Résumé du Cours de Statistique Descriptive

Yves Tillé

15 décembre 2010

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).

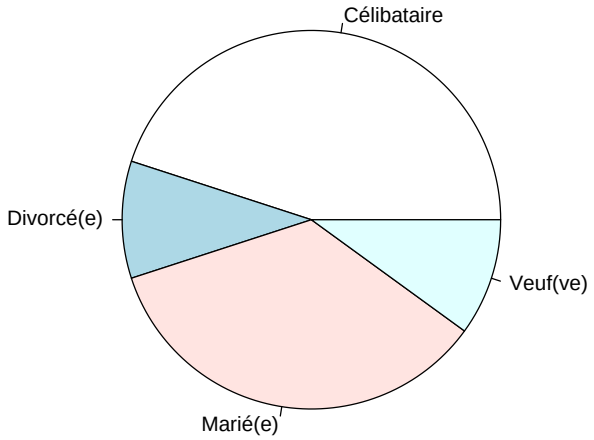


FIGURE 1.1 – Diagramme en secteurs des fréquences

En langage R

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

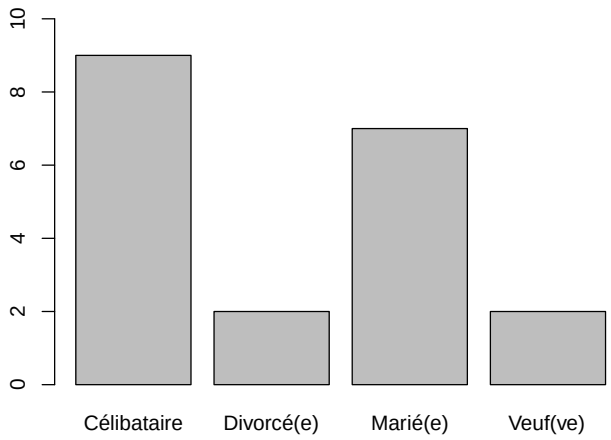


FIGURE 1.2 – Diagramme en barres des effectifs

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",
  "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).

