Objectif et moyens

Objectifs du cours

- Apprendre les principales techniques de statistique descriptive univariée et bivariée.
- Être capable de mettre en oeuvre ces techniques de manière appropriée dans un contexte donné.
- Être capable d'utiliser les commandes de base du Language R. Pouvoir appliquer les techniques de statistiques descriptives au moyen du language R.
- Références
  Dodge Y.(2003), *Premiers pas en statistique*, Springer.
  Droesbeke J.-J. (1997), *Éléments de statistique*, Editions de l'Université libre de Bruxelles/Ellipses.

Moyens

- 2 heures de cours par semaine.
- 2 heures de TP par semaine, répartis en TP théoriques et applications en Language R.

Le language R

- Shareware : gratuit et installé en 10 minutes.
- Open source (on sait ce qui est réellement calculé).
- Développé par la communauté des chercheurs, contient énormément de fonctionnalités.
- Possibilité de programmer.
- Désavantage : pas très convivial.
- Manuel :
  `http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf`

Table des matières

Chapitre 1

Variables, données statistiques, tableaux, effectifs

1.1 Définitions fondamentales

1.1.1 La science statistique

- Méthode scientifique du traitement des données quantitatives.
- Etymologiquement : science de l'état.
- La statistique s'applique à la plupart des disciplines : agronomie, biologie, démographie, économie, sociologie, linguistique, psychologie, ...

1.1.2 Mesure et variable

- On s'intéresse à des *unités statistiques* ou *unités d'observation* : par exemple des individus, des entreprises, des ménages. En sciences humaines, on s'intéresse dans la plupart des cas à un nombre fini d'unités.
- Sur ces unités, on mesure un caractère ou une *variable*, le chiffre d'affaires de l'entreprise, le revenu du ménage, l'âge de la personne, la catégorie socioprofessionnelle d'une personne. On suppose que la variable prend toujours une seule valeur sur chaque unité. Les variables sont désignées par simplicité par une lettre (X, Y, Z).
- Les *valeurs possibles* de la variable, sont appelées *modalités*.
- L'ensemble des valeurs possibles ou des modalités est appelé le *domaine* de la variable.

1.1.3 Typologie des variables

- *Variable qualitative* : La variable est dite qualitative quand les modalités sont des catégories.
  - *Variable qualitative nominale* : La variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées.
  - *Variable qualitative ordinale* : La variable est dite qualitative ordinale quand les modalités peuvent être ordonnées. Le fait de pouvoir ou non ordonner les modalités est parfois discutable. Par exemple : dans les catégories socioprofessionnelles, on admet d'ordonner les modalités : 'ouvriers', 'employés', 'cadres'. Si on ajoute les modalités 'sans profession', 'enseignant', 'artisan', l'ordre devient beaucoup plus discutable.
- *Variable quantitative* : Une variable est dite quantitative si toute ses valeurs possibles sont numériques.
  - *Variable quantitative discrète* : Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable.
  - *Variable quantitative continue* : Une variable est dite continue, si l'ensemble des valeurs possibles est continu.

Remarque 1.1 Ces définitions sont à relativiser, l'âge est théoriquement une variable quantitative continue, mais en pratique, l'âge est mesuré dans le meilleur des cas au jour près. Toute mesure est limitée en précision !

Exemple 1.1 Les modalités de la variable *sexe* sont *masculin* (codé M) et *féminin* (codé F). Le domaine de la variable est $\{M, F\}$.

Exemple 1.2 Les modalités de la variable nombre d'enfants par famille sont 0,1,2,3,4,5,...C'est une variable quantitative discrète.

1.1.4 Série statistique

On appelle *série statistique* la suite des valeurs prises par une variable X sur les unités d'observation.

Le nombre d'unités d'observation est noté n.

Les valeurs de la variable X sont notées

$$x_1, \ldots, x_i, \ldots, x_n.$$

Exemple 1.3 On s'intéresse à la variable 'état-civil' notée X et à la série statistique des valeurs prises par X sur 20 personnes. La codification est

C :	célibataire,
M :	marié(e),
V :	veuf(ve),
D :	divorcée.

Le domaine de la variable X est $\{C, M, V, D\}$. Considérons la série statistique suivante :

M	M	D	C	C	M	C	C	C	M
C	M	V	M	V	D	C	C	C	M

Ici, $n = 20$,

$$x_1 = M, x_2 = M, x_3 = D, x_4 = C, x_5 = C, \ldots, x_{20} = M.$$

1.2 Variable qualitative nominale

1.2.1 Effectifs, fréquences et tableau statistique

Une variable qualitative nominale a des valeurs distinctes qui ne peuvent pas être ordonnées. On note J le nombre de valeurs distinctes ou modalités. Les valeurs distinctes sont notées $x_1, \ldots, x_j, \ldots, x_J$. On appelle *effectif* d'une modalité ou d'une valeur distincte, le nombre de fois que cette modalité (ou valeur distincte) apparaît. On note n_j l'effectif de la modalité x_j. La fréquence d'une modalité est l'effectif divisé par le nombre d'unités d'observation.

$$f_j = \frac{n_j}{n}, j = 1, \ldots, J.$$

Exemple 1.4 Avec la série de l'exemple précédent, on obtient le tableau statistique :

x_j	n_j	f_j
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
	$n = 20$	1

En langage R

```
>X=c('Marié(e)','Marié(e)','Divorcé(e)','Célibataire','Célibataire','Marié(e)','Cél
     'Célibataire','Célibataire','Marié(e)','Célibataire','Marié(e)','Veuf(ve)','Ma
     'Veuf(ve)','Divorcé(e)','Célibataire','Célibataire','Célibataire','Marié(e)')
> T1=table(X)
> V1=c(T1)
> data.frame(Eff=V1,Freq=V1/sum(V1))
            Eff Freq
Célibataire   9 0.45
Divorcé(e)    2 0.10
Marié(e)      7 0.35
Veuf(ve)      2 0.10
```

1.2.2 Diagramme en secteurs et diagramme en barres

Le tableau statistique d'une variable qualitative nominale peut être représenté par deux types de graphique. Les effectifs sont représentés par un diagramme en barres et les fréquences par un diagramme en secteurs (ou camembert ou *piechart* en anglais) (voir Figures 1.1 et 1.2).

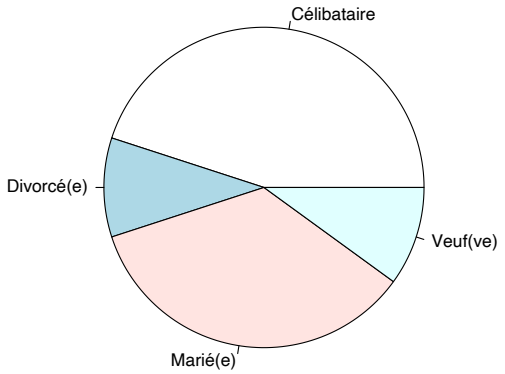


FIGURE 1.1 – Diagramme en secteurs des fréquences

En langage R

```
> pie(T1,radius=1.0)
```

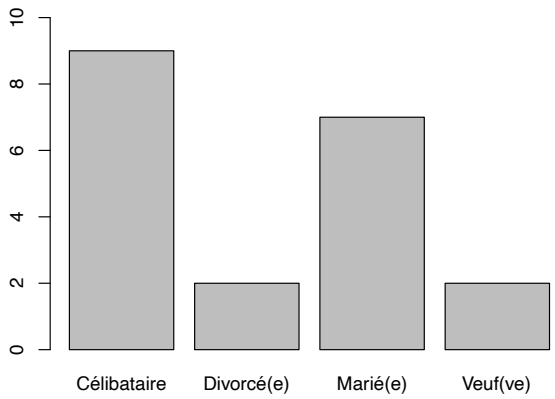


FIGURE 1.2 – Diagramme en barres des effectifs

En langage R

```
>m=max(V1)
>barplot(T1, ylim=c(0,m+1))
```

1.3 Variable qualitative ordinale

1.3.1 Le tableau statistique

Les valeurs distinctes d'une variable ordinale peuvent être ordonnées, ce qu'on écrit

$$x_1 \prec x_2 \prec \cdots \prec x_{j-1} \prec x_j \prec \cdots \prec x_{J-1} \prec x_J.$$

La notation $x_1 \prec x_2$ se lit x_1 précède x_2.

Si la variable est ordinale, on peut calculer les effectifs cumulés :

$$N_j = \sum_{k=1}^{j} n_k, j = 1, \ldots, J.$$

On a $N_1 = n_1$ et $N_J = n$. On peut également calculer les fréquences cumulées

$$F_j = \frac{N_j}{n} = \sum_{k=1}^{j} f_k, j = 1, \ldots, J.$$

Exemple 1.5 On interroge 50 personnes sur leur dernier diplôme obtenu (variable Y). La codification a été faite selon le Tableau 1.1. On a obtenu la série

TABLE 1.1 – Codification de la variable Y

Dernier diplôme obtenu	x_j
Sans diplôme	Sd
Primaire	P
Secondaire	Se
Supérieur non-universitaire	Su
Universitaire	U

TABLE 1.2 – Série statistique de la variable Y

Sd	Sd	Sd	Sd	P	P	P	P	P	P	P	P	P	P	P	Se	Se
Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Su	Su	Su	Su	Su
Su	Su	Su	Su	U	U	U	U	U	U	U	U	U	U	U	U	

TABLE 1.3 – Tableau statistique complet

x_j	n_j	N_j	f_j	F_j
Sd	4	4	0.08	0.08
P	11	15	0.22	0.30
Se	14	29	0.28	0.58
Su	9	38	0.18	0.76
U	12	50	0.24	1.00
	50		1.00	

statistique présentée dans le tableau 1.2. Finalement, on obtient le tableau statistique complet présenté dans le Tableau 1.3.

En langage R

```
> YY=c("Sd","Sd","Sd","Sd","P","P","P","P","P","P","P","P","P","P","P",
  "Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se","Se",
  "Su","Su","Su","Su","Su","Su","Su","Su","Su",
  "U","U","U","U","U","U","U","U","U","U","U","U")
  YF=factor(YY,levels=c("Sd","P","Se","Su","U"))
  T2=table(YF)
  V2=c(T2)
> data.frame(Eff=V2,EffCum=cumsum(V2),Freq=V2/sum(V2),FreqCum=cumsum(V2/sum(V2)))
   Eff EffCum Freq FreqCum
Sd   4      4 0.08    0.08
```

```
P   11      15 0.22    0.30
Se  14      29 0.28    0.58
Su   9      38 0.18    0.76
U   12      50 0.24    1.00
```

1.3.2 Diagramme en secteurs

Les fréquences d'une variable qualitative ordinale sont représentées au moyen d'un diagramme en secteurs (voir Figure 1.3).

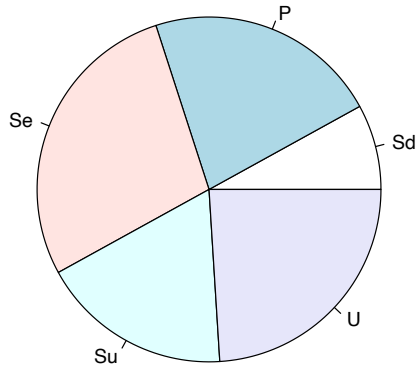


FIGURE 1.3 – Diagramme en secteurs des fréquences

En langage R

```
> pie(T2,radius=1)
```

1.3.3 Diagramme en barres des effectifs

Les effectifs d'une variable qualitative ordinale sont représentés au moyen d'un diagramme en barres (voir Figure 1.4).

En langage R

```
> barplot(T2)
```

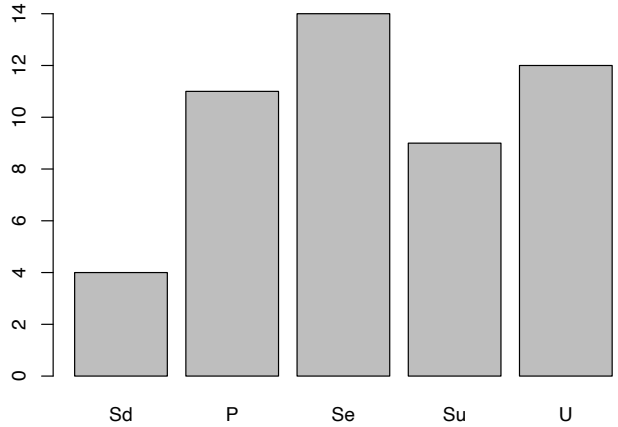



FIGURE 1.4 – Diagramme en barres des effectifs

1.3.4 Diagramme en barres des effectifs cumulés

Les effectifs cumulés d'une variable qualitative ordinale sont représentés au moyen d'un diagramme en barres (voir Figure 1.5).

