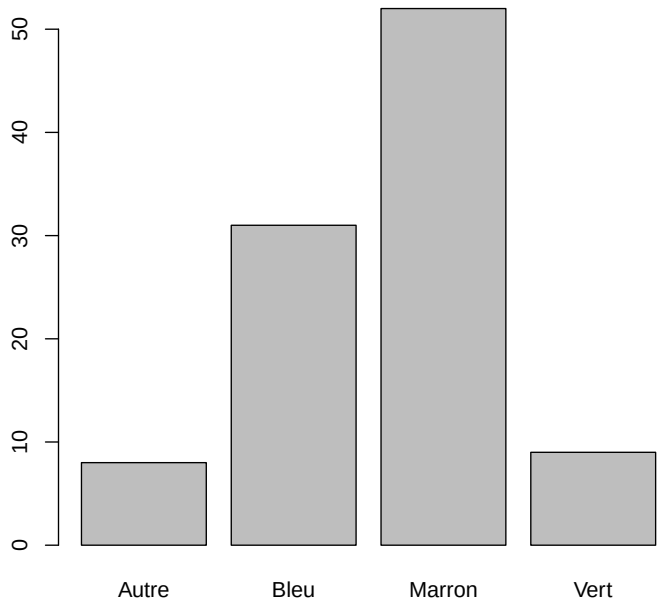


b. Les diagrammes en barre

servent à représenter les effectifs ou les fréquences d'une variable qualitative.

$\longrightarrow$ Principe : on porte les modalités de la variable le long d'une ligne horizontale ; les modalités sont régulièrement espacées sur cet axe. L'axe des ordonnées 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.

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



Étudier les liens entre deux caractères

- On considère :
  - une population Ω (ou un échantillon) de taille N
  - deux variables statistiques X et Y, définies sur Ω, qui peuvent être quantitatives ou qualitatives (ordinales ou nominales).

Étudier les liens entre deux caractères

- On considère :
  - une population Ω (ou un échantillon) de taille N
  - deux variables statistiques X et Y, définies sur Ω, qui peuvent être quantitatives ou qualitatives (ordinales ou nominales).

- Le but est d'analyser la liaison (dépendance) entre ces deux variables en donnant tout d'abord un sens à cette notion.

Étudier les liens entre deux caractères

- On considère :
  - une population Ω (ou un échantillon) de taille N
  - deux variables statistiques X et Y, définies sur Ω, qui peuvent être quantitatives ou qualitatives (ordinales ou nominales).

- Le but est d'analyser la liaison (dépendance) entre ces deux variables en donnant tout d'abord un sens à cette notion.

- 3 cas donc 3 outils spécifiques suivant que les variables X et Y sont
  - toutes les deux qualitatives : le khi-deux
  - toutes les deux quantitatives : la corrélation linéaire
  - l'une quantitative et l'autre qualitative : le rapport de corrélation

Étudier les liens entre deux caractères

- On considère :
  - une population Ω (ou un échantillon) de taille N
  - deux variables statistiques X et Y, définies sur Ω, qui peuvent être quantitatives ou qualitatives (ordinales ou nominales).

- Le but est d'analyser la liaison (dépendance) entre ces deux variables en donnant tout d'abord un sens à cette notion.

- 3 cas donc 3 outils spécifiques suivant que les variables X et Y sont
  - toutes les deux qualitatives : le khi-deux
  - toutes les deux quantitatives : la corrélation linéaire
  - l'une quantitative et l'autre qualitative : le rapport de corrélation

Remarque

Le khi-deux peut également être utilisé pour des variables quelconques (quantitatives ou qualitatives)

Deux variables qualitatives - le khi-deux

- Les variables X et Y sont deux variables qualitatives définies sur Ω et observées sur N individus.
- $x_1, \ldots, x_k$ sont les modalités de la variable X
- $y_1, \ldots, y_m$ celles de la variable Y
- X a donc k modalités et Y a m modalités
- Les deux variables X et Y ont été mesurées sur l'échantillon et pour chaque individu on a donc deux mesures. Ainsi, la série statistique brute est constituée de deux suites de N mesures chacune. Comme dans le cas d'une série statistique simple (une seule variable) on regroupe les observations suivant leurs valeurs : puisque X a k valeurs différentes $x_1, \ldots, x_k$ et Y a m valeurs différentes $y_1, \ldots, y_m$, l'observation conjointe de X et de Y donne $k \times m$ couples différents :

$$\overbrace{\begin{array}{|l} (x_1, y_1), \quad \ldots, \quad (x_1, y_m), \\ (x_2, y_1), \quad \ldots, \quad (x_2, y_m), \\ \vdots \quad \vdots \quad \vdots \\ (x_k, y_1), \quad \ldots, \quad (x_k, y_m) \end{array}}^{m \text{ colonnes}} \Bigg\} k \text{ lignes}$$

Effectifs conjoints du couple (x_i, y_j)

nombre d'individus, noté n_{ij}, ayant pris la modalité x_i de la variable X et la modalité y_j de la variable Y.

- La somme de tous les effectifs conjoints est égale à N :

$$n_{11} + \ldots + n_{k1} + n_{21} + \ldots + n_{k,m} = N$$

Fréquences conjointes du couple (x_i, y_j)

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 :

$$f_{ij} = \frac{n_{ij}}{N} \text{ (effectif conjoint divisé par N)}$$

- La somme de toutes les fréquences conjointes vaut 1 :

$$f_{11} + \ldots + f_{k1} + f_{21} + \ldots + f_{k,m} = 1$$

Effectif marginal de la modalité x_i de la variable X

c'est le nombre d'individus, noté $n_{i.}$, ayant pris la modalité x_i.

$$\boxed{n_{i.} = n_{i1} + \ldots + n_{im}}$$ (somme des effectifs conjoints de la ligne i)

Effectif marginal de la modalité y_j de la variable Y

c'est le nombre d'individus, noté $n_{.j}$, ayant pris la modalité y_j.

$$\boxed{n_{.j} = n_{1j} + \ldots + n_{kj}}$$ (somme des effectifs conjoints de la colonne j)

Fréquence marginale de la modalité x_i de la variable X

c'est la proportion d'individus, noté $f_{i.}$, ayant pris la modalité x_i.