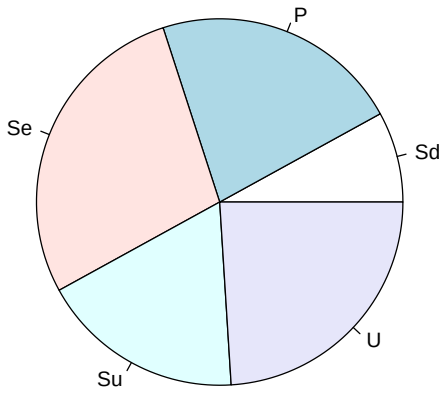


FIGURE 1.3 – Diagramme en secteurs des fréquences

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)
```

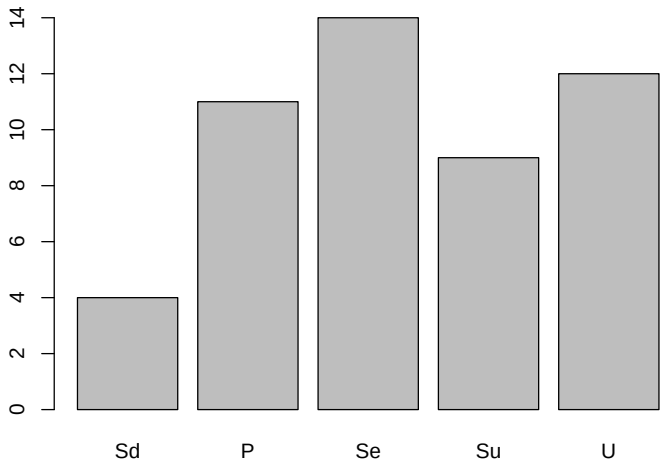


FIGURE 1.4 – Diagramme en barres des effectifs

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).

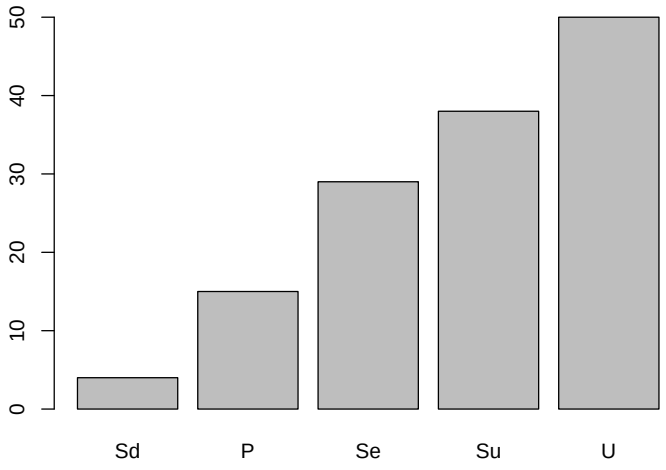


FIGURE 1.5 – Diagramme en barres des effectifs cumulés

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(1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,
+   4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,8,8)
> 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).

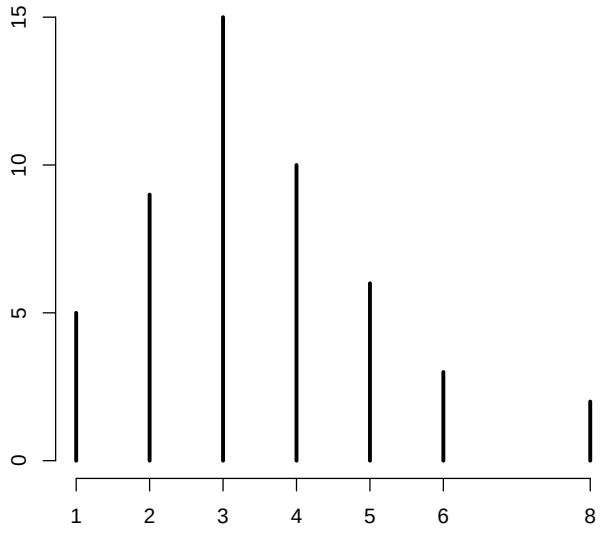


FIGURE 1.6 – Diagramme en bâtonnets des effectifs pour une variable quantitative discrète

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 :

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

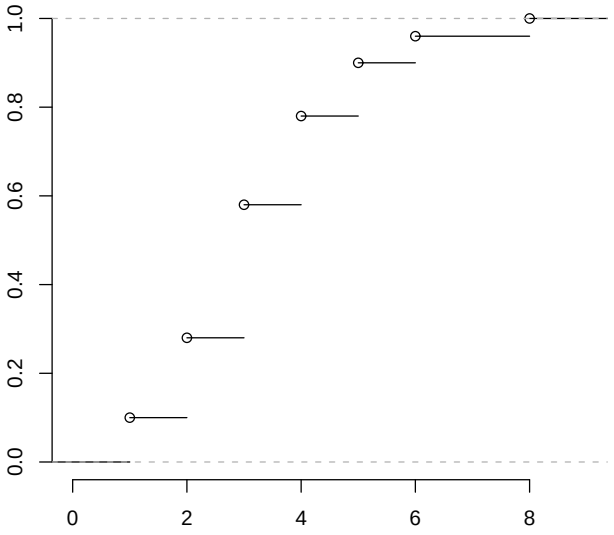


FIGURE 1.7 – Fonction de répartition d'une variable quantitative discrète

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 :

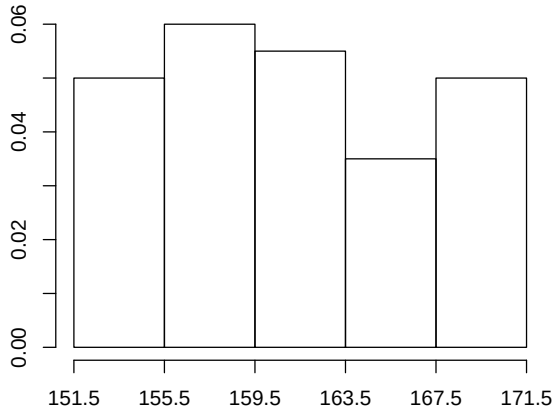


FIGURE 1.8 – Histogramme des fréquences

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))
```