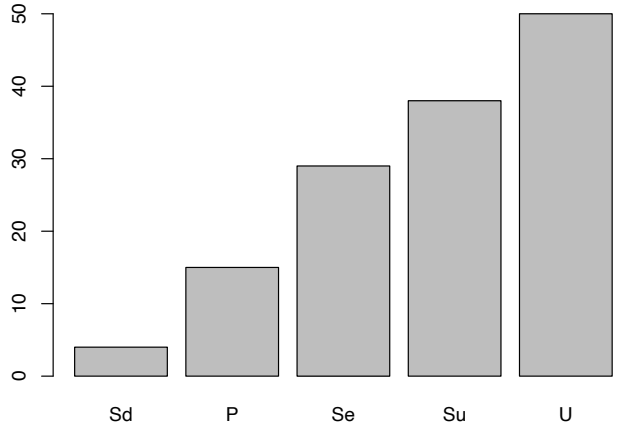


FIGURE 1.5 – Diagramme en barres des effectifs cumulés

En langage R

```
> T3=cumsum(T2)
> barplot(T3)
```

1.4 Variable quantitative discrète

1.4.1 Le tableau statistique

Une variable discrète a un domaine dénombrable.

Exemple 1.6 Un quartier est composé de 50 ménages, et la variable Z représente le nombre de personnes par ménage. Les valeurs de la variable sont

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

Comme pour les variables qualitatives ordinales, on peut calculer les effectifs, les effectifs cumulés, les fréquences, les fréquences cumulées. À nouveau, on peut construire le tableau statistique :

x_j	n_j	N_j	f_j	F_j
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1.00
	50		1.0	

En langage R

```
> Z=c(1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,
+   4,4,4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,8,8)
> T4=table(Z)
> T4c=c(T4)
> data.frame(Eff=T4c,EffCum=cumsum(T4c),Freq=T4c/sum(T4c),FreqCum=cumsum(T4c/sum(T4c)))
  Eff EffCum Freq FreqCum
```



```
1   5       5 0.10    0.10
2   9      14 0.18    0.28
3  15      29 0.30    0.58
4  10      39 0.20    0.78
5   6      45 0.12    0.90
6   3      48 0.06    0.96
8   2      50 0.04    1.00
```

1.4.2 Diagramme en bâtonnets des effectifs

Quand la variable est discrète, les effectifs sont représentés par des bâtonnets (voir Figure 1.6).

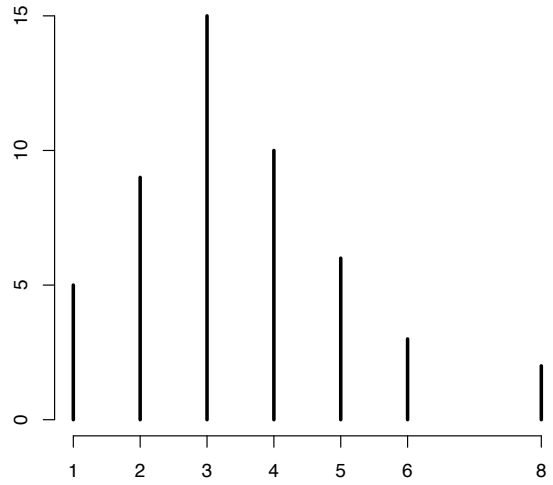


FIGURE 1.6 – Diagramme en bâtonnets des effectifs pour une variable quantitative discrète

En langage R

```
> plot(T4,type="h",xlab="",ylab="",main="",frame=0,lwd=3)
```

1.4.3 Fonction de répartition

Les fréquences cumulées sont représentées au moyen de la fonction de répartition. Cette fonction, présentée en Figure 1.7,est définie de $\mathbb{R}$ dans $[0,1]$ et vaut :

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_j & x_j \leq x < x_{j+1} \\ 1 & x_J \leq x. \end{cases}$$

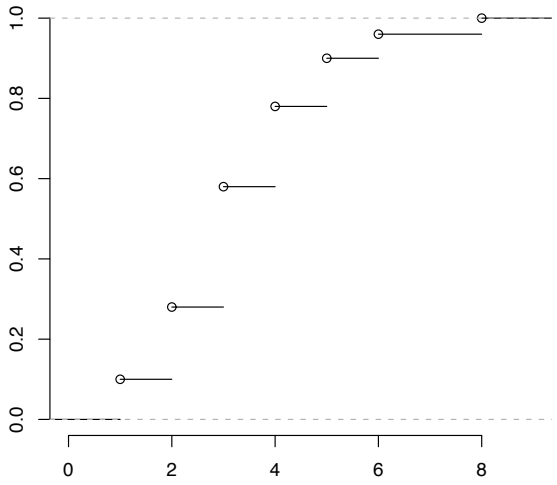


FIGURE 1.7 – Fonction de répartition d'une variable quantitative discrète

En langage R

```
> plot(ecdf(Z),xlab="",ylab="",main="",frame=0)
```

1.5 Variable quantitative continue

1.5.1 Le tableau statistique

Une variable quantitative continue peut prendre une infinité de valeurs possibles. Le domaine de la variable est alors $\mathbb{R}$ ou un intervalle de $\mathbb{R}$. En pratique, une mesure est limitée en précision. La taille peut être mesurée en centimètres, voire en millimètres. On peut alors traiter les variables continues comme des variables discrètes. Cependant, pour faire des représentations graphiques et

construire le tableau statistique, il faut procéder à des regroupements en classes. Le tableau regroupé en classe est souvent appelé *distribution groupée*. Si $[c_j^-\,;c_j^+[$ designe la classe j, on note, de manière générale :

- c_j^- la borne inférieure de la classe j,
- c_j^+ la borne supérieure de la classe j,
- $c_j = (c_j^+ + c_j^-)/2$ le centre de la classe j,
- $a_j = c_j^+ - c_j^-$ l'amplitude de la classe j,
- n_j l'effectif de la classe j,
- N_j l'effectif cumulé de la classe j,
- f_j la fréquence de la classe j,
- F_j la fréquence cumulée de la classe j.

La répartition en classes des données nécessite de définir *a priori* le nombre de classes J et donc l'amplitude de chaque classe. En règle générale, on choisit au moins cinq classes de même amplitude. Cependant, il existent des formules qui nous permettent d'établir le nombre de classes et l'intervalle de classe (l'amplitude) pour une série statistique de n observations.

- La règle de Sturge : $J = 1 + (3.3\log_{10}(n))$.
- La règle de Yule : $J = 2.5\sqrt[4]{n}$.

L'intervalle de classe est obtenue ensuite de la manière suivante : longueur de l'intervalle $= (x_{max} - x_{min})/J$, où x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

Remarque 1.2 Il faut arrondir le nombre de classe J à l'entier le plus proche. Par commodité, on peut aussi arrondir la valeur obtenue de l'intervalle de classe.

A partir de la plus petite valeur observée, on obtient les bornes de classes en additionnant successivement l'intervalle de classe (l'amplitude).

Exemple 1.7 On mesure la taille en centimetres de 50 élèves d'une classe :

152	152	152	153	153
154	154	154	155	155
156	156	156	156	156
157	157	157	158	158
159	159	160	160	160
161	160	160	161	162
162	162	163	164	164
164	164	165	166	167
168	168	168	169	169
170	171	171	171	171

On a les classes de tailles définies préablement comme il suit :

$[151,5; 155,5[$
$[155,5; 159,5[$
$[159,5; 163,5[$
$[163,5; 167,5[$
$[167,5; 171,5[$

On construit le tableau statistique.

$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[151,5; 155,5[$	10	10	0.20	0.20
$[155,5; 159,5[$	12	22	0.24	0.44
$[159,5; 163,5[$	11	33	0.22	0.66
$[163,5; 167,5[$	7	40	0.14	0.80
$[167,5; 171,5[$	10	50	0.20	1.00
	50		1.00	

En langage R

```
> S=c(152,152,152,153,153,154,154,154,155,155,156,156,156,156,156,
+ 157,157,157,158,158,159,159,160,160,160,161,160,160,161,162, +
162,162,163,164,164,164,164,165,166,167,168,168,168,169,169, +
170,171,171,171,171)
> T5=table(cut(S, breaks=c(151,155,159,163,167,171)))
> T5c=c(T5)
> data.frame(Eff=T5c,EffCum=cumsum(T5c),Freq=T5c/sum(T5c),FreqCum=cumsum(T5c/sum(T5c)))
          Eff EffCum Freq FreqCum
(151,155]  10     10 0.20    0.20 (155,159]  12     22 0.24 0.44
(159,163]  11     33 0.22    0.66 (163,167]   7     40 0.14 0.80
(167,171]  10     50 0.20    1.00
```

1.5.2 Histogramme

L'histogramme consiste à représenter les effectifs (resp. les fréquences) des classes par des rectangles contigus dont la surface (et non la hauteur) représente l'effectif (resp. la fréquence). Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe j est donc donnée par :

$$h_j = \frac{n_j}{a_j}$$

– On appelle h_j la densité d'effectif.

– L'aire de l'histogramme est égale à l'effectif total n, puisque l'aire de chaque rectangle est égale à l'effectif de la classe j : $a_j \times h_j = n_j$.

Pour un histogramme des fréquences on a

$$d_j = \frac{f_j}{a_j}$$

– On appelle d_j la densité de fréquence.
– L'aire de l'histogramme est égale à 1, puisque l'aire de chaque rectangle est égale à la fréquence de la classe j : $a_j \times d_j = f_j$.

Figure 1.8 représente l'histogramme des fréquences de l'exemple précedent :

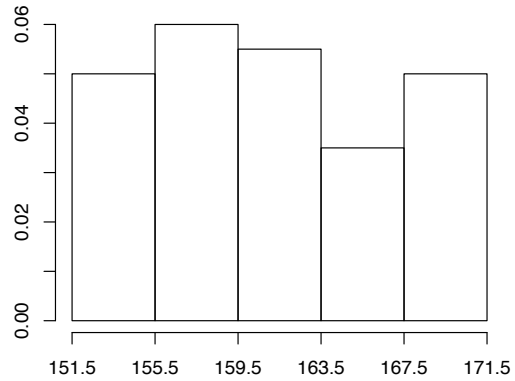


FIGURE 1.8 – Histogramme des fréquences

En langage R

```
> hist(S,breaks=c(151.5,155.5,159.5,163.5,167.5,171.5), freq=FALSE,
  xlab="",ylab="",main="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,167.5,171.5))
```

Si les deux dernières classes sont agrégées, comme dans la Figure 1.9, la surface du dernier rectangle est égale à la surface des deux derniers rectangles de l'histogramme de la Figure 1.8.

En langage R

```
> hist(S,breaks=c(151.5,155.5,159.5,163.5,171.5),
  xlab="",ylab="",main="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,171.5))
```

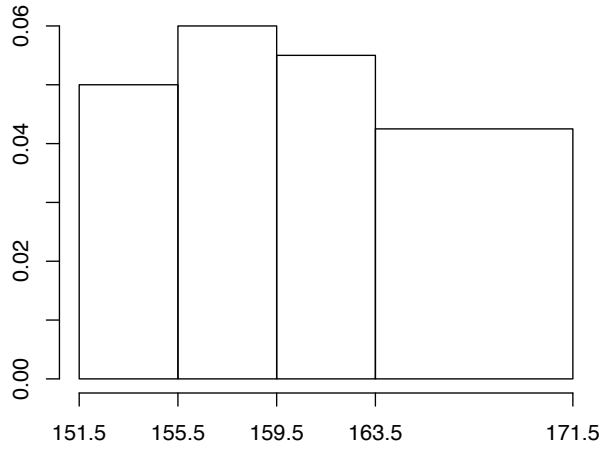


FIGURE 1.9 – Histogramme des fréquences avec les deux dernières classes agrégées

Remarque 1.3 Dans le cas de classes de même amplitude certains auteurs et logiciels représentent l'histogramme avec les effectifs (resp. les fréquences) reportés en ordonnée, l'aire de chaque rectangle étant proportionnelle à l'effectif (resp. la fréquence) de la classe.