$$\boxed{f_{i.} = f_{i1} + \ldots + f_{im} = \frac{n_{i.}}{N}}$$ (somme des fréquences conjointes de la ligne i ou effectif marginal sur N)

Fréquence marginale de la modalité y_j de la variable Y

c'est la proportion d'individus, noté $f_{.j}$, ayant pris la modalité y_j. On a :

$$\boxed{f_{.j} = f_{1j} + \ldots + f_{kj} = \frac{n_{.j}}{N}}$$ (somme des fréquences conjointes de la colonne j ou effectif marginal sur N)

La table de contingence

- Tableau constitué de k lignes et m colonnes
- sur les lignes sont portées les modalités de la variable X

modalités de la variable X

X						
x_1						
$\vdots$						
x_i						
$\vdots$						
x_k						

La table de contingence

- Tableau constitué de k lignes et m colonnes
- sur les lignes sont portées les modalités de la variable X
- sur les colonnes sont portées les modalités de la variable Y

modalités de la variable Y

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_m	
x_1						
$\vdots$						
x_i						
$\vdots$						
x_k						

La table de contingence

- Tableau constitué de k lignes et m colonnes
- sur les lignes sont portées les modalités de la variable X
- sur les colonnes sont portées les modalités de la variable Y
- à l'intersection de la ligne i et de la colonne j est porté l'effectif conjoint du couple (x_i, y_j).
-
-

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_m	
x_1						
$\vdots$						
x_i			n_{ij}			
$\vdots$						
x_k						

Effectif conjoint du couple (x_i, y_j)

La table de contingence

- Tableau constitué de k lignes et m colonnes
- sur les lignes sont portées les modalités de la variable X
- sur les colonnes sont portées les modalités de la variable Y
- à l'intersection de la ligne i et de la colonne j est porté l'effectif conjoint du couple (x_i, y_j).
- colonne supplémentaire avec les effectifs marginaux de la variable X
-

X \ Y	y_1	$\dots$	y_j	$\dots$	y_m	Somme X
x_1						
$\vdots$						
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{im}	$n_{i.}$
$\vdots$						
x_k						
						N

Effectif marginal de x_i

La table de contingence

- Tableau constitué de k lignes et m colonnes
- sur les lignes sont portées les modalités de la variable X
- sur les colonnes sont portées les modalités de la variable Y
- à l'intersection de la ligne i et de la colonne j est porté l'effectif conjoint du couple (x_i, y_j).
- colonne supplémentaire avec les effectifs marginaux de la variable X
- ligne supplémentaire avec les effectifs marginaux de la variable Y.

X \ Y	y_1	$\dots$	y_j	$\dots$	y_m	Somme X
x_1			n_{1j}			
$\vdots$			$\vdots$			
x_i			n_{ij}			
$\vdots$			$\vdots$			
x_k			n_{kj}			
Somme Y			$n_{.j}$			N

Effectif marginal de y_j

Exemple

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

X				
N'ont pas fleuri				
Ont fleuri				

modalités de la variable floraison

Exemple

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

modalités de la variable type d'engrais

Y / X	Engrais A	Engrais B	Engrais C	
N'ont pas fleuri				
Ont fleuri				

Exemple

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

X \ Y	Engrais A	Engrais B	Engrais C	
N'ont pas fleuri	16	12	12	
Ont fleuri	34	73	63	

Exemple

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

X \ Y	Engrais A	Engrais B	Engrais C	Somme X
N'ont pas fleuri	16	12	12	40
Ont fleuri	34	73	63	170
Somme Y	50	85	75	210