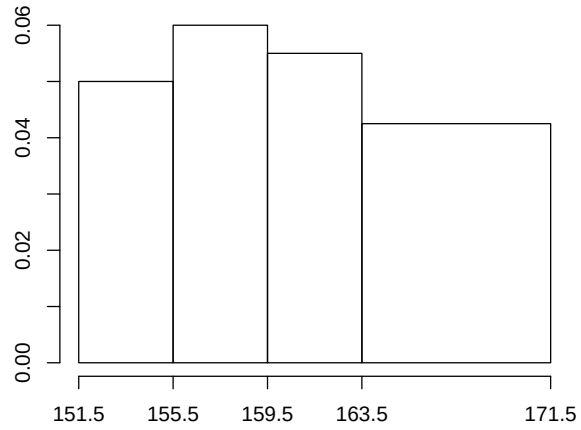


FIGURE 1.9 – Histogramme des fréquences avec les deux dernières classes agrégées

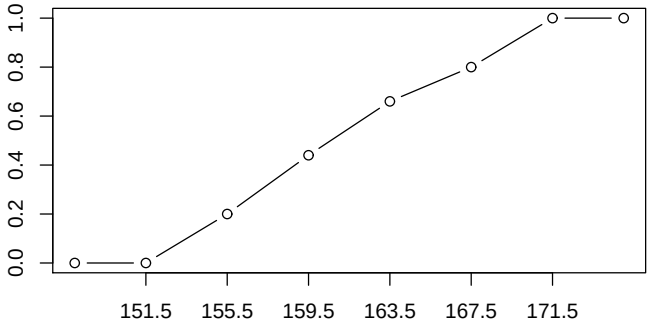
Remarque 1.3 Dans le cas de classes de même amplitude certains auteurs et logiciels représentent l'histogramme avec les effectifs (resp. les fréquences) reportés en ordonnée, l'aire de chaque rectangle étant proportionnelle à l'effectif (resp. la fréquence) de la classe.

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 :

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

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 $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$)

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

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 :

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

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 $x_{1/2} = 1$:

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

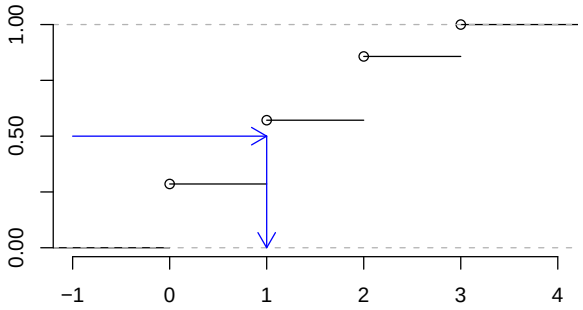
La Figure 2.1 montre la fonction de répartition de la série. La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur 1/2 :

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

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 :

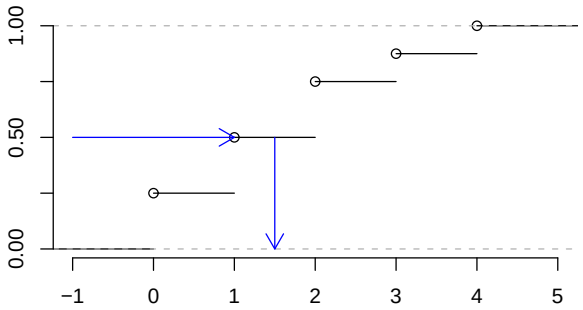
$$x_{1/2} = \frac{1+2}{2} = 1.5.$$

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.

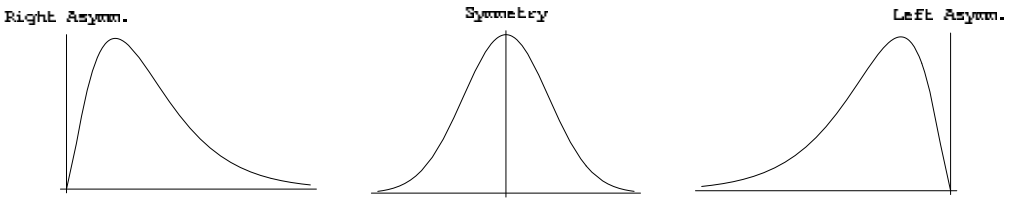


FIGURE 2.3 – Asymétrie d'une distribution

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.