1.5.3 La fonction de répartition

La fonction de répartition $F(x)$ est une fonction de $\mathbb{R}$ dans $[0,1]$, qui est définie par

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x \end{cases}$$

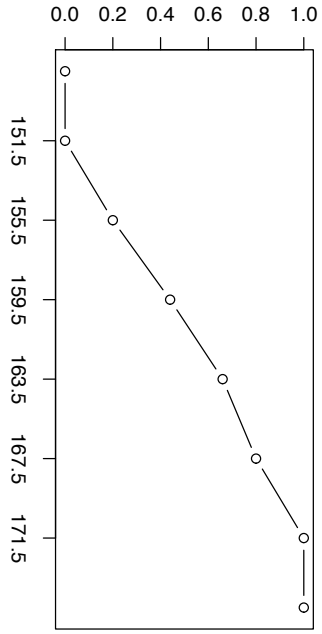


FIGURE 1.10 – Fonction de répartition d'une distribution groupée

En langage R

```
> y=c(0,0,cumsum(T5c/sum(T5c)),1)
> x=c(148,151.5,155.5,159.5,163.5,167.5,171.5,175)
> plot(x,y,type="b",xlab="",ylab="",xaxt = "n")
> axis(1, c(151.5,155.5,159.5,163.5,167.5,171.5))
```

Chapitre 2

Statistique descriptive univariée

2.1 Paramètres de position

2.1.1 Le mode

Le mode est la valeur distincte correspondant à l'effectif le plus élevé ; il est noté x_M.

Si on reprend la variable 'Etat civil' , dont le tableau statistique est le suivant :

x_j	n_j	f_j
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
	$n = 20$	1

le mode est C : célibataire.

Remarque 2.1
- Le mode peut être calculé pour tous les types de variable, quantitative et qualitative.
- Le mode n'est pas nécessairement unique.
- Quand une variable continue est découpée en classes, on peut définir une classe modale (classe correspondant à l'effectif le plus élevé).

2.1.2 La moyenne

La *moyenne* ne peut être définie que sur une variable *quantitative*.

La moyenne est la somme des valeurs observées divisée par leur nombre, elle est notée $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La moyenne peut être calculée à partir des valeurs distinctes et des effectifs

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j.$$

Exemple 2.1 Les nombres d'enfants de 8 familles sont les suivants $0, 0, 1, 1, 1, 2, 3, 4$. La moyenne est

$$\bar{x} = \frac{0+0+1+1+1+2+3+4}{8} = \frac{12}{8} = 1.5.$$

On peut aussi faire les calculs avec les valeurs distinctes et les effectifs. On considère le tableau :

x_j	n_j
0	2
1	3
2	1
3	1
4	1
	8

$$\begin{aligned} \bar{x} &= \frac{2\times 0 + 3\times 1 + 1\times 2 + 1\times 3 + 1\times 4}{8} \\ &= \frac{3+2+3+4}{8} \\ &= 1.5. \end{aligned}$$

Remarque 2.2 La moyenne n'est pas nécessairement une valeur possible.

En langage R

```
E=c(0,0,1,1,1,2,3,4)
n=length(E)
xb=sum(E)/n
xb
xb=mean(E)
xb
```

2.1.3 Remarques sur le signe de sommation $\sum$

Définition 2.1

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

1. En statistique les x_i sont souvent les valeurs observées.
2. L'indice est muet : $\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$.
3. Quand il n'y a pas de confusion possible, on peut écrire $\sum_i x_i$.

Exemple 2.2

1. $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$.
2. $\sum_{i=3}^{5} x_{i2} = x_{32} + x_{42} + x_{52}$.
3. $\sum_{i=1}^{3} i = 1 + 2 + 3 = 6$.
4. On peut utiliser plusieurs sommations emboîtées, mais il faut bien distinguer les indices :

$$\begin{aligned} \sum_{i=1}^{3}\sum_{j=1}^{2} x_{ij} &= x_{11} + x_{12} \qquad (i=1) \\ &+ x_{21} + x_{22} \qquad (i=2) \\ &+ x_{31} + x_{32} \qquad (i=3) \end{aligned}$$

5. On peut exclure une valeur de l'indice.

$$\sum_{\substack{i=1 \\ i\neq 3}}^{5} x_i = x_1 + x_2 + x_4 + x_5.$$

Propriété 2.1

1. Somme d'une constante

$$\sum_{i=1}^{n} a = \underbrace{a + a + \cdots + a}_{n \text{ fois}} = na \qquad \text{(a constante)}.$$

Exemple

$$\sum_{i=1}^{5} 3 = 3 + 3 + 3 + 3 + 3 = 5 \times 3 = 15.$$

2. Mise en évidence

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i \qquad \text{(a constante)}.$$

Exemple

$$\sum_{i=1}^{3} 2 \times i = 2(1 + 2 + 3) = 2 \times 6 = 12.$$

3. Somme des n premiers entiers

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

4. Distribution

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

5. Distribution

$$\sum_{i=1}^{n} (x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i.$$

Exemple (avec $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)

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

6. Somme de carrés

$$\sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i y_i + y_i^2) = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$$

C'est une application de la formule

$$(a - b)^2 = a^2 - 2ab + b^2.$$

2.1.4 Moyenne géométrique

Si $x_i \geq 0$, on appelle moyenne géométrique la quantité

$$G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 \times x_2 \times \cdots \times x_n)^{1/n}.$$

On peut écrire la moyenne géométrique comme l'exponentielle de la moyenne arithmétique des logarithmes des valeurs observées

$$G = \exp\log G = \exp\log\left(\prod_{i=1}^{n} x_i\right)^{1/n} = \exp\frac{1}{n}\log\prod_{i=1}^{n} x_i = \exp\frac{1}{n}\sum_{i=1}^{n}\log x_i.$$

La moyenne géométrique s'utilise, par exemple, quand on veut calculer la moyenne de taux d'intérêt.

Exemple 2.3 Supposons que les taux d'intérêt pour 4 années consécutives soient respectivement de $5, 10, 15$, et 10%. Que va-t-on obtenir après 4 ans si je place 100 francs ?

- Après 1 an on a, $100 \times 1.05 = 105$ Fr.
- Après 2 ans on a, $100 \times 1.05 \times 1.1 = 115.5$ Fr.
- Après 3 ans on a, $100 \times 1.05 \times 1.1 \times 1.15 = 132.825$ Fr.
- Après 4 ans on a, $100 \times 1.05 \times 1.1 \times 1.15 \times 1.1 = 146.1075$ Fr.

Si on calcule la moyenne arithmétique des taux on obtient

$$\bar{x} = \frac{1.05 + 1.10 + 1.15 + 1.10}{4} = 1.10.$$

Si on calcule la moyenne géométrique des taux, on obtient

$$G = (1.05 \times 1.10 \times 1.15 \times 1.10)^{1/4} = 1.099431377.$$

Le bon taux moyen est bien G et non $\bar{x}$, car si on applique 4 fois le taux moyen G aux 100 francs, on obtient

$$100 \text{ Fr} \times G^4 = 100 \times 1.099431377^4 = 146.1075 \text{ Fr}.$$

2.1.5 Moyenne harmonique

Si $x_i \geq 0$, on appelle moyenne harmonique la quantité

$$H = \frac{n}{\sum_{i=1}^{n} 1/x_i}.$$

Il est judicieux d'appliquer la moyenne harmonique sur des vitesses.

Exemple 2.4 Un cycliste parcourt 4 étapes de 100km. Les vitesses respectives pour ces étapes sont de 10 km/h, 30 km/h, 40 km/h, 20 km/h. Quelle a été sa vitesse moyenne ?

- Un raisonnement simple nous dit qu'il a parcouru la première étape en 10h, la deuxième en 3h20 la troisième en 2h30 et la quatrième en 5h. Il a donc parcouru le total des 400km en

  $$10 + 3h20 + 2h30 + 5h = 20h50 = 20.8333h,$$

  sa vitesse moyenne est donc

  $$\text{Moy} = \frac{400}{20.8333} = 19.2 \text{ km/h}.$$

- Si on calcule la moyenne arithmétique des vitesses, on obtient

  $$\bar{x} = \frac{10 + 30 + 40 + 20}{4} = 25 \text{ km/h}.$$

- Si on calcule la moyenne harmonique des vitesses, on obtient

  $$H = \frac{4}{\frac{1}{10} + \frac{1}{30} + \frac{1}{40} + \frac{1}{20}} = 19.2 \text{ km/h}.$$

  La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

Remarque 2.3 Il est possible de montrer que la moyenne harmonique est toujours inférieure ou égale à la moyenne géométrique qui est toujours inférieure ou égale à la moyenne arithmétique

$$H \leq G \leq \bar{x}.$$

2.1.6 Moyenne pondérée

Dans certains cas, on n'accorde pas le même poids à toutes les observations. Par exemple, si on calcule la moyenne des notes pour un programme d'étude, on peut pondérer les notes de l'étudiant par le nombre de crédits ou par le nombre d'heures de chaque cours. Si $w_i > 0, i = 1, \ldots, n$ sont les poids associés à chaque observation, alors la moyenne pondérée par w_i est définie par :

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Exemple 2.5 Supposons que les notes soient pondérées par le nombre de crédits, et que les notes de l'étudiant soient les suivantes :

Note	5	4	3	6	5
Crédits	6	3	4	3	4

La moyenne pondérée des notes par les crédits est alors

$$\bar{x}_w = \frac{6\times 5+3\times 4+4\times 3+3\times 6+4\times 5}{6+3+4+3+4} = \frac{30+12+12+18+20}{20} = \frac{92}{20} = 4.6.$$

2.1.7 La médiane

La médiane, notée $x_{1/2}$, est une valeur centrale de la série statistique obtenue de la manière suivante :

– On trie la série statistique par ordre croissant des valeurs observées. Avec la série observée :

$$3 \quad 2 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2,$$

on obtient :

$$0 \quad 0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3.$$

– La médiane $x_{1/2}$ est la valeur qui se trouve au milieu de la série ordonnée :

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3. \\ & & & \uparrow & & & & \end{array}$$

On note alors $x_{1/2} = 1$.

Nous allons examiner une manière simple de calculer la médiane. Deux cas doivent être distingués.

– Si n est impair, il n'y a pas de problème (ici avec $n = 7$), alors $x_{1/2} = 1$:

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3. \\ & & & \uparrow & & & & \end{array}$$

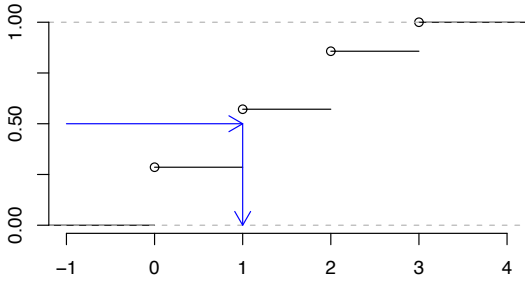
La Figure 2.1 montre la fonction de répartition de la série. La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur 1/2 :

$$x_{1/2} = F^{-1}(0.5).$$

En langage R

```
x=c(0 , 0 , 1 , 1 , 2 , 2 , 3)
median(x)
plot(ecdf(x),xlab="",ylab="",main="",frame=FALSE,yaxt = "n")
axis(2, c(0.0,0.25,0.50,0.75,1.00))
arrows(-1,0.5,1,0.50,length=0.14,col="blue")
arrows(1,0.50,1,0,length=0.14,col="blue")
```


FIGURE 2.1 – Médiane quand n est impair



– Si n est pair, deux valeurs se trouvent au milieu de la série (ici avec $n = 8$)

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 2 & 2 & 3 & 4 \\ & & & \uparrow & \uparrow & & & \end{array}$$

La médiane est alors la moyenne de ces deux valeurs :

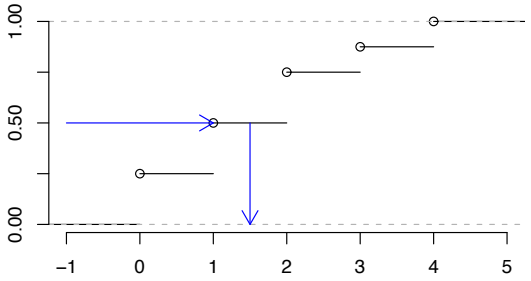
$$x_{1/2} = \frac{1+2}{2} = 1.5.$$

La Figure 2.2 montre la fonction de répartition de la série de taille paire. La médiane peut toujours être définie comme l'inverse de la fonction de répartition pour la valeur 1/2 :

$$x_{1/2} = F^{-1}(0.5).$$

Cependant, la fonction de répartition est discontinue par 'palier'. L'inverse de la répartition correspond exactement à un 'palier'.