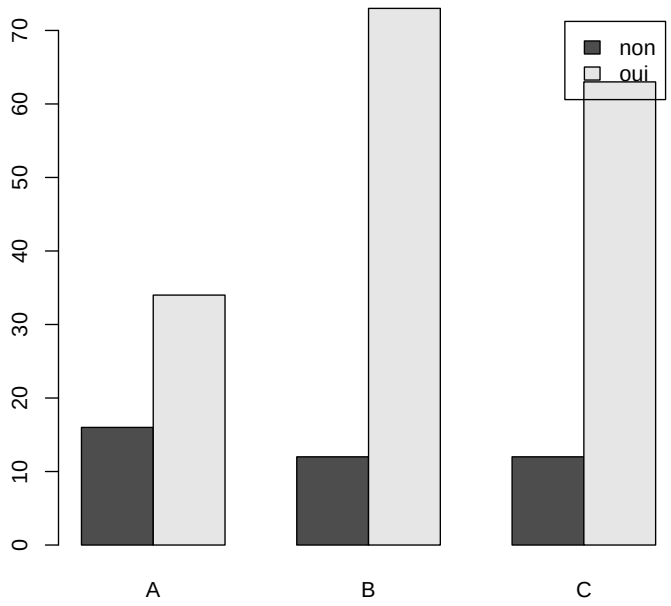
Les données ont été recensées dans le data.frame plante dans le logiciel R. Celui-ci contient deux colonnes ”floraison” et ”engrais” de longueur $N = 210$.

```
> 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

Fréquences conditionnelles

On appelle fréquence de la modalité y_j conditionnellement à la modalité x_i la proportion d'individus ayant pris la modalité y_j parmi tous les individus ayant pris la modalité x_i. C'est donc la quantité définie par :

$$\boxed{\frac{n_{ij}}{n_{i.}}} \text{ (effectif conjoint de } (x_i, y_j) \text{ divisé par l'effectif marginal de } x_i)$$

On définit ainsi le $i^{\text{ème}}$ profil-ligne par :

$$\boxed{\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \ldots, \frac{n_{im}}{n_{i.}}} \text{ (fréquences conditionnelles à la modalité } x_i)$$

- On définit de même la fréquence d'une modalité x_i de X conditionnellement à une modalité y_j de Y :

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

ainsi que le $j^{\text{ème}}$ profil-colonne :

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

Calcul des profils-lignes :

X \ Y	Engrais A	Engrais B	Engrais C	Somme (X)
N'ont pas fleuri	16	12	12	40

÷40

X \ Y	Engrais A	Engrais B	Engrais C	Somme
N'ont pas fleuri	0,4	0,3	0,3	1

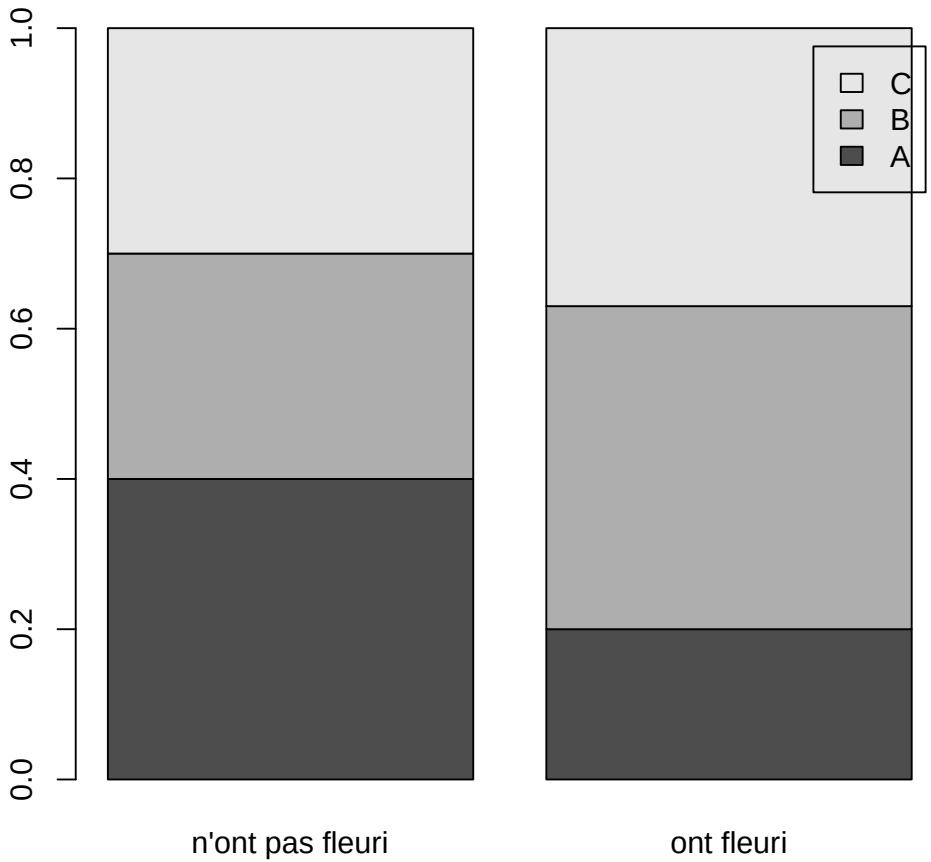
Calcul des profils-lignes :

X \ Y	Engrais A	Engrais B	Engrais C	Somme (X)
N'ont pas fleuri	16	12	12	40
Ont fleuri	34	73	63	170

$\div 170$

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

Représentation graphique



Étude de la liaison entre deux variables qualitatives

- Le calcul des profils lignes dans l'exemple montrent que ceux-ci présentent des différences selon
- Ces différences indiquent-elles une liaison entre les deux variables ?

Les deux exemples précédents montrent qu'il est naturel de baser l'analyse de la liaison entre deux variables qualitatives sur les profils-lignes ou, de manière équivalente, sur les profils-colonnes. On a ainsi la définition suivante :

État de non liaison

On dit que les variables statistiques (qualitatives) X et Y sont **non liées** si et seulement si 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 les différentes **profils-colonnes** sont égaux.

État de non liaison - traduction mathématique (1)

- Égalité des profils-lignes :

	X \ Y	y_1	$\dots$	y_j	$\dots$	y_m	Somme
=	$\vdots$	=		=		=	
ligne i	x_i	$\frac{n_{i1}}{n_{i.}}$		$\frac{n_{ij}}{n_{i.}}$		$\frac{n_{im}}{n_{i.}}$	1
=	$\vdots$	=		=		=	
ligne l	x_l	$\frac{n_{l1}}{n_{l.}}$		$\frac{n_{lj}}{n_{l.}}$		$\frac{n_{lm}}{n_{l.}}$	1
=	$\vdots$	=		=		=	

s'écrit :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{lj}}{n_{l.}},\ j = 1, \dots, m,\ i = 1, \dots, k,\ l = 1, \dots, k \qquad (1)$$