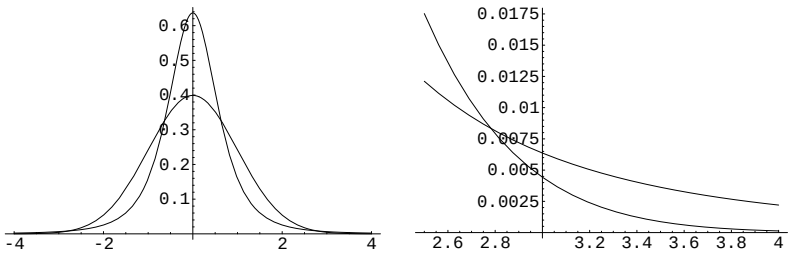


FIGURE 2.4 – Distributions mésokurtique et leptokurtique

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

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

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

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

On note que

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

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

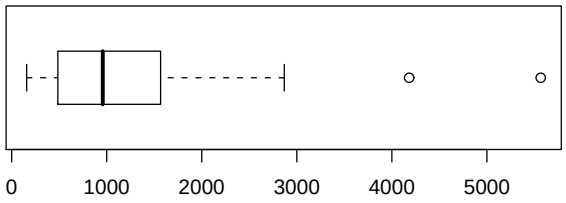


FIGURE 2.5 – Boîtes à moustaches pour la variable superficie en hectares (HApoly) des communes du canton de Neuchâtel

```
# 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

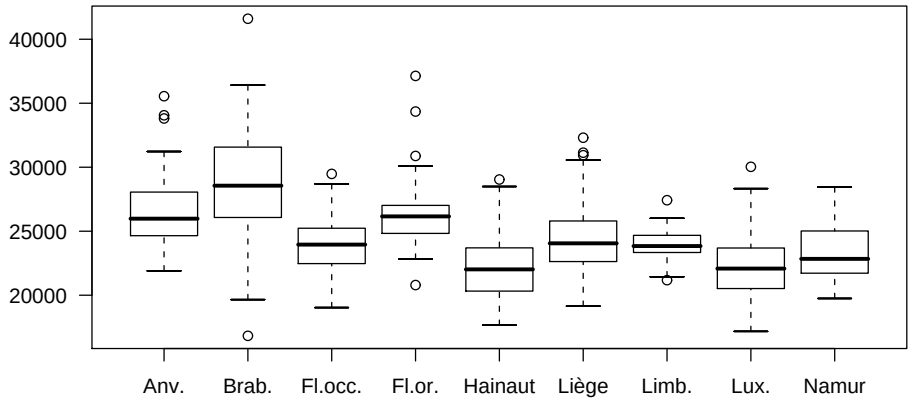


FIGURE 2.6 – Boîtes à moustaches du "revenu moyen des habitants" des communes selon les provinces belges

```
# 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

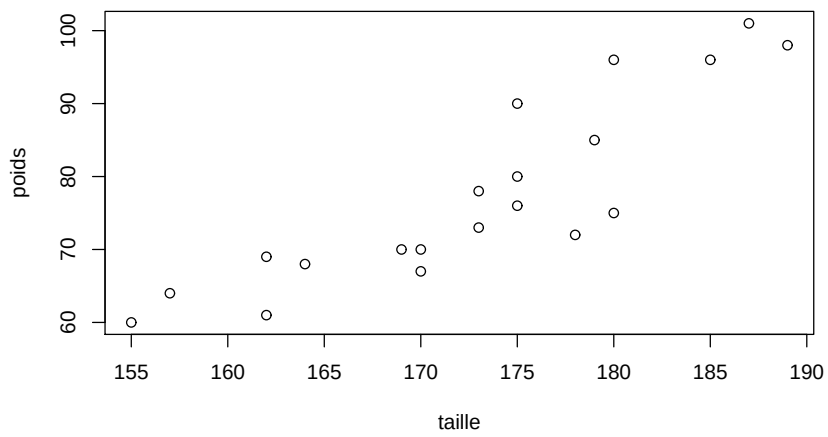


FIGURE 3.1 – Le nuage de points

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.