FIGURE 2.2 – Médiane quand n est pair



En langage R

```
x=c(0 , 0 , 1 , 1 , 2 , 2 , 3 , 4)
median(x)
plot(ecdf(x),xlab="",ylab="",main="",frame=FALSE,yaxt = "n")
axis(2, c(0.0,0.25,0.50,0.75,1.00))
arrows(-1,0.5,1,0.50,length=0.14,col="blue")
arrows(1.5,0.50,1.5,0,,length=0.14,col="blue")
```

En général on note

$$x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(n)}$$

la série ordonnée par ordre croissant. On appelle cette série ordonnée la statistique d'ordre. Cette notation, très usuelle en statistique, permet de définir la médiane de manière très synthétique.

– Si n est impair

$$x_{1/2} = x_{\left(\frac{n+1}{2}\right)}$$

– Si n est pair

$$x_{1/2} = \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\}.$$

Remarque 2.4 La médiane peut être calculée sur des variables quantitatives et sur des variables qualitatives ordinales.

2.1.8 Quantiles

La notion de quantile d'ordre p (où $0 < p < 1$) généralise la médiane. Formellement un quantile est donné par l'inverse de la fonction de répartition :

$$x_p = F^{-1}(p).$$

Si la fonction de répartition était continue et strictement croissante, la définition du quantile serait sans équivoque. La fonction de répartition est cependant discontinue et "par palier". Quand la fonction de répartition est par palier, il existe au moins 9 manières différentes de définir les quantiles selon que l'on fasse ou non une interpolation de la fonction de répartition. Nous présentons une de ces méthodes, mais il ne faut pas s'étonner de voir les valeurs des quantiles différer légèrement d'un logiciel statistique à l'autre.

– Si np est un nombre entier, alors

$$x_p = \frac{1}{2}\left\{x_{(np)} + x_{(np+1)}\right\}.$$

– Si np n'est pas un nombre entier, alors

$$x_p = x_{(\lceil np \rceil)},$$

où $\lceil np \rceil$ représente le plus petit nombre entier supérieur ou égal à np.

Remarque 2.5

– La médiane est le quantile d'ordre $p = 1/2$.
– On utilise souvent
  $x_{1/4}$ le premier quartile,
  $x_{3/4}$ le troisième quartile,
  $x_{1/10}$ le premier décile ,
  $x_{1/5}$ le premier quintile,
  $x_{4/5}$ le quatrième quintile,
  $x_{9/10}$ le neuvième décile,
  $x_{0.05}$ le cinquième percentile ,
  $x_{0.95}$ le nonante-cinquième percentile.
– Si $F(x)$ est la fonction de répartition, alors $F(x_p) \geq p$.

Exemple 2.6 Soit la série statistique 12, 13, 15, 16, 18, 19, 22, 24, 25, 27, 28, 34 contenant 12 observations ($n = 12$).

– Le premier quartile : Comme $np = 0.25 \times 12 = 3$ est un nombre entier, on a

$$x_{1/4} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{15+16}{2} = 15.5.$$

– La médiane : Comme $np = 0.5 \times 12 = 6$ est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(6)} + x_{(7)}\right\} = (19+22)/2 = 20.5.$$

– Le troisième quartile : Comme $np = 0.75 \times 12 = 9$ est un nombre entier, on a

$$x_{3/4} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{25+27}{2} = 26.$$

En langage R

```
x=c(12,13,15,16,18,19,22,24,25,27,28,34)
quantile(x,type=2)
```

Exemple 2.7 Soit la série statistique 12, 13, 15, 16, 18, 19, 22, 24, 25, 27 contenant 10 observations ($n = 10$).

– Le premier quartile : Comme $np = 0.25 \times 10 = 2.5$ n'est pas un nombre entier, on a

$$x_{1/4} = x_{(\lceil 2.5 \rceil)} = x_{(3)} = 15.$$

– La médiane : Comme $np = 0.5 \times 10 = 5$ est un nombre entier, on a

$$x_{1/2} = \frac{1}{2}\left\{x_{(5)} + x_{(6)}\right\} = (18+19)/2 = 18.5.$$

– Le troisième quartile : Comme $np = 0.75 \times 10 = 7.5$ n'est pas un nombre entier, on a

$$x_{3/4} = x_{(\lceil 7.5 \rceil)} = x_{(8)} = 24.$$

En langage R

```
x=c(12,13,15,16,18,19,22,24,25,27)
quantile(x,type=2)
```

2.2 Paramètres de dispersion

2.2.1 L'étendue

L'*étendue* est simplement la différence entre la plus grande et la plus petite valeur observée.

$$\mathrm{E} = x_{(n)} - x_{(1)}.$$

2.2.2 La distance interquartile

La distance interquartile est la différence entre le troisième et le premier quartile :

$$IQ = x_{3/4} - x_{1/4}.$$

2.2.3 La variance

La *variance* est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations :

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

Théorème 2.1 *La variance peut aussi s'écrire*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{2.1}$$

Démonstration

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

□

La variance peut également être définie à partir des effectifs et des valeurs distinctes :

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(x_j - \bar{x})^2.$$

La variance peut aussi s'écrire

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2.$$

Quand on veut estimer une variance d'une variable X à partir d'un échantillon (une partie de la population sélectionnée au hasard) de taille n, on utilise la variance "corrigée" divisée par $n-1$.

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2\frac{n}{n-1}.$$

La plupart des logiciels statistiques calculent S_x^2 et non s_x^2.

2.2.4 L'écart-type

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

$$s_x = \sqrt{s_x^2}.$$

Quand on veut estimer l'écart-type d'une variable X partir d'un échantillon de taille n, utilise la variance "corrigée" pour définir l'écart type

$$S_x = \sqrt{S_x^2} = s_x\sqrt{\frac{n}{n-1}}.$$

La plupart des logiciels statistiques calculent S_x et non s_x.

Exemple 2.8 Soit la série statistique $2, 3, 4, 4, 5, 6, 7, 9$ de taille 8. On a

$$\bar{x} = \frac{2+3+4+4+5+6+7+9}{8} = 5,$$

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{8}\left[(2-5)^2+(3-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2+(9-5)^2\right] \\ &= \frac{1}{8}\left[9+4+1+1+0+1+4+16\right] \\ &= \frac{36}{8} \\ &= 4.5. \end{aligned}$$

On peut également utiliser la formule (2.1) de la variance, ce qui nécessite moins de calcul (surtout quand la moyenne n'est pas un nombre entier).

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2 \\ &= \frac{1}{8}(2^2+3^2+4^2+4^2+5^2+6^2+7^2+9^2)-5^2 \\ &= \frac{1}{8}(4+9+16+16+25+36+49+81)-25 \\ &= \frac{236}{8}-25 \\ &= 29.5-25=4.5. \end{aligned}$$

En langage R

```
> x=c(2,3,4,4,5,6,7,9)
> n=length(x)
> s2=sum((x-mean(x))^2)/n
> s2
[1] 4.5
> S2=s2*n/(n-1)
> S2
[1] 5.142857
> S2=var(x)
> S2
[1] 5.142857
> s=sqrt(s2)
> s
[1] 2.121320
> S=sqrt(S2)
> S
[1] 2.267787
> S=sd(x)
```



```
> S
[1] 2.267787
> E=max(x)-min(x)
> E
[1] 7
```

2.2.5 L'écart moyen absolu

L'*écart moyen absolu* est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observations :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|\,.$$

2.2.6 L'écart médian absolu

L'*écart médian absolu* est la somme des valeurs absolues des écarts à la médiane divisée par le nombre d'observations :

$$e_{med} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i-x_{1/2}\right|.$$

2.3 Moments

Définition 2.2 *On appelle moment à l'origine d'ordre $r \in \mathbb{N}$ le paramètre*

$$m'_r = \frac{1}{n}\sum_{i=1}^{n}x_i^r.$$

Définition 2.3 *On appelle moment centré d'ordre $r \in \mathbb{N}$ le paramètre*

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r.$$

Les moments généralisent la plupart des paramètres. On a en particulier

- $m'_1 = \bar{x}$,
- $m_1 = 0$,
- $m'_2 = \frac{1}{n}\sum_i x_i^2 = s_x^2+\bar{x}^2$,
- $m_2 = s_x^2$.

Nous verrons plus loin que des moments d'ordres supérieurs (r=3,4) sont utilisés pour mesurer la symétrie et l'aplatissement.

2.4 Paramètres de forme

2.4.1 Coefficient d'asymétrie de Fisher (skewness)

Le *moment centré d'ordre trois* est défini par

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Il peut prendre des valeurs positives, négatives ou nulles. L'asymétrie se mesure au moyen du coefficient d'asymétrie de Fisher

$$g_1 = \frac{m_3}{s_x^3},$$

où s_x^3 est le cube de l'écart-type.

2.4.2 Coefficient d'asymétrie de Yule

Le coefficient d'asymétrie de Yule est basé sur les positions des 3 quartiles (1er quartile, médiane et troisième quartile), et est normalisé par la distance interquartile :

$$A_Y = \frac{x_{3/4} + x_{1/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}}.$$

2.4.3 Coefficient d'asymétrie de Pearson

Le coefficient d'asymétrie de Pearson est basé sur une comparaison de la moyenne et du mode, et est standardisé par l'écart-type :

$$A_P = \frac{\bar{x} - x_M}{s_x}.$$

Tous les coefficients d'asymétrie ont les mêmes propriétés, ils sont nuls si la distribution est symétrique, négatifs si la distribution est allongée à gauche (left asymmetry), et positifs si la distribution est allongée à droite (right asymmetry) comme montré dans la Figure 2.3.

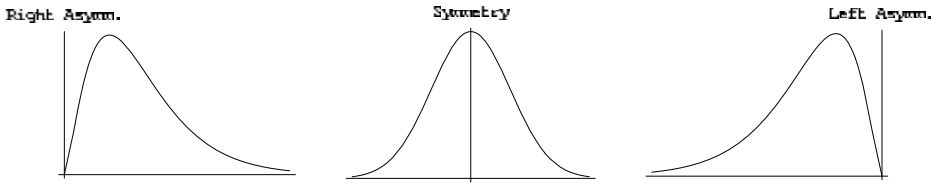


FIGURE 2.3 – Asymétrie d'une distribution

Remarque 2.6 Certaines variables sont toujours très asymétriques à droite, comme les revenus, les tailles des entreprises, ou des communes. Une méthode simple pour rendre une variable symétrique consiste alors à prendre le logarithme de cette variable.

2.5 Paramètre d'aplatissement (kurtosis)

L'aplatissement est mesuré par le coefficient d'aplatissement de Pearson

$$\beta_2 = \frac{m_4}{s_x^4},$$

ou le coefficient d'aplatissement de Fisher

$$g_2 = \beta_2 - 3 = \frac{m_4}{s_x^4} - 3,$$

où m_4 est le moment centré d'ordre 4, et s_x^4 est le carré de la variance.

- Une courbe mésokurtique si $g_2 \approx 0$.
- Une courbe leptokurtique si $g_2 > 0$. Elle est plus pointue et possède des queues plus longues.
- Une courbe platykurtique si $g_2 < 0$. Elle est plus arrondie et possède des queues plus courtes.

Dans la Figure 2.4, on présente un exemple de deux distributions de même moyenne et de même variance. La distribution plus pointue est leptokurtique, l'autre est mésokurtique. La distribution leptokurtique a une queue plus épaisse.