- L'égalité des différents profils-colonnes s'écrit :

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{il}}{n_{.l}},\ i = 1, \dots, k,\ j = 1, \dots, m,\ l = 1, \dots, m \qquad (2)$$

État de non liaison - traduction mathématique (2)

- On a vu que :

$$\text{égalité des profils-lignes} \Longleftrightarrow \text{égalité des profils-colonnes}$$
$$(1) \Longleftrightarrow (2)$$

- On montre que ces deux assertions sont encore équivalentes à :

  les effectifs conjoints sont égaux aux produits des effectifs marginaux divisé par la taille de l'échantillon N

  ce qui s'écrit :

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

- On utilise cette dernière propriété pour construire un indice mesurant la liaison entre les variables X et Y : cet indice est basé sur la comparaison entre l'effectif observé n_{ij} et l'effectif $\frac{n_{i.}n_{.j}}{N}$ que l'on devrait obtenir si il n'y a pas de liaison entre X et Y.

Effectifs conjoints théoriques

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

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

- On utilise cette dernière propriété pour construire un indice mesurant la liaison entre les variables X et Y : cet indice est basé sur la comparaison entre l'effectif observé n_{ij} et l'effectif $\frac{n_{i.}n_{.j}}{N}$ que l'on devrait obtenir si il n'y a pas de liaison entre X et Y.

Effectifs conjoints théoriques

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

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

X \ Y	$\dots$	y_j	$\dots$	Somme (X)
$\vdots$				
x_i		$t_{ij} = \frac{n_{i.}n_{.j}}{N}$		eff. marginal $n_{i.}$
$\vdots$				
Somme (Y)		eff. marginal $n_{.j}$		N

Table de contingence des effectifs théoriques

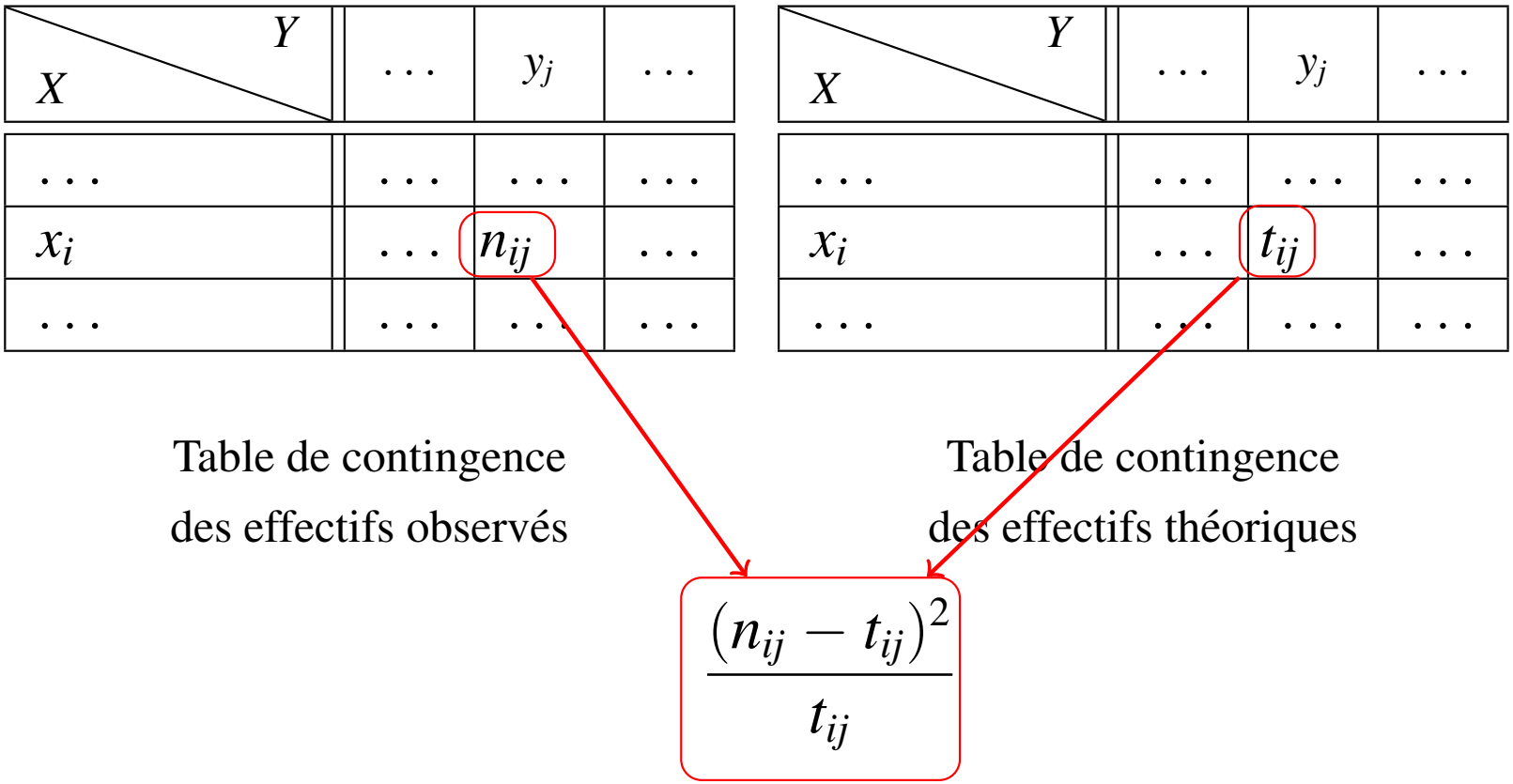
X \ Y	Engrais A	Engrais B	Engrais C	Totaux
Ont fleuri	40,476	68,810	60,714	170
N'ont pas fleuri	9,524	16,190	14,286	40
Totaux	50	85	75	210

Table de contingence des effectifs théoriques

X \ Y	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

- $40,476 = \frac{170 \times 50}{210}$

Définition du khi-deux - comparaison de deux tables



Définition du khi-deux

- Comme vu plus haut, l'indice de liaison entre X et Y est basé sur la comparaison entre effectifs conjoints observés et effectifs conjoints théoriques, c'est-à-dire sur les différences $n_{ij} - \frac{n_{i.} n_{.j}}{N}$ pout tout i et j. Dans le cas d'une absence de liaison, ces différences sont nulles.

Le χ^2 de contingence

Le χ^2 de contingence est la quantité définie par :

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

Calcul du χ^2.

- On calcule les effectifs conjoints théoriques
- Pour chaque cellule du tableau, on calcule les quantités : $\frac{(n_{ij}-t_{ij})^2}{t_{ij}}$
- On calcule la somme de ces $k \times m$ quantités.

Exemple. $\chi^2 =$ 7,332