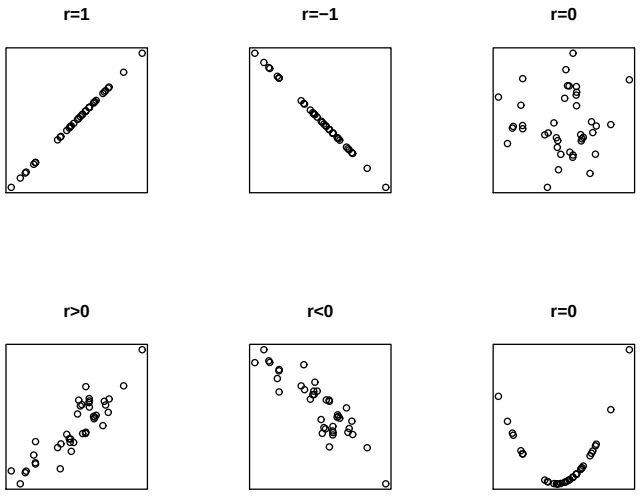


FIGURE 3.2 – Exemples de nuages de points et coefficients de corrélation

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.

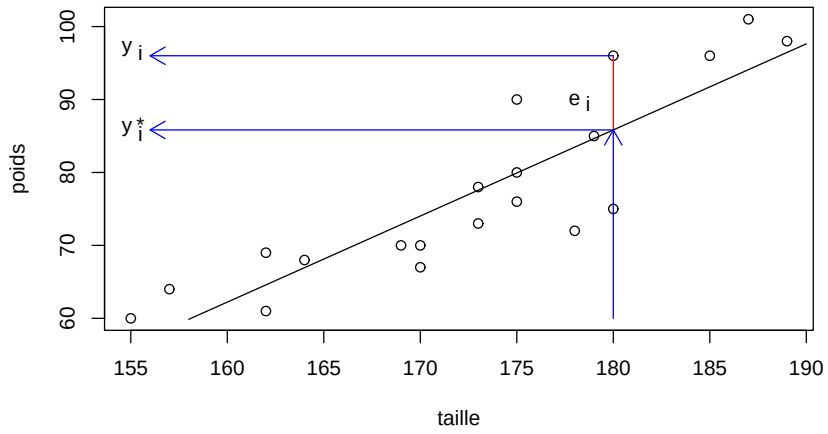


FIGURE 3.3 – Le nuage de points, le résidu

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

$$\left\{\begin{array}{l} \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} a - b\frac{1}{n}\sum_{i=1}^{n} x_i = 0 \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} a x_i - \frac{1}{n}\sum_{i=1}^{n} b x_i^2 = 0, \end{array}\right.$$

ce qui s'écrit aussi

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

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

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

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

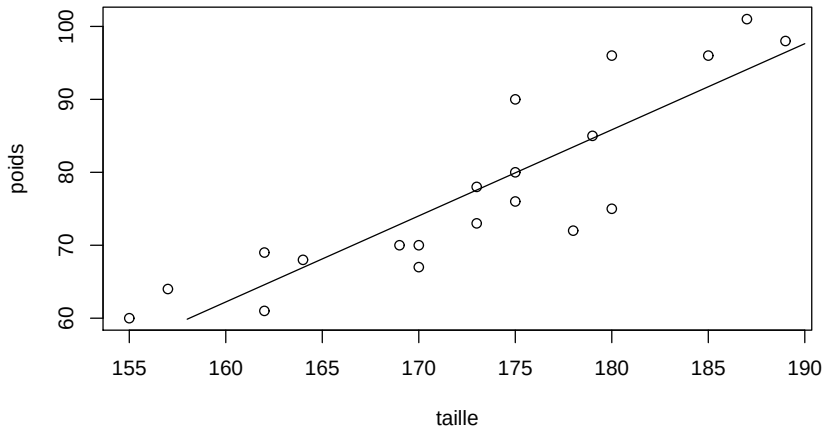
$$y = a + bx = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} + \frac{s_{xy}}{s_x^2}x,$$