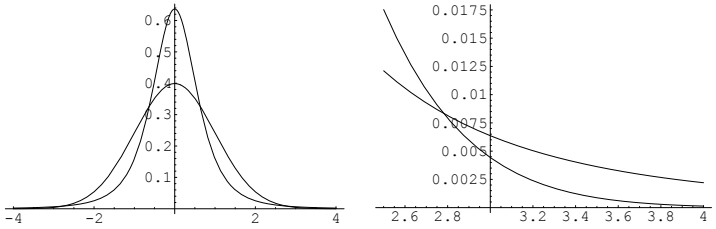


FIGURE 2.4 – Distributions mésokurtique et leptokurtique

2.6 Changement d'origine et d'unité

Définition 2.4 *On appelle changement d'origine l'opération consistant à ajouter (ou soustraire) la même quantité $a \in \mathbb{R}$ à toutes les observations*

$$y_i = a + x_i, i = 1, \ldots, n$$

Définition 2.5 *On appelle changement d'unité l'opération consistant à multiplier (ou diviser) par la même quantité $b \in \mathbb{R}$ toutes les observations*

$$y_i = bx_i, i = 1, \ldots, n.$$

Définition 2.6 *On appelle changement d'origine et d'unité l'opération consistant à multiplier toutes les observations par la même quantité $b \in \mathbb{R}$ puis à ajouter la même quantité $a \in \mathbb{R}$ à toutes les observations :*

$$y_i = a + bx_i, i = 1, \ldots, n.$$

Théorème 2.2 *Si on effectue un changement d'origine et d'unité sur une variable X, alors sa moyenne est affectée du même changement d'origine et d'unité.*

Démonstration Si $y_i = a + bx_i$, alors

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\frac{1}{n}\sum_{i=1}^{n}x_i = a + b\bar{x}.$$

□

Théorème 2.3 *Si on effectue un changement d'origine et d'unité sur une variable X, alors sa variance est affectée par le carré du changement d'unité et pas par le changement d'origine.*

Démonstration Si $y_i = a + bx_i$, alors

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^2 = b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 s_x^2.$$

□

Remarque 2.7

1. Les paramètres de position sont tous affectés par un changement d'origine et d'unité.
2. Les paramètres de dispersion sont tous affectés par un changement d'unité mais pas par un changement d'origine.
3. Les paramètres de forme et d'aplatissement ne sont affectés ni par un changement d'unité ni par un changement d'origine.

2.7 Moyennes et variances dans des groupes

Supposons que les n observations soient réparties dans deux groupes G_A et G_B. Les n_A premières observations sont dans le groupe G_A et les n_B dernières observations sont dans le groupe G_B, avec la relation

$$n_A + n_B = n.$$

On suppose que la série statistique contient d'abord les unités de G_A puis les unités de G_B :

$$\underbrace{x_1, x_2, \ldots, x_{n_A-1}, x_{n_A}}_{\text{observations de } G_A}, \underbrace{x_{n_A+1}, x_{n_A+2}, \ldots, x_{n-1}, x_n}_{\text{observations de } G_B}.$$

On définit les moyennes des deux groupes :

- la moyenne du premier groupe $\bar{x}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} x_i$,
- la moyenne du deuxième groupe $\bar{x}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n} x_i$.

La moyenne générale est une moyenne pondérée par la taille des groupes des moyennes des deux groupes. En effet

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n_A} x_i + \sum_{i=n_A+1}^{n} x_i\right) = \frac{1}{n}\left(n_A\bar{x}_A + n_B\bar{x}_B\right).$$

On peut également définir les variances des deux groupes :

- la variance du premier groupe $s_A^2 = \frac{1}{n_A}\sum_{i=1}^{n_A}(x_i - \bar{x}_A)^2$,
- la variance du deuxième groupe $s_B^2 = \frac{1}{n_B}\sum_{i=n_A+1}^{n}(x_i - \bar{x}_B)^2$.

Théorème 2.4 *(de Huygens) La variance totale, définie par*

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

se décompose de la manière suivante :

$$s_x^2 = \underbrace{\frac{n_A s_A^2 + n_B s_B^2}{n}}_{\textit{variance intra-groupes}} + \underbrace{\frac{n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2}{n}}_{\textit{variance inter-groupes}}.$$

Démonstration

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[\sum_{i=1}^{n_A}(x_i - \bar{x})^2 + \sum_{i=n_A+1}^{n}(x_i - \bar{x})^2\right] \qquad (2.2)$$

On note que

$$\begin{aligned}
&\sum_{i=1}^{n_A}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n_A}(x_i-\bar{x}_A+\bar{x}_A-\bar{x})^2 \\
&= \sum_{i=1}^{n_A}(x_i-\bar{x}_A)^2+\sum_{i=1}^{n_A}(\bar{x}_A-\bar{x})^2+\underbrace{2\sum_{i=1}^{n_A}(x_i-\bar{x}_A)(\bar{x}_A-\bar{x})}_{=0} \\
&= n_A s_A^2+n_A(\bar{x}_A-\bar{x})^2.
\end{aligned}$$

On a évidemment la même relation dans le groupe G_B :

$$\sum_{i=n_A+1}^{n}(x_i-\bar{x})^2 = n_B s_B^2+n_B(\bar{x}_B-\bar{x})^2.$$

En revenant à l'expression (2.2), on obtient

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\left[\sum_{i=1}^{n_A}(x_i-\bar{x})^2+\sum_{i=n_A+1}^{n}(x_i-\bar{x})^2\right] \\
&= \frac{1}{n}\left[n_A s_A^2+n_A(\bar{x}_A-\bar{x})^2+n_B s_B^2+n_B(\bar{x}_B-\bar{x})^2\right] \\
&= \frac{n_A s_A^2+n_B s_B^2}{n}+\frac{n_A(\bar{x}_A-\bar{x})^2+n_B(\bar{x}_B-\bar{x})^2}{n}.
\end{aligned}$$

□

2.8 Diagramme en tiges et feuilles

Le diagramme en tiges et feuilles ou *Stem and leaf diagram* est une manière rapide de présenter une variable quantitative. Par exemple, si l'on a la série statistique ordonnée suivante :

$$15, 15, 16, 17, 18, 20, 21, 22, 23, 23, 23, 24, 25, 25, 26,$$

$$26, 27, 28, 28, 29, 30, 30, 32, 34, 35, 36, 39, 40, 43, 44,$$

la tige du diagramme sera les dizaines et les feuilles seront les unités. On obtient le graphique suivant.

```
The decimal point is 1 digit(s) to the right of the |

  1 | 55678
  2 | 012333455667889
  3 | 0024569
  4 | 034
```


Ce diagramme permet d'avoir une vue synthétique de la distribution. Évidemment, les tiges peuvent être définies par les centaines, ou des millers, selon l'ordre de grandeur de la variable étudiée.

En langage R

```
#
# Diagramme en tige et feuilles
#
X=c(15,15,16,17,18,20,21,22,23,23,23,24,25,25,26,26,
27,28,28,29,30,30,32,34,35,36,39,40,43,44)
stem(X,0.5)
```

2.9 La boîte à moustaches

La boîte à moustaches, ou diagramme en boîte, ou encore *boxplot* en anglais, est un diagramme simple qui permet de représenter la distribution d'une variable. Ce diagramme est composé de :

- Un rectangle qui s'étend du premier au troisième quartile. Le rectangle est divisé par une ligne correspondant à la médiane.
- Ce rectangle est complété par deux segments de droites.
  - Pour les dessiner, on calcule d'abord les bornes

$$b^- = x_{1/4} - 1.5IQ \quad \text{et} \quad b^+ = x_{3/4} + 1.5IQ,$$

    où IQ est la distance interquartile.
  - On identifie ensuite la plus petite et la plus grande observation comprise entre ces bornes. Ces observations sont appelées "valeurs adjacentes".
  - On trace les segments de droites reliant ces observations au rectangle.
  - Les valeurs qui ne sont pas comprises entre les valeurs adjacentes, sont représentées par des points et sont appelées "valeurs extrêmes".

Exemple 2.9 On utilise une base de données de communes suisses de 2003 fournie par l'Office fédéral de la statistique (OFS) contenant un ensemble de variables concernant la population et l'aménagement du territoire. L'objectif est d'avoir un aperçu des superficies des communes du canton de Neuchâtel. On s'intéresse donc à la variable HApoly donnant la superficie en hectares des 62 communes neuchâteloises. La boîte à moustaches est présentée en Figure 2.5. L'examen du graphique indique directement une dissymétrie de la distribution, au sens où il y a beaucoup de petites communes et peu de grandes communes. Le graphique montre aussi que deux communes peuvent être considérées communes des points extrêmes, car elles ont plus de 3000 hectares. Il s'agit de la Brévine (4182ha) et de la Chaux-de-Fonds (5566ha).

En langage R

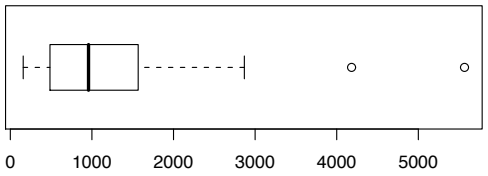


FIGURE 2.5 – Boîtes à moustaches pour la variable superficie en hectares (HApoly) des communes du canton de Neuchâtel

```
# Étape 1: installation du package sampling
#          dans lequel se trouve la base de données des communes belges
#          choisir "sampling" dans la liste
utils:::menuInstallPkgs()
# Etape 2: charge le package sampling
#          choisir "sampling" dans la liste
local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
# Utilisation des données
data(swissmunicipalities)
attach(swissmunicipalities)
# boxplot de la sélection des communes neuchâteloises
# le numéro du canton est 24
boxplot(HApoly[CT==24],horizontal=TRUE)
% selection des communes neuchâteloises de plus de 3000 HA
data.frame(Nom=Nom[HApoly>3000 & CT==24],Superficie=HApoly[HApoly>3000 & CT==24])
```

Exemple 2.10 On utilise une base de données belges fournie par l'Institut National (belge) de Statistique contenant des informations sur la population et les revenus des personnes physiques dans les communes. On s'intéresse à la variable "revenu moyen en euros par habitant en 2004" pour chaque commune (variable averageincome) et l'on aimerait comparer les 9 provinces belges : Anvers, Brabant, Flandre occidentale, Flandre orientale, Hainaut, Liège, Limboug, Luxembourg, Namur. La Figure 2.6 contient les boîtes à moustaches de chaque province. Les communes ont été triées selon les provinces belges. De ce graphique, on peut directement voir que la province du Brabant contient à la fois la commune la plus riche (Lasne) et la plus pauvre (Saint-Josse-ten-Noode). On voit également une dispersion plus importante dans la province du Brabant.

En langage R

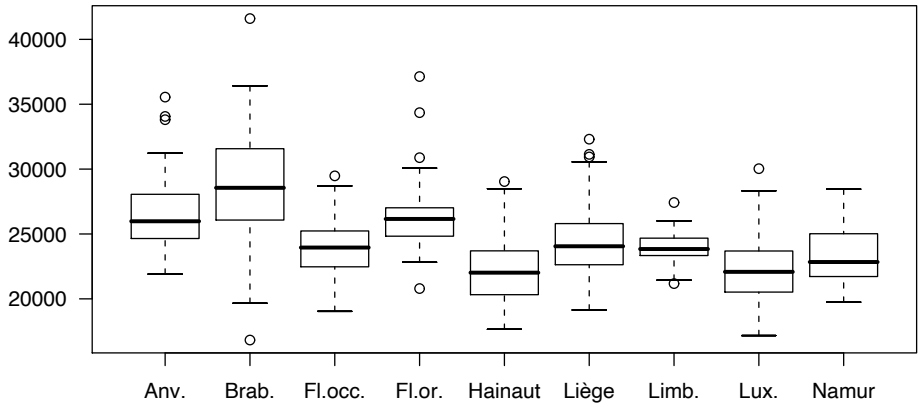


FIGURE 2.6 – Boîtes à moustaches du "revenu moyen des habitants" des communes selon les provinces belges

```
# Utilisation des données
data(belgianmunicipalities)
attach(belgianmunicipalities)
# Construction d'une liste avec les noms des provinces
b=list(
"Anv."=averageincome[Province==1],
"Brab."=averageincome[Province==2],
"Fl.occ."=averageincome[Province==3],
"Fl.or."=averageincome[Province==4],
"Hainaut"=averageincome[Province==5],
"Liège"=averageincome[Province==6],
"Limb."=averageincome[Province==7],
"Lux."=averageincome[Province==8],
"Namur"=averageincome[Province==9]
)
boxplot(b)
```

Exercices

Exercice 2.1 On pèse les 50 élèves d'une classe et nous obtenons les résultats résumés dans le tableau suivant :

43	43	43	47	48
48	48	48	49	49
49	50	50	51	51
52	53	53	53	54
54	56	56	56	57
59	59	59	62	62
63	63	65	65	67
67	68	70	70	70
72	72	73	77	77
81	83	86	92	93

1. De quel type est la variable poids ?
2. Construisez le tableau statistique en adoptant les classes suivantes : [40 ;45]]45 ;50]]50 ;55]]55 ;60]]60 ;65]]65 ;70]]70 ;80]]80 ;100]
3. Construisez l'histogramme des effectifs ainsi que la fonction de répartition.