```
> 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
```

Remarques sur le khi-deux

- Absence de liaison : 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) ou qu'il y a absence de liaison.

Remarques sur le khi-deux

- Absence de liaison : 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) ou qu'il y a absence de liaison.
- Liaison parfaite : on parle de liaison parfaite lorsque les individus prenant une modalité quelconque de la variable X prennent de manière certaine et exclusive certaines modalités déterminées de la variable Y. Dans ce cas, le χ^2 atteint sa **valeur maximale**, à savoir :

$$\boxed{\chi^2_{max} = N(min(k,m) - 1)}$$

où $min(k,m)$ est la plus petite valeur entre k et m.

Remarques sur le khi-deux

- Absence de liaison : 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) ou qu'il y a absence de liaison.
- Liaison parfaite : on parle de liaison parfaite lorsque les individus prenant une modalité quelconque de la variable X prennent de manière certaine et exclusive certaines modalités déterminées de la variable Y. Dans ce cas, le χ^2 atteint sa **valeur maximale**, à savoir :

$$\boxed{\chi^2_{max} = N(min(k,m) - 1)}$$

  où $min(k,m)$ est la plus petite valeur entre k et m.
- Liaison ni absente ni parfaite : dans la pratique, les cas les plus fréquemment rencontrés sont ceux où il n'y a ni absence de liaison ni liaison parfaite. Le $\chi 2$ aura donc le plus souvent une valeur intermédiaire entre 0 et χ^2_{max} et selon qu'il sera proche de 0 ou de χ^2_{max} la liaison sera faible ou forte.

Le C de Cramer ou φ de contingence

On a vu plus haut que le χ^2 est compris entre 0 et χ^2_{max}. On construit alors une mesure de liaison normalisée permettant de mesurer l'intensité de la liaison entre X et Y.

C de Cramer

On appelle C de Cramer (ou parfois φ de contingence) la quantité définie par :

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

- 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. on a $C = 0{,}19$

Tableau des échelles d'intensité de la liaison

Cette échelle est donnée à titre indicatif, l'appréciation étant fonction du contexte de l'étude.

C de Cramer	Liaison
0	Nulle
0 à moins de 0,2	Faible
0,2 à moins de 0,4	Moyenne
0,4 à moins de 0,7	Forte
0,7 à moins de 1	Très forte
1	Parfaite

Deux variables quantitatives - la corrélation linéaire

- Les variables X et Y sont deux variables quantitatives définies sur Ω et observées sur N individus.
- Pour chaque individu i on dispose de deux mesures x_i et y_i.
- On a ainsi N couples :

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

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

$$\overline{X} = \frac{x_1 + \ldots + x_N}{N}, \qquad \overline{Y} = \frac{y_1 + \ldots + y_N}{N}$$

- Les variances de X et de Y sont données par :

$$\sigma_X^2 = \frac{\left(x_1 - \overline{X}\right)^2 + \ldots + \left(x_N - \overline{X}\right)^2}{N}$$

$$\sigma_Y^2 = \frac{\left(y_1 - \overline{Y}\right)^2 + \ldots + \left(y_N - \overline{Y}\right)^2}{N}$$

ou de manière alternative :

$$\sigma_X^2 = \frac{x_1^2 + \ldots + x_N^2}{N} - \overline{X}^2$$

$$\sigma_Y^2 = \frac{y_1^2 + \ldots + y_N^2}{N} - \overline{Y}^2$$

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

La covariance et le coefficient de corrélation linéaire

La covariance

La covariance entre deux variables quantitatives X et Y est la quantité notée c_{XY} définie par :

$$c_{XY} = \frac{\left(x_1 - \overline{X}\right)\left(y_1 - \overline{Y}\right) + \ldots + \left(x_N - \overline{X}\right)\left(y_N - \overline{Y}\right)}{N}$$

Autre expression de la covariance. On utilise le plus souvent une expression qui s'avère souvent plus commode à calculer :

$$c_{XY} = \underbrace{\frac{x_1 y_1 + \ldots + x_N y_N}{N}}_{\substack{\text{moyenne} \\ \text{des produits}}} - \underbrace{\overline{X}\,\overline{Y}}_{\substack{\text{produit} \\ \text{des moyennes}}}$$

Remarques sur la covariance

- La covariance est un indice de dispersion conjointe entre les variables X et Y.
- La covariance peut être positive ou négative.
- La covariance est symétrique, c'est-à-dire que :

$$c_{XY} = c_{YX}$$

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

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

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

$$\boxed{c_{XY}^2 \leq \sigma_X^2 \sigma_Y^2}$$

ou encore :

$$\boxed{-\sigma_X \sigma_Y \leq c_{XY} \leq \sigma_X \sigma_Y}$$

Exemple. (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
```