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

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

□

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

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

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

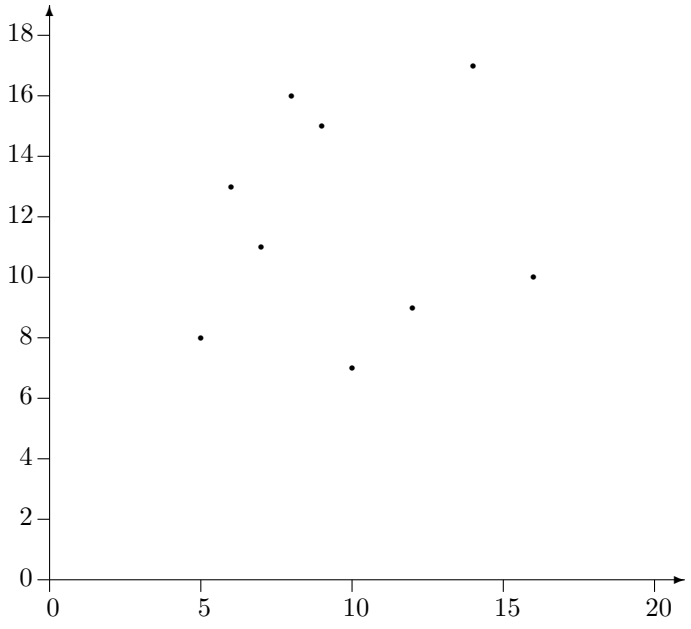
$$\rho = 0.77562456, b = 3.107356119, a = 206.8621479, y_1^* = 334.2637488, e_1 = 51.73625123,$$

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	$x_i y_i$
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	$x_i y_i$
	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

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

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

$$\sum_{i=1}^{n} x_i y_i = 1473$$

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

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

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

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

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)

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

Chapitre 4

Théorie des indices, mesures d'inégalité

4.1 Nombres indices

4.2 Définition

Un indice est la valeur d'une grandeur par rapport à une valeur de référence. Prenons l'exemple du tableau 4.1 contenant le prix (fictif) d'un bien de consommation de 2000 à 2006. Le temps varie de $0, 1, 2, \ldots, 6$ et 0 est considéré comme le temps de référence par rapport auquel l'indice est calculé.

TABLE 4.1 – Tableau du prix d'un bien de consommation de 2000 à 2006

année	t	prix p_t
2000	0	2.00
2001	1	2.30
2002	2	2.40
2003	3	2.80
2004	4	3.00
2005	5	3.50
2006	6	4.00

L'indice simple est défini par

$$I(t/t') = 100 \times \frac{p_t}{p_{t'}}, t, t' = 0, 1, \ldots, 6.$$

Le tableau 4.2 contient la matrice des indices de prix du bien. Par exemple de 2000 à 2006, le prix a doublé, donc $I(6/0) = 200$.

TABLE 4.2 – Tableau de l'indice simple du prix du tableau 4.1

	$t=0$	1	2	3	4	5	6
$t'=0$	100.00	115.00	120.00	140.00	150.00	175.00	200.00
1	86.96	100.00	104.35	121.74	130.43	152.17	173.91
2	83.33	95.83	100.00	116.67	125.00	145.83	166.67
3	71.43	82.14	85.71	100.00	107.14	125.00	142.86
4	66.67	76.67	80.00	93.33	100.00	116.67	133.33
5	57.14	65.71	68.57	80.00	85.71	100.00	114.29
6	50.00	57.50	60.00	70.00	75.00	87.50	100.00

4.2.1 Propriétés des indices

Considérons un indice quelconque $I(t/0)$. On dit que cet indice possède les propriétés de

- *réversibilité* si $I(t/0) = 100^2 \times \frac{1}{I(0/t)}$,
- *identité* si $I(t/t) = 100$,
- *circularité (ou transitivité)* si $I(t/u) \times I(u/v) = 100 \times I(t/v)$.

Il est facile de montrer que ces trois propriétés sont satisfaites pour un indice simple.

4.2.2 Indices synthétiques

Quand on veut calculer un indice à partir de plusieurs prix, le problème devient sensiblement plus compliqué. Un indice synthétique est une grandeur d'un ensemble de biens par rapport à une année de référence. On ne peut pas construire un indice synthétique en additionnant simplement des indices simples. Il faut, en effet, tenir compte des quantités achetées.

Pour calculer un indice de prix de n biens de consommation étiquetés de $1, 2, \ldots, n$, on utilise la notation suivante :

- p_{ti} représente le prix du bien de consommation i au temps t,
- q_{ti} représente la quantité de biens i consommée au temps t.

Considérons par exemple le Tableau 4.3 qui contient 3 biens de consommation et pour lesquels ont connaît les prix et les quantités achetées.

Il existe deux méthodes fondamentales pour calculer les indices de prix, l'indice de Paasche et l'indice de Laspeyres.