Solution

1. La variable poids est de type quantitative continue.
2.

$[c_j^-, c_j^+]$	n_j	N_j	f_j	F_j
$[40; 45]$	3	3	0.06	0.06
$]45; 50]$	10	13	0.20	0.26
$]50; 55]$	8	21	0.16	0.42
$]55; 60]$	7	28	0.14	0.56
$]60; 65]$	6	34	0.12	0.68
$]65; 70]$	6	40	0.12	0.80
$]70; 80]$	5	45	0.10	0.90
$]80; 100]$	5	50	0.10	1.00
	50		1	

3.

Exercice 2.2 Calculez tous les paramètres (de position, de dispersion et de forme) à partir du tableau de l'exemple 1.7 sans prendre en compte les classes.

Solution

- Médiane : Comme n est pair,

$$x_{1/2} = \frac{1}{2}(x_{25} + x_{26}) = \frac{1}{2}(160 + 160) = 160.$$

- quantiles
  - Premier quartile :

$$x_{1/4} = x_{13} = 156$$

  - Deuxième quartile :

$$x_{3/4} = x_{38} = 165$$

- Étendue :

$$E = 171 - 152 = 19.$$

- Distance interquartile :

$$IQ = x_{3/4} - x_{1/4} = 165 - 156 = 9$$

- Variance :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{50} \times 1668 = 33,36.$$

- Écart type :

$$s_x = \sqrt{s_x^2} = 5,7758.$$

- Écart moyen absolu :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{50} \times 245,2 = 4,904.$$

- Écart médian absolu :

$$e_{med} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{1/2}| = \frac{1}{50} \times 242 = 4,84.$$

- Moment centré d'ordre trois :

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{50} \times 2743,2 = 54,864.$$

Exercice 2.3

1. Montrez que

$$s_x^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

2. Montrez que

$$s_x \leq E_t \sqrt{\frac{n-1}{2n}}.$$

3. Montrez que, si $x_i > 0$,

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \leq 2\bar{x}.$$

Solution

1.

$$\begin{aligned}
&\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^2 + x_j^2 - 2x_i x_j) \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i^2 + \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_j^2 - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}2x_i x_j \\
&= \frac{1}{2n}\sum_{i=1}^{n}x_i^2 + \frac{1}{2n}\sum_{j=1}^{n}x_j^2 - \frac{1}{n}\sum_{i=1}^{n}x_i\frac{1}{n}\sum_{j=1}^{n}x_j \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}x_i\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 \\
&= s_x^2.
\end{aligned}$$

2.

$$\begin{aligned}
s_x^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(x_i - x_j)^2 \\
&\leq \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(x_{(1)} - x_{(n)})^2 \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}E_t^2 \\
&= \frac{1}{2n^2}n(n-1)E_t^2 \\
&= \frac{n-1}{2n}E_t^2.
\end{aligned}$$

Donc,

$$s_x \leq E\sqrt{\frac{n-1}{2n}}.$$

Chapitre 3

Statistique descriptive bivariée

3.1 Série statistique bivariée

On s'intéresse à deux variables x et y. Ces deux variables sont mesurées sur les n unités d'observation. Pour chaque unité, on obtient donc deux mesures. La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu :

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Chacune des deux variables peut être, soit quantitative, soit qualitative. On examine deux cas.

- Les deux variables sont quantitatives.
- Les deux variables sont qualitatives.

3.2 Deux variables quantitatives

3.2.1 Représentation graphique de deux variables

Dans ce cas, chaque couple est composé de deux valeurs numériques. Un couple de nombres (entiers ou réels) peut toujours être représenté comme un point dans un plan

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Exemple 3.1 On mesure le poids Y et la taille X de 20 individus.

y_i	x_i	y_i	x_i
60	155	75	180
61	162	76	175
64	157	78	173
67	170	80	175
68	164	85	179
69	162	90	175
70	169	96	180
70	170	96	185
72	178	98	189
73	173	101	187

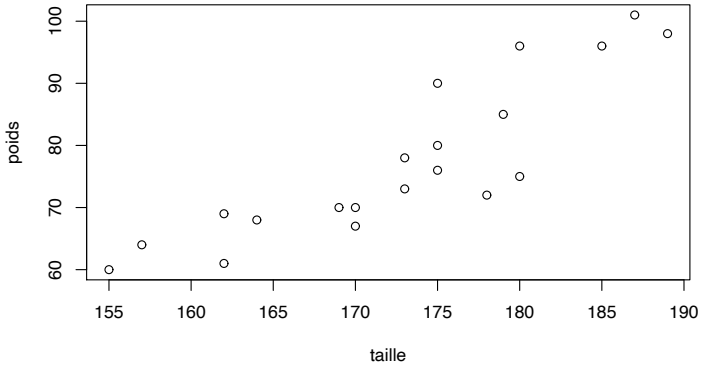


FIGURE 3.1 – Le nuage de points

En langage R

```
# nuage de points
poids=c(60,61,64,67,68,69,70,70,72,73,75,76,78,80,85,90,96,96,98,101)
taille=c(155,162,157,170,164,162,169,170,178,173,180,175,173,175,179,175,180,185,18
plot(taille,poids)
```

3.2.2 Analyse des variables

Les variables x et y peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Ces paramètres sont appelés *paramètres marginaux* : *variances marginales*, *moyennes marginales*, *écarts-types marginaux*, quantiles marginaux, etc.. . .

3.2.3 Covariance

La *covariance* est définie

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Remarque 3.1

- La covariance peut prendre des valeurs positives, négatives ou nulles.
- Quand $x_i = y_i$, pour tout $i = 1, \ldots, n$, la covariance est égale à la variance.

Théorème 3.1 *La covariance peut également s'écrire :*

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Démonstration

$$\begin{aligned}
s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - y_i\bar{x} - \bar{y}x_i + \bar{x}\bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\bar{x} - \frac{1}{n}\sum_{i=1}^{n}\bar{y}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.
\end{aligned}$$

□

3.2.4 Corrélation

Le *coefficient de corrélation* est la covariance divisée par les deux écart-types marginaux :

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Le *coefficient de détermination* est le carré du coefficient de corrélation :

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Remarque 3.2

- Le coefficient de corrélation mesure la dépendance linéaire entre deux variables :
- $-1 \leq r_{xy} \leq 1$,
- $0 \leq r_{xy}^2 \leq 1$.
- Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante.
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante.
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.

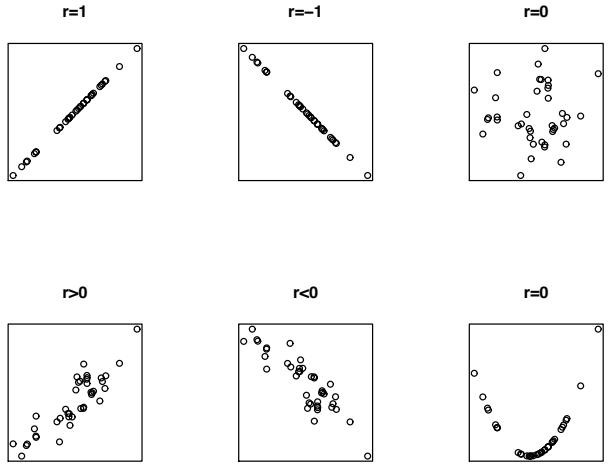


FIGURE 3.2 – Exemples de nuages de points et coefficients de corrélation

3.2.5 Droite de régression

La *droite de régression* est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés.

On considère que la variable X est explicative et que la variable Y est dépendante. L'équation d'une droite est

$$y = a + bx.$$

Le problème consiste à identifier une droite qui ajuste bien le nuage de points. Si les coefficients a et b étaient connus, on pourrait calculer les résidus de la régression définis par :

$$e_i = y_i - a - bx_i.$$

Le résidu e_i est l'erreur que l'on commet (voir Figure 3.3) en utilisant la droite de régression pour prédire y_i à partir de x_i. Les résidus peuvent être positifs ou négatifs.

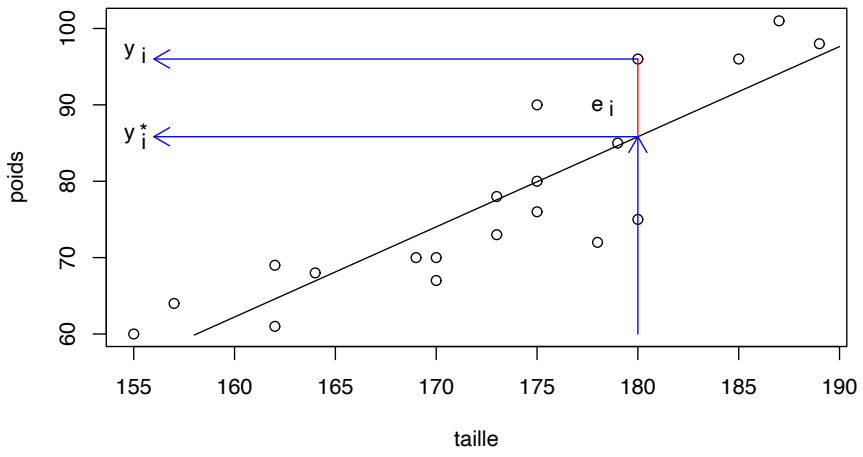


FIGURE 3.3 – Le nuage de points, le résidu

En langage R

```
# Graphique avec le résidus
plot(taille,poids)
segments(158,a+b*158,190,a+b*190)
segments(180,a+b*180,180,96,col="red")
#
text(178,90,expression(e))
text(178.7,89.5,"i")
#
arrows(180,a+b*180,156,a+b*180,col="blue",length=0.14)
arrows(180,60,180,a+b*180,col="blue",length=0.14)
arrows(180,96,156,96,col="blue",length=0.14)
#
text(154.8,86,expression(y))
text(155.5,85.5,"i")
#
text(154.8,97,expression(y))
text(155.5,97.8,"*")
text(155.5,96.5,"i")
```

Pour déterminer la valeur des coefficients a et b on utilise le principe des *moindres carrés* qui consiste à chercher la droite qui minimise la somme des carrés des résidus :

$$M(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

Théorème 3.2 *Les coefficients a et b qui minimisent le critère des moindres carrés sont donnés par :*

$$b = \frac{s_{xy}}{s_x^2} \quad et \quad a = \bar{y} - b\bar{x}.$$

Démonstration Le minimum $M(a,b)$ en (a,b) s'obtient en annulant les dérivées partielles par rapport à a et b.

$$\begin{cases} \dfrac{\partial M(a,b)}{\partial a} = -\displaystyle\sum_{i=1}^{n} 2\,(y_i - a - bx_i) = 0 \\ \dfrac{\partial M(a,b)}{\partial b} = -\displaystyle\sum_{i=1}^{n} 2\,(y_i - a - bx_i)\,x_i = 0 \end{cases}$$

On obtient un système de deux équations à deux inconnues. En divisant les deux équations par $-2n$, on obtient :

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (y_i - a - bx_i)\,x_i = 0, \end{cases}$$

ou encore

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} a - b\frac{1}{n}\sum_{i=1}^{n} x_i = 0 \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} a x_i - \frac{1}{n}\sum_{i=1}^{n} b x_i^2 = 0, \end{cases}$$

ce qui s'écrit aussi

$$\begin{cases} \bar{y} = a + b\bar{x} \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - a\bar{x} - \frac{1}{n}\sum_{i=1}^{n} b x_i^2 = 0. \end{cases}$$

La première équation montre que la droite passe par le point $(\bar{x}, \bar{y})$. On obtient

$$a = \bar{y} - b\bar{x}.$$

En remplaçant a par $\bar{y} - b\bar{x}$ dans la seconde équation, on a

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y} - b\bar{x})\bar{x} - b\frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ = & \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - b\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) \\ = & s_{xy} - b s_x^2 \\ = & 0, \end{aligned}$$

ce qui donne

$$s_{xy} - b s_x^2 = 0.$$

Donc

$$b = \frac{s_{xy}}{s_x^2}.$$

On a donc identifié les deux paramètres

$$\begin{cases} b = \frac{s_{xy}}{s_x^2} \text{ (la pente)} \\ a = \bar{y} - b\bar{x} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \text{ (la constante).} \end{cases}$$

On devrait en outre vérifier qu'il s'agit bien d'un minimum en montrant que la matrice des dérivées secondes est définie positive. □

La droite de régression est donc

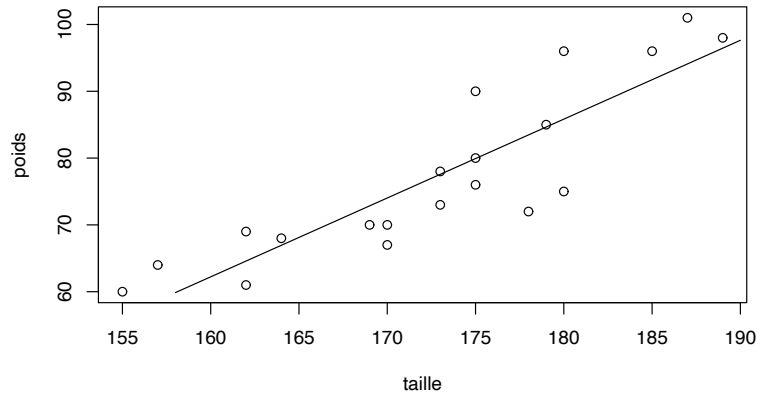
$$y = a + bx = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x,$$

ce qui peut s'écrire aussi

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Remarque 3.3 La droite de régression de y en x n'est pas la même que la droite de régression de x en y.