Le coefficient de corrélation linéaire

Le coefficient de corrélation linéaire

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

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

Exemple. (mêmes données que précédemment).

```
> cor(salaire$sal_min,salaire$sal_moy)
[1] 0.9657632
```

Remarques sur le coefficient de corrélation linéaire

- Le coefficient de corrélation linéaire est symétrique :

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

- L'inégalité de Cauchy-Schwartz donne :

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

En effet :

$$-\sigma_X \sigma_Y \leq c_{XY} \leq \sigma_X \sigma_Y \Longrightarrow -1 \leq \frac{c_{XY}}{\sigma_X \sigma_Y} \leq 1$$

Signification

- Si $r_{XY} = \pm 1$, alors : $c_{XY}^2 = \sigma_X^2 \sigma_Y^2$. Cette égalité est vérifiée 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 : pour tout individu i

  $$y_i = ax_i + b$$

  Si $r_{XY} = 1$ alors a est strictement positif et si $r_{XY} = -1$ alors a est strictement négatif.
- Si $r_{XY} = 0$ (ce qui signifie que $c_{XY} = 0$) il n'existe aucune forme de liaison linéaire entre X et Y.
- En dehors de ces valeurs, $|r_{XY}|$ est d'autant plus proche de 1 que la liaison linéaire entre X et Y est grande.

Régression linéaire entre deux variables

- Deux variables quantitatives X et Y observées sur une population Ω (ou un échantillon) de taille N. On suppose qu'il existe une "causalité" entre X et Y: X est la "cause" de Y.

- Y a-t-il une liaison entre X et Y du type $y_i \simeq ax_i + b$ ou encore $y_i = ax_i + b + e_i$ avec e_i "petit" ?

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

Régression linéaire entre deux variables

- Deux variables quantitatives X et Y observées sur une population Ω (ou un échantillon) de taille N. On suppose qu'il existe une "causalité" entre X et Y: X est la "cause" de Y.

- Y a-t-il une liaison entre X et Y du type $y_i \simeq ax_i + b$ ou encore $y_i = ax_i + b + e_i$ avec e_i "petit" ?