FIGURE 3.4 – La droite de régression



3.2.6 Résidus et valeurs ajustées

Les *valeurs ajustées* sont obtenues au moyen de la droite de régression :

$$y_i^* = a + bx_i.$$

Les valeurs ajustées sont les 'prédictions' des y_i réalisées au moyen de la variable x et de la droite de régression de y en x.

Remarque 3.4 La moyenne des valeurs ajustées est égale à la moyenne des valeurs observées $\bar{y}$. En effet,

$$\frac{1}{n}\sum_{i=1}^{n} y_i^* = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}.$$

Or, $\bar{y} = a + b\bar{x}$, car le point $(\bar{x}, \bar{y})$ appartient à la droite de régression.

Les résidus sont les différences entre les valeurs observées et les valeurs ajustées de la variable dépendante.

$$e_i = y_i - y_i^*.$$

Les résidus représentent la partie inexpliquée des y_i par la droite de régression.

Remarque 3.5

– La moyenne des résidus est nulle. En effet

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*) = \bar{y} - \bar{y} = 0.$$

– De plus,

$$\sum_{i=1}^{n} x_i e_i = 0.$$

La démonstration est un peu plus difficile.

3.2.7 Sommes de carrés et variances

Définition 3.1 *On appelle somme des carrés totale la quantité*

$$\mathrm{SC}_{TOT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

La variance marginale peut alors être définie par

$$s_y^2 = \frac{\mathrm{SC}_{TOT}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Définition 3.2 *On appelle somme des carrés de la régression la quantité*

$$\mathrm{SC}_{REGR} = \sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

Définition 3.3 *La variance de régression est la variance des valeurs ajustées.*

$$s_{y^*}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

Définition 3.4 *On appelle somme des carrés des résidus (ou résiduelle) la quantité*

$$\mathrm{SC}_{RES} = \sum_{i=1}^{n} e_i^2.$$

Définition 3.5 *La variance résiduelle est la variance des résidus.*

$$s_e^2 = \frac{\mathrm{SC}_{RES}}{n} = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

Note : Il n'est pas nécessaire de centrer les résidus sur leurs moyennes pour calculer la variance, car la moyenne des résidus est nulle.

Théorème 3.3

$$\mathrm{SC}_{TOT} = \mathrm{SC}_{REGR} + \mathrm{SC}_{RES}.$$

Démonstration

$$\begin{aligned}
\mathrm{SC}_{TOT} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - y_i^* + y_i^* - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) \\
&= \mathrm{SC}_{RES} + \mathrm{SC}_{REGR} + 2\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}).
\end{aligned}$$

Le troisième terme est nul. En effet,

$$\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) = \sum_{i=1}^{n}(y_i - a - bx_i)(a + bx_i - \bar{y})$$

En remplaçant a par $\bar{y} - b\bar{x}$, on obtient

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) &= \sum_{i=1}^{n}\left[y_i - \bar{y} - b(x_i - \bar{x}))\right] b(x_i - \bar{x}) \\
&= \sum_{i=1}^{n}\left[(y_i - \bar{y}) - b(x_i - \bar{x})\right] b(x_i - \bar{x}) \\
&= b\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - b^2\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\
&= bns_{xy} - b^2 ns_x^2 \\
&= \frac{s_{xy}}{s_x^2} ns_{xy} - \frac{s_{xy}^2}{s_x^4} ns_x^2 \\
&= 0.
\end{aligned}$$

□

3.2.8 Décomposition de la variance

Théorème 3.4 *La variance de régression peut également s'écrire*

$$s_{y^*}^2 = s_y^2 r^2,$$

où r^2 est le coefficient de détermination.

Démonstration

$$\begin{aligned} s^2_{y^*} &= \frac{1}{n}\sum_{i=1}^n (y_i^* - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ \bar{y} + \frac{s_{xy}}{s_x^2}(x_i - \bar{x}) - \bar{y} \right\}^2 \\ &= \frac{s_{xy}^2}{s_x^4}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{s_{xy}^2}{s_x^2} \\ &= s_y^2 \frac{s_{xy}^2}{s_x^2 s_y^2} \\ &= s_y^2 r^2. \end{aligned}$$

□

La *variance résiduelle* est la variance des résidus.

$$s_e^2 = \frac{1}{n}\sum_{i=1}^n e_i^2.$$

Théorème 3.5 *La variance résiduelle peut également s'écrire*

$$s_e^2 = s_y^2(1 - r^2),$$

où r^2 est le coefficient de détermination.

Démonstration

$$\begin{aligned} s_e^2 &= \frac{1}{n}\sum_{i=1}^n e_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n (y_i - y_i^*)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ y_i - \bar{y} - \frac{s_{xy}}{s_x^2}(x_i - \bar{x}) \right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 - 2\frac{s_{xy}}{s_x^2}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= s_y^2 + \frac{s_{xy}^2}{s_x^2} - 2\frac{s_{xy}^2}{s_x^2} \\ &= s_y^2 \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right). \end{aligned}$$

□

Théorème 3.6 *La variance marginale est la somme de la variance de régression et de la variance résiduelle,*

$$s_y^2 = s_{y^*}^2 + s_e^2.$$

La démonstration découle directement des deux théorèmes précédents.

3.3 Deux variables qualitatives

3.3.1 Données observées

Si les deux variables x et y sont qualitatives, alors les données observées sont une suite de couples de variables

$$(x_1, y_1), \ldots, (x_i, y_j), \ldots, (x_n, y_n),$$

chacune des deux variables prend comme valeurs des modalités qualitatives.

Les valeurs distinctes de x et y sont notées respectivement

$$x_1, \ldots, x_j, \ldots, x_J$$

et

$$y_1, \ldots, y_k, \ldots, y_K.$$

3.3.2 Tableau de contingence

Les données observées peuvent être regroupées sous la forme d'un *tableau de contingence*

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jK}	$n_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	n_{J1}	$\cdots$	n_{Jk}	$\cdots$	n_{JK}	$n_{J.}$
total	$n_{.1}$	$\cdots$	$n_{.k}$		$n_{.K}$	n

Les $n_{j.}$ et $n_{.k}$ sont appelés les effectifs marginaux. Dans ce tableau,

- $n_{j.}$ représente le nombre de fois que la modalité x_j apparaît,
- $n_{.k}$ représente le nombre de fois que la modalité y_k apparaît,
- n_{jk} représente le nombre de fois que les modalités x_j et y_k apparaissent ensemble.

On a les relations

$$\sum_{j=1}^J n_{jk} = n_{.k}, \text{ pour tout } k = 1, \ldots, K,$$

$$\sum_{k=1}^K n_{jk} = n_{j.}, \text{ pour tout } j = 1, \ldots, J,$$

et

$$\sum_{j=1}^{J} n_{j.} = \sum_{k=1}^{K} n_{.k} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = n \qquad .$$

Exemple 3.2 On s'intéresse à une éventuelle relation entre le sexe de 200 personnes et la couleur des yeux. Le Tableau 3.1 reprend le tableau de contingence.

TABLE 3.1 – Tableau des effectifs n_{jk}

	Bleu	Vert	Marron	Total
Homme	10	50	20	80
Femme	20	60	40	120
Total	30	110	60	200

3.3.3 Tableau des fréquences

Le *tableau de fréquences* s'obtient en divisant tous les effectifs par la taille de l'échantillon :

$$f_{jk} = \frac{n_{jk}}{n}, j = 1, \ldots, J, k = 1, \ldots, K$$

$$f_{j.} = \frac{n_{j.}}{n}, j = 1, \ldots, J,$$

$$f_{.k} = \frac{n_{.k}}{n}, k = 1, \ldots, K.$$

Le tableau des fréquences est

	y_1	$\cdots$	y_k	$\cdots$	y_K	total
x_1	f_{11}	$\cdots$	f_{1k}	$\cdots$	f_{1K}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_j	f_{j1}	$\cdots$	f_{jk}	$\cdots$	f_{jK}	$f_{j.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_J	f_{J1}	$\cdots$	f_{Jk}	$\cdots$	f_{JK}	$f_{J.}$
total	$f_{.1}$	$\cdots$	$f_{.k}$		$f_{.K}$	1

Exemple 3.3 Le Tableau 3.2 reprend le tableau des fréquences.

TABLE 3.2 – Tableau des fréquences

	Bleu	Vert	Marron	Total
Homme	0.05	0.25	0.10	0.40
Femme	0.10	0.30	0.20	0.60
Total	0.15	0.55	0.30	1.00

3.3.4 Profils lignes et profils colonnes

Un tableau de contingence s'interprète toujours en comparant des fréquences en lignes ou des fréquences en colonnes (appelés aussi *profils lignes* et *profils colonnes*).

Les profils lignes sont définis par

$$f_k^{(j)} = \frac{n_{jk}}{n_{j.}} = \frac{f_{jk}}{f_{j.}}, k = 1, \ldots, K, j = 1, \ldots, J,$$

et les profils colonnes par

$$f_j^{(k)} = \frac{n_{jk}}{n_{.k}} = \frac{f_{jk}}{f_{.k}}, j = 1, \ldots, J, k = 1, \ldots, K.$$

Exemple 3.4 Le Tableau 3.3 reprend le tableau des profils lignes, et le Tableau 3.4 reprend le tableau des profils colonnes.

TABLE 3.3 – Tableau des profils lignes

	Bleu	Vert	Marron	Total
Homme	0.13	0.63	0.25	1.00
Femme	0.17	0.50	0.33	1.00
Total	0.15	0.55	0.30	1.00

TABLE 3.4 – Tableau des profils colonnes

	Bleu	Vert	Marron	Total
Homme	0.33	0.45	0.33	0.40
Femme	0.67	0.55	0.67	0.60
Total	1.00	1.00	1.00	1.00

3.3.5 Effectifs théoriques et khi-carré

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien, on construit un tableau d'effectifs théoriques qui représente la situation où les variables ne sont pas liées (indépendance). Ces *effectifs théoriques* sont construits de la manière suivante :

$$n^*_{jk} = \frac{n_{j.} n_{.k}}{n}.$$

Les effectifs observés n_{jk} ont les mêmes marges que les effectifs théoriques n^*_{jk}.

Enfin, les *écarts à l'indépendance* sont définis par

$$e_{jk} = n_{jk} - n^*_{jk}.$$

– La dépendance du tableau se mesure au moyen du khi-carré défini par

$$\chi^2_{obs} = \sum_{k=1}^{K}\sum_{j=1}^{J} \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}} = \sum_{k=1}^{K}\sum_{j=1}^{J} \frac{e^2_{jk}}{n^*_{jk}}. \tag{3.1}$$

– Le khi-carré peut être normalisé pour ne plus dépendre du nombre d'observations. On définit le phi-deux par :

$$\phi^2 = \frac{\chi^2_{obs}}{n}.$$

Le ϕ^2 ne dépend plus du nombre d'observations. Il est possible de montrer que

$$\phi^2 \leq \min(J-1, K-1).$$

– Le V de Cramer est définit par

$$V = \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{\chi^2_{obs}}{n \min(J-1, K-1)}}.$$

Le V de Cramer est compris entre 0 et 1. Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si $V \approx 0$, les deux variables sont indépendantes. Si $V = 1$, il existe une relation fonctionnelle entre les variables, ce qui signifie que chaque ligne et chaque colonne du tableau de contingence ne contiennent qu'un seul effectif différent de 0 (il faut que le tableau ait le même nombre de lignes que de colonnes).

Exemple 3.5 Le Tableau 3.5 reprend le tableau des effectifs théoriques, le Tableau 3.6 reprend le tableau des écarts à l'indépendance. Enfin, les e^2_{jk}/n^*_{jk} sont présentés dans le tableau 3.7.

– Le khi-carré observé vaut $\chi^2_{obs} = 3.03$.
– Le phi-deux vaut $\phi^2 = 0.01515$.
– Comme le tableau a deux lignes $\min(J-1, K-1) = \min(2-1, 3-1) = 1$. Le V de Cramer est égal à $\sqrt{\phi^2}$.

TABLE 3.5 – Tableau des effectifs théoriques n^*_{jk}

	Bleu	Vert	Marron	Total
Homme	12	44	24	80
Femme	18	66	36	120
Total	30	110	60	200

TABLE 3.6 – Tableau des écarts à l'indépendance e_{jk}

	Bleu	Vert	Marron	Total
Homme	-2	6	-4	0
Femme	2	-6	4	0
Total	0	0	0	0

TABLE 3.7 – Tableau des e^2_{jk}/n^*_{jk}

	Bleu	Vert	Marron	Total
Homme	0.33	0.82	0.67	1.82
Femme	0.22	0.55	0.44	1.21
Total	0.56	1.36	1.11	3.03

– On a $V = 0.123$. La dépendance entre les deux variables est très faible.

En langage R

```
yeux= c(rep("bleu",times=10),rep("vert",times=50),rep("marron",times=20),
      rep("bleu",times=20),rep("vert",times=60),rep("marron",times=40))
sexe= c(rep("homme",times=80),rep("femme",times=120))
yeux=factor(yeux,levels=c("bleu","vert","marron"))
sexe=factor(sexe,levels=c("homme","femme"))
T=table(sexe,yeux)
T
plot(T,main="")
summary(T)
```

Exemple 3.6 Le tableau suivant est extrait de Boudon (1979, p. 57). La variable X est le niveau d'instruction du fils par rapport au père (plus élevé,

égal, inférieur), et la variable Y est le statut professionnel du fils par rapport au père (plus élevé, égal, inférieur).

TABLE 3.8 – Tableau de contingence : effectifs n_{jk}

Niveau d'instruction du fils par rapport au père	Statut professionnel du fils par rapport au père			
	Plus élevé	Egal	inférieur	total
plus élevé	134	96	61	291
égal	23	33	24	80
inférieur	7	16	22	45
total	164	145	107	416

TABLE 3.9 – Tableau des fréquences f_{jk}

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	0.322	0.231	0.147	0.700
égal	0.055	0.079	0.058	0.192
inférieur	0.017	0.038	0.053	0.108
total	0.394	0.349	0.257	1.000

TABLE 3.10 – Tableau des profils lignes

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	0.460	0.330	0.210	1
égal	0.288	0.413	0.300	1
inférieur	0.156	0.356	0.489	1
total	0.394	0.349	0.257	1

TABLE 3.11 – Tableau des profils colonnes

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	0.817	0.662	0.570	0.700
égal	0.140	0.228	0.224	0.192
inférieur	0.043	0.110	0.206	0.108
total	1	1	1	1

TABLE 3.12 – Tableau des effectifs théoriques n^*_{jk}

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	114.72	101.43	74.85	291
égal	31.54	27.88	20.58	80
inférieur	17.74	15.69	11.57	45
total	164	145	107	416

TABLE 3.13 – Tableau des écarts à l'indépendance e_{jk}

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	19.28	−5.43	−13.85	0
égal	−8.54	5.12	3.42	0
inférieur	−10.74	0.31	10.43	0
total	0	0	0	0

TABLE 3.14 – Tableau des e^2_{jk}/n^*_{jk}

$X\backslash Y$	Plus élevé	Egal	inférieur	total
plus élevé	3.24	0.29	2.56	6.09
égal	2.31	0.94	0.57	3.82
inférieur	6.50	0.01	9.39	15.90
total	12.05	1.24	12.52	$\chi^2_{obs} = 25.81$

On a donc

$$\begin{aligned} \chi^2_{obs} &= 25.81 \\ \phi^2 &= \frac{\chi^2_{obs}}{n} = \frac{25.81}{416} = 0.062 \\ V &= \sqrt{\frac{\phi^2}{\min(J-1, K-1)}} = \sqrt{\frac{0.062}{2}} = 0.176. \end{aligned}$$

Exercices

Exercice 3.1 La consommation de crèmes glacées par individus a été mesurée pendant 30 périodes. L'objectif est déterminé si la consommation dépend de la température. Les données sont dans le tableau 3.15. On sait en outre que

TABLE 3.15 – Consommation de crèmes glacées

consommation y	température x	consommation y	température x	consommation y	température x
386	41	286	28	319	44
374	56	298	26	307	40
393	63	329	32	284	32
425	68	318	40	326	27
406	69	381	55	309	28
344	65	381	63	359	33
327	61	470	72	376	41
288	47	443	72	416	52
269	32	386	67	437	64
256	24	342	60	548	71

$$\sum_{i=i}^{n} y_i = 10783, \qquad \sum_{i=i}^{n} x_i = 1473,$$

$$\sum_{i=i}^{n} y_i^2 = 4001293, \qquad \sum_{i=i}^{n} x_i^2 = 80145,$$

$$\sum_{i=i}^{n} x_i y_i = 553747,$$

1. Donnez les moyennes marginales, les variances marginales et la covariance entre les deux variables.
2. Donnez la droite de régression, avec comme variable dépendante la consommation de glaces et comme variable explicative la température.
3. Donnez la valeur ajustée et le résidu pour la première observation du tableau 3.15.

Solution

$$\bar{y} = 359.4333333, \bar{x} = 49.1,$$

$$\sigma_y^2 = 4184.112222, \sigma_x^2 = 260.69, \sigma_{xy}^2 = 810.0566667,$$

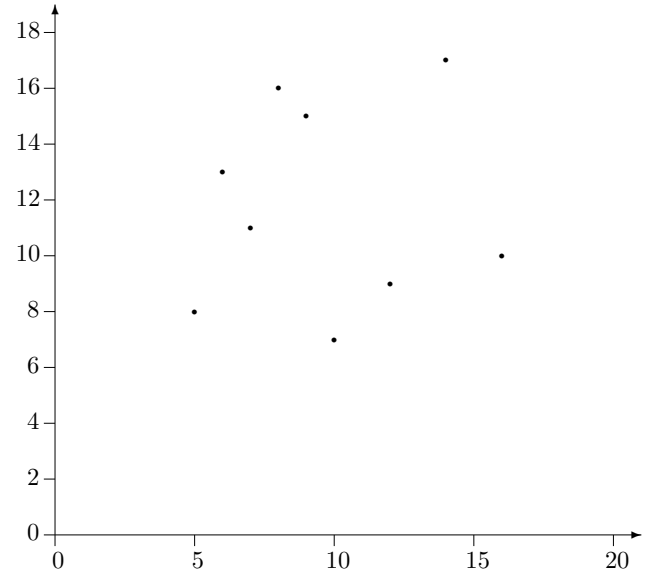
$$\rho = 0.77562456, b = 3.107356119, a = 206.8621479, y_1^* = 334.2637488, e_1 = 51.73625123,$$

Exercice 3.2 Neuf étudiants émettent un avis pédagogique vis-à-vis d'un professeur selon une échelle d'appréciation de 1 à 20. On relève par ailleurs la note obtenue par ces étudiants l'année précédente auprès du professeur.

	Etudiants								
y = Avis	5	7	16	6	12	14	10	9	8
x = Résultat	8	11	10	13	9	17	7	15	16

1. Représentez graphiquement les deux variables.
2. Déterminez le coefficient de corrélation entre les variables X et Y. Ensuite, donnez une interprétation de ce coefficient.
3. Déterminez la droite de régression Y en fonction de X.
4. Établissez, sur base du modèle, l'avis pour un étudiant ayant obtenu 12/20.
5. Calculez la variance résiduelle et le coefficient de détermination.

Solution



y_i	x_i	y_i^2	x_i^2	x_iy_i
5	8	25	64	40
7	11	49	121	77
16	10	256	100	160
6	13	36	169	78
12	9	144	81	108
14	17	196	289	238
10	7	100	49	70
9	15	81	225	135
8	16	64	256	128
87	106	951	1354	1034

$$\bar{y} = \frac{87}{9} = 9,667$$

$$s_y^2 = \frac{951}{9} - 9,667^2 = 12,22$$

$$\bar{x} = \frac{106}{9} = 11,78$$

$$s_x^2 = \frac{1354}{9} - 11,78^2 = 11,73$$

$$s_{xy} = \frac{1034}{9} - 9,667 \times 11,78 = 1,037$$

$$r_{xy} = \frac{1,037}{\sqrt{1}2,22\ \ 11,73} = 0.087$$

Ajustement linéaire de y en x

$$D_{y|x}\ : y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

$$D_{y|x}\ : y = 0.088x + 8,625$$

Valeur ajustée pour une cote de 12/20, (x=12)

$$y = 0.088 \times 12 + 8,625 = 9,686$$

Mesure de la qualité du modèle :
Variance résiduelle

$$\begin{aligned} s_{y|x}^2 &= s_y^2(1 - r^2) \\ &= 12,22(1 - 0.087^2) \\ &= 12,13 \text{ à comparer avec } s_y^2 = 12,22 \end{aligned}$$

Coefficient de détermination

$$r^2 = 0.087^2 = 0.008$$

ce coefficient représente la proportion de variance expliquée par le modèle (ici 0.8% faible).

Exercice 3.3 Considérons un échantillon de 10 fonctionnaires (ayant entre 40 et 50 ans) d'un ministère. Soit X le nombre d'années de service et Y le nombre de jours d'absence pour raison de maladie (au cours de l'année précédente) déterminé pour chaque personne appartenant à cet échantillon.

x_i	2	14	16	8	13	20	24	7	5	11
y_i	3	13	17	12	10	8	20	7	2	8

1. Représentez le nuage de points.
2. Calculez le coefficient de corrélation entre X et Y.
3. Déterminez l'équation de la droite de régression de Y en fonction de X.
4. Déterminez la qualité de cet ajustement.
5. Établissez, sur base de ce modèle, le nombre de jours d'absence pour un fonctionnaire ayant 22 ans de service.

Solution
2)

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	2	3	4	9	6
	14	13	196	169	182
	16	17	256	289	272
	8	12	64	144	96
	13	10	169	100	130
	20	8	400	64	160
	24	20	576	400	480
	7	7	49	49	49
	5	2	25	4	10
	11	8	121	64	88
somme	120	100	1860	1292	1473
moyenne	12.00	10.00	186.00	129.20	147.30

$$\sum_{i=l}^{n} x_i = 120; \sum_{i=l}^{n} y_i = 100;$$

$$\sum_{i=l}^{n} x_i^2 = 1860; \sum_{i=l}^{n} y_i^2 = 1292;$$

$$\sum_{i=1}^{n} x_iy_i = 1473$$

$$\bar{x} = 120/10 = 12; \quad \bar{y} = 100/10 = 10;$$

$$s_x^2 = (1860/10) - 12^2 = 42; s_y^2 = (1292/10) - 10^2 = 29,2$$

$$s_{xy} = (1473/10) - (10.12) = 27,3$$

$$r_{xy} = \frac{27,3}{\sqrt{42 \times 29,2}} = 0.78$$

3)

$$D_{xy} \equiv y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

$$D_{xy} \equiv y - 10 = \frac{27,3}{42}(x - 12)$$

$$D_{xy} \equiv y = 0.65x + 2,2$$

4)

$$r_2 = 60.8\%;$$

$$s_e^2 = s_y^2(1 - r^2) = 29,2 \times (1 - 0.608) = 11,43$$

$s_e^2 = 11,43$ est beaucoup plus petit que $S_y^2 = 29,2$

5)

$$y = 0.65 \times 22 + 2,2 = 16,5 \text{ jours.}$$