- Si oui, quelles sont les valeurs de a et de b?
- Deux outils : le nuage de points et le coefficient de corrélation linéaire

Le nuage de points

Le nuage de points

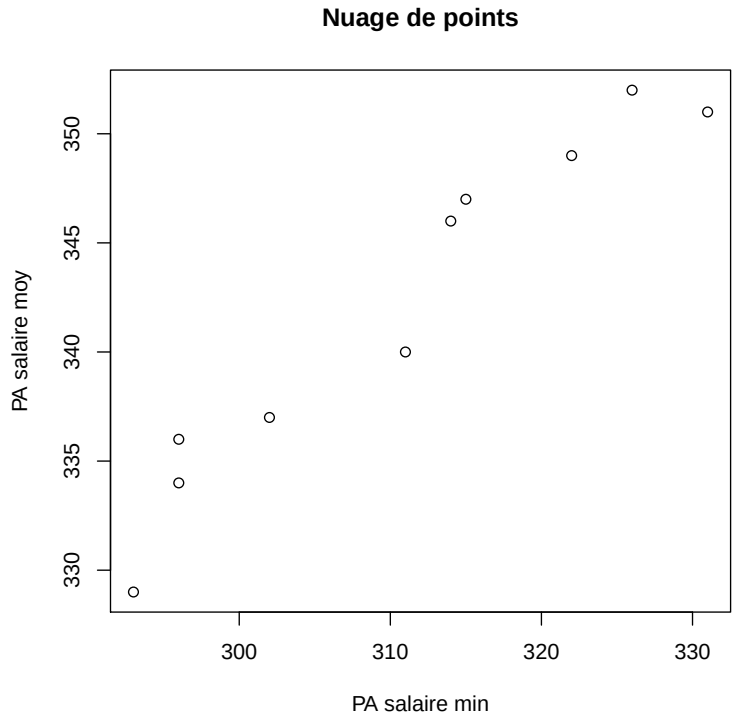
Le nuage de points est la représentation dans un repère orthogonal des N points de coordonnées $(x_i, y_i),\ i = 1, \ldots, N$. L'axe des abscisses est l'axe des valeurs de la variable X et l'axe des ordonnées est l'axe des valeurs de la variable Y.

Dans le cas d'une liaison linéaire forte entre X et Y :

1. les points doivent être relativement alignés ou le nuage de points doit être "étiré",
2. Le coefficient de corrélation linéaire est proche de 1 en valeur absolue.

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

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



1. Le nuage de points ci-dessus est relativement "étiré",
2. $r_{XY} = 0,97$ est proche de 1,

il y a (pour l'échantillon concerné par l'étude) une liaison linéaire forte entre salaire minimum et salaire moyen (par ailleurs on peut supposer qu'il existe une causalité entre salaire minimum et salaire moyen).

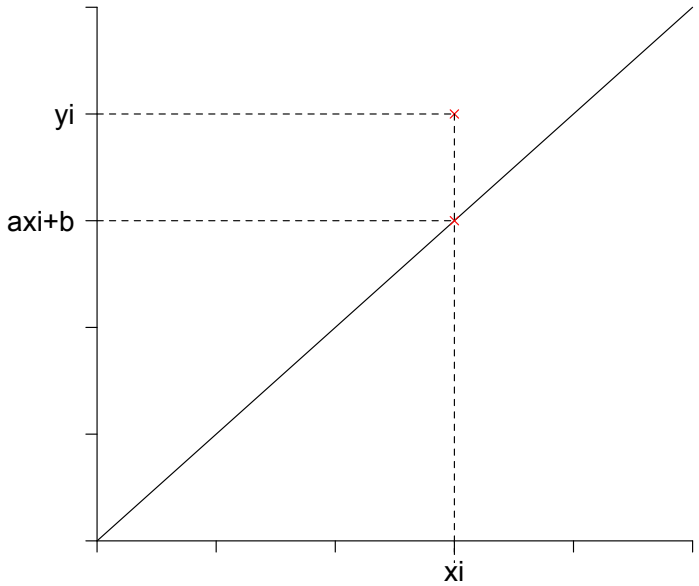
- On suppose qu'il existe une liaison linéaire forte entre X et Y : $y_i \simeq ax_i + b$, pout tout individu i où les coefficients a et b sont inconnus.
- Problème : proposer des valeurs pour ces coefficients.
- Lorsque les points du nuage ne sont pas alignés, il y a une infinité de candidats pour a et b (en effet, les points du nuage sont rarement parfaitement alignés : ils sont "proches" d'une droite)
- Solution unique : méthode des moindres carrés.

La droite de régression

La méthode des moindres carrés.

- Idée : chercher les valeurs de a et b qui rendent globalement les distances entre y_i (valeur observée) et $ax_i + b$ (valeur ajustée) les plus petites possibles.
- Pour chaque individu i on forme le carré de la distance entre y_i et $ax_i + b$. On cherche ensuite les valeurs de a et b qui minimisent "globalement" ces carrés, c'est-à-dire leur somme sur tous les individus. Ainsi, a et b minimisent l'expression :

$$(y_1 - ax_1 - b)^2 + \ldots + (y_N - ax_N - b)^2$$



Méthode des moindres carrés - solution

- On monte qu'il existe un couple unique de réels a et b minimisant le critère des moindres carrés :
- Solution :

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

et

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

Remarques

- Les nombres a et b sont appelés coefficients de régression linéaire.
- La droite d'équation $y = ax + b$ est appelée droite de régression.
- Les valeurs $\widehat{y}_i = ax_i + b$, $i = 1, \ldots, N$, sont appelées les valeurs ajustées
- Les nombres $e_i = y_i - (ax_i + b)$, $i = 1, \ldots, N$ sont appelés les résidus.
  On montre facilement que la moyenne des résidus est nulle.

Remarques

- Les nombres a et b sont appelés coefficients de régression linéaire.
- La droite d'équation $y = ax + b$ est appelée droite de régression.
- Les valeurs $\widehat{y}_i = ax_i + b$, $i = 1, \ldots, N$, sont appelées les valeurs ajustées
- Les nombres $e_i = y_i - (ax_i + b)$, $i = 1, \ldots, N$ sont appelés les résidus.
  On montre facilement que la moyenne des résidus est nulle.

Exemple.

```
> 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

- Problème : Ayant observé pour un individu $N+1$ de la population Ω (n'appartenant pas à l'échantillon) la valeur x_{N+1} de la variable X, peut-on donner une estimation $\widehat{y}_{N+1}$ de Y pour cet individu?
- En utilisant la droite de régression : la vraie valeur (inconnue) y_{N+1} est, lorsque le modèle de régression est "bon", proche de $ax_{N+1}+b$. Une estimation de la valeur y_{N+1} est ainsi donnée par :

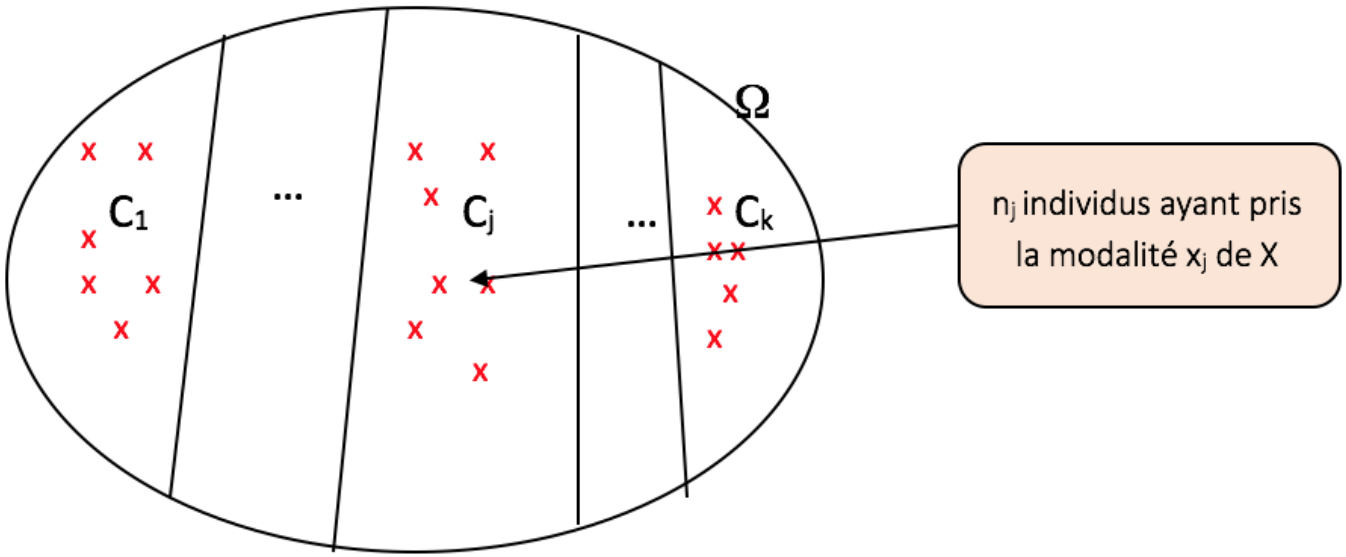
$$\widehat{y}_{N+1} = ax_{N+1}+b.$$

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

- La variable qualitative X et la variable quantitative Y sont observées sur Ω (ou un échantillon) de taille N.
- La variable X a k modalités $x_1, \ldots, x_k$;
  n_j est l'effectif de la modalité x_j.

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

- La variable qualitative X et la variable quantitative Y sont observées sur Ω (ou un échantillon) de taille N.
- La variable X a k modalités $x_1, \ldots, x_k$;
  n_j est l'effectif de la modalité x_j.
- La modalité x_j de la variable X définit une classe C_j de n_j individus, c'est-à-dire les n_j individus ayant pris la modalité x_j.



Exemple. 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

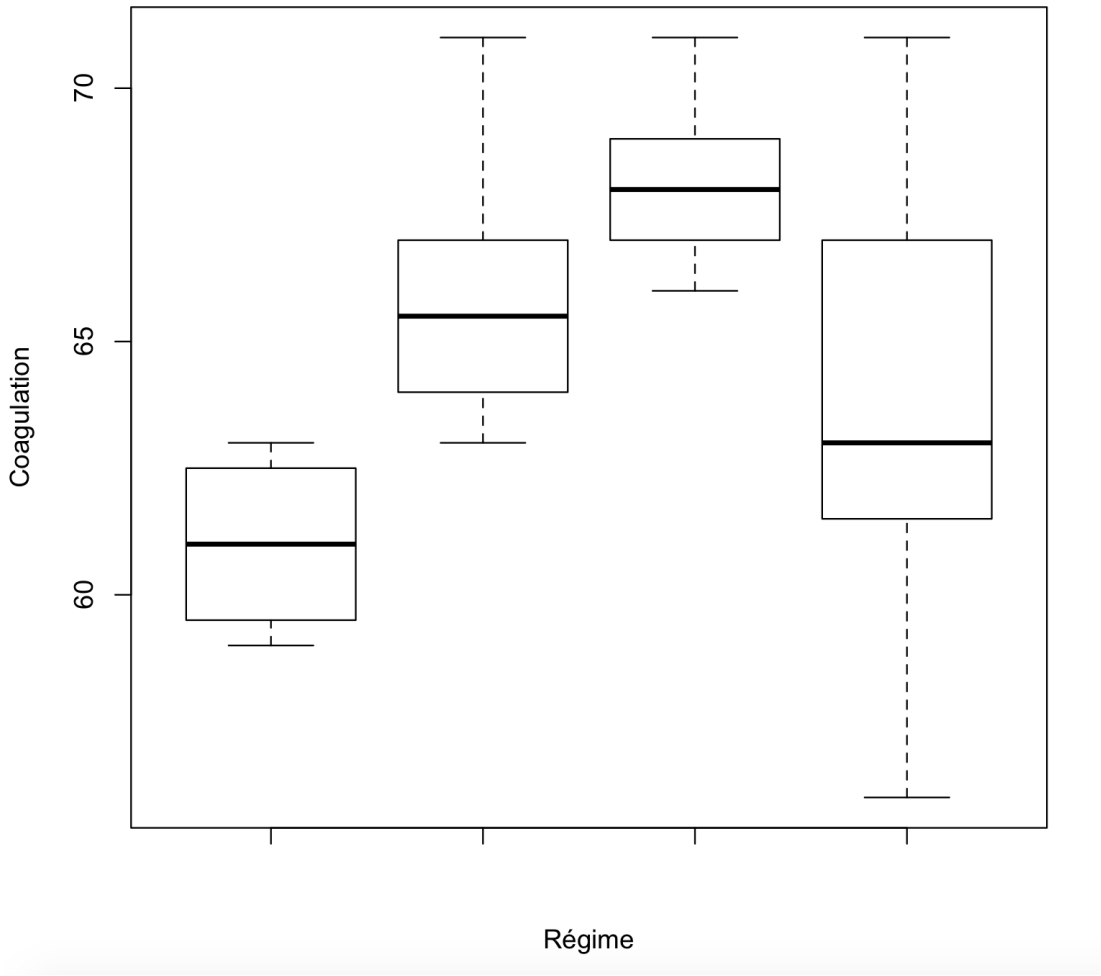
- Nombre d'individus : $N = 24$
- La variable qualitative X, "type de régime" a $k = 4$ modalités : $n_1 = 4,\ n_2 = n_3 = 6, n_4 = 8$
- Les valeurs de Y sont "empilées" sur 4 colonnes correspondant à chaque classe
  par exemple C_1 est la classe des $n_1 = 4$ individus ayant suivi le régime 1

Notion de liaison entre X et Y

- La liaison entre les variables X et Y est basée sur la comparaison des valeurs de la variable Y sur chaque classe C_j (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 au moins en moyenne).

- La liaison est basée sur :
  - la représentation en parallèles des diagrammes en boîtes de Y sur chaque classe ;
  - la comparaison des moyennes de Y sur chaque classe (première indication) ;
  - le calcul d'un indice de liaison, le rapport de corrélation, basée sur la décomposition de la variance de Y.

Représentation graphique - Boîtes parallèles



Coagulation en fonction du régime. Boîtes parallèles.

Moyennes et variances partielles et totales

- Sur chaque classe C_j on calcule la moyenne $\overline{y}_j$ et la variance σ_j^2 de la variable quantitative Y :

$$\boxed{\overline{Y}_j = \frac{1}{n_j} \sum_{i \in C_j} y_i}$$ (somme des observations y_i sur la classe C_j divisée par n_j)

$$\boxed{\sigma_j^2 = \frac{1}{n_j} \sum_{i \in C_j} (Y_i - \overline{Y}_j)^2}$$ (moyenne des écarts à la moyenne au carré sur la classe C_j)

Moyennes et variances partielles et totales

- Sur chaque classe C_j on calcule la moyenne $\overline{y}_j$ et la variance σ_j^2 de la variable quantitative Y :

$$\boxed{\overline{Y}_j = \frac{1}{n_j} \sum_{i \in C_j} y_i}$$ (somme des observations y_i sur la classe C_j divisée par n_j)

$$\boxed{\sigma_j^2 = \frac{1}{n_j} \sum_{i \in C_j} (Y_i - \overline{Y}_j)^2}$$ (moyenne des écarts à la moyenne au carré sur la classe C_j)

- Exemple. Pour la classe C_1 :

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

Moyennes et variances partielles et totales

- Sur chaque classe C_j on calcule la moyenne $\overline{y}_j$ et la variance σ_j^2 de la variable quantitative Y :

$$\boxed{\overline{Y}_j = \frac{1}{n_j}\sum_{i\in C_j} y_i}$$ (somme des observations y_i sur la classe C_j divisée par n_j)

$$\boxed{\sigma_j^2 = \frac{1}{n_j}\sum_{i\in C_j}(Y_i - \overline{Y}_j)^2}$$ (moyenne des écarts à la moyenne au carré sur la classe C_j)

- Exemple. Pour la classe C_1 :

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

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

$$\boxed{\overline{Y} = \frac{y_1 + \ldots + y_N}{N}}$$

$$\boxed{\sigma_Y^2 = \frac{(y_1 - \overline{Y})^2 + \ldots + (y_N - \overline{Y})^2}{N}}$$

Comparaison des moyennes

Exemple.

	C_1	C_2	C_3	C_4	Popu^{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.

Étude de la liaison entre X et Y

- Outre la comparaison des boîtes parallèles et des moyennes sur chaque classe qui fournissent une première analyse de la liaison entre X et Y, on définit un indice de liaison : le rapport de corrélation.

Étude de la liaison entre X et Y

- Outre la comparaison des boîtes parallèles et des moyennes sur chaque classe 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 totale σ_Y^2.

Étude de la liaison entre X et Y

- Outre la comparaison des boîtes parallèles et des moyennes sur chaque classe 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 totale σ_Y^2.
- Notons tout d'abord que moyenne totale et moyennes partielles sont liées par la relation :

$$\overline{Y} = \frac{n_1\overline{Y_1} + \ldots + n_k\overline{Y_k}}{N} \text{ ($\overline{Y}$ est la moyenne pondérée des moyennes partielles)}$$

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 :

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

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 :

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

Variance expliquée par la partition ou variance inter-classes : σ_E^2. 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 $n_1, \ldots, n_j$. 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.

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 :

$$\sigma_Y^2 = \underbrace{\frac{1}{N}\sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2}_{\sigma_E^2} + \underbrace{\frac{1}{N}\sum_{j=1}^{k} n_j\sigma_j^2}_{\sigma_R^2}$$

Variance expliquée par la partition ou variance inter-classes : σ_E^2. 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 $n_1, \ldots, n_j$. 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 : σ_R^2. Ce terme mesure la variation à l'intérieur des classes.

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 :

$$\sigma_Y^2 = \underbrace{\frac{1}{N}\sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2}_{\sigma_E^2} + \underbrace{\frac{1}{N}\sum_{j=1}^{k} n_j\sigma_j^2}_{\sigma_R^2}$$

Variance expliquée par la partition ou variance inter-classes : σ_E^2. 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 $n_1, \ldots, n_j$. 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 : σ_R^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.
  Plus le premier terme σ_E^2 est grand comparativement au second terme σ_R^2 plus les variables X et Y sont liées (suivant le principe de comparaison des moyennes).

Le rapport de corrélation

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

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

- Le coefficient $C_{Y|X}$ n'est pas symétrique.
- $0 \leq C_{Y|X} \leq 1$ (propriété 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 chaque classe). Dans ce cas, la connaissance de X est suffisante pour connaître Y. Il y a liaison totale entre X et Y.

Tableau des intensités de liaison

$C_{Y\|X}$	Liaison
0	Nulle
0 à moins de 0,3	Faible
0,3 à moins de 0,6	Moyenne
0,6 à moins de 0,8	Forte
0,8 à moins de 1	Très forte
1	Parfaite

Exemple. on trouve :

$$\sigma_E^2 \simeq 9,85$$

et

$$C_{Y|X} = \sqrt{\frac{9,85}{14,54}} \simeq \boxed{0{,}82}$$

ce qui indique une liaison très forte entre X et Y.