4.2.3 Indice de Laspeyres

L'indice de Laspeyres, est défini par

$$L(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{0i} p_{ti}}{\sum_{i=1}^{n} q_{0i} p_{0i}}.$$

On utilise pour le calculer, les quantités q_{0i} du temps de référence.

TABLE 4.3 – Exemple : prix et quantités de trois bien pendant 3 ans

Temps	0		1		2	
	Prix (p_{0i})	Quantités (q_{0i})	Prix (p_{1i})	Quantités (q_{1i})	Prix (p_{2i})	Quantités (q_{2i})
Bien 1	100	14	150	10	200	8
Bien 2	60	10	50	12	40	14
Bien 3	160	4	140	5	140	5

L'indice de Laspeyres peut aussi être présenté comme une moyenne pondérée des indices simples. Soient l'indice simple du bien i :

$$I_i(t/0) = 100 \times \frac{p_{ti}}{p_{0i}},$$

et le poids w_{0i} correspondant à la recette totale du bien i au temps 0

$$w_{0i} = p_{0i}q_{0i}.$$

L'indice de Laspeyres peut alors être défini comme une moyenne des indices simples pondérés par les recettes au temps 0 :

$$L(t/0) = \frac{\sum_{i=1}^n w_{0i} I_i(t/0)}{\sum_{i=1}^n w_{0i}} = \frac{\sum_{i=1}^n p_{0i}q_{0i} 100 \times \frac{p_{ti}}{p_{0i}}}{\sum_{i=1}^n p_{0i}q_{0i}} = 100 \times \frac{\sum_{i=1}^n q_{0i}p_{ti}}{\sum_{i=1}^n p_{0i}q_{0i}}.$$

L'indice de Laspeyres ne possède ni la propriété de circularité ni de réversibilité. L'indice de Laspeyres est facile à calculer, car seules les quantités q_{0i} du temps de référence sont nécessaires pour le calculer.

Exemple 4.1 Si on utilise les données du tableau 4.3, les indices de Laspeyres sont les suivants

$$L(1/0) = 100 \times \frac{\sum_{i=1}^n q_{0i}p_{1i}}{\sum_{i=1}^n q_{0i}p_{0i}} = 100 \times \frac{14 \times 150 + 10 \times 50 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 119.6970,$$

$$L(2/0) = 100 \times \frac{\sum_{i=1}^n q_{0i}p_{2i}}{\sum_{i=1}^n q_{0i}p_{0i}} = 100 \times \frac{14 \times 200 + 10 \times 40 + 4 \times 140}{14 \times 100 + 10 \times 60 + 4 \times 160} = 142.4242,$$

$$L(2/1) = 100 \times \frac{\sum_{i=1}^n q_{1i}p_{2i}}{\sum_{i=1}^n q_{1i}p_{1i}} = 100 \times \frac{10 \times 200 + 12 \times 40 + 5 \times 140}{10 \times 150 + 12 \times 50 + 5 \times 140} = 113.5714.$$

4.2.4 Indice de Paasche

L'indice de Paasche, est défini par

$$P(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}.$$

On utilise, pour le calculer, les quantités q_{ti} du temps par rapport auquel on veut calculer l'indice.

L'indice de Paasche peut aussi être présenté comme une moyenne harmonique pondérée des indices simples. Soient l'indice simple du bien i :

$$I_i(t/0) = 100 \times \frac{p_{ti}}{p_{0i}},$$

et le poids w_{ti} correspondant à la recette totale du bien i au temps t

$$w_{ti} = p_{ti} q_{ti}.$$

L'indice de Paasche peut alors être défini comme une moyenne harmonique des indices simples pondérés par les recettes au temps t :

$$P(t/0) = \frac{\sum_{i=1}^{n} w_{ti}}{\sum_{i=1}^{n} w_{ti}/I_i(t/0)} = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{ti} q_{ti} \frac{p_{0i}}{100 \times p_{ti}}} = 100 \times \frac{\sum_{i=1}^{n} q_{ti} p_{ti}}{\sum_{i=1}^{n} q_{ti} p_{0i}}.$$

L'indice de Paasche ne possède ni la propriété de circularité ni de réversibilité. L'indice de Paasche est plus difficile à calculer que l'indice de Laspeyres, car on doit connaître les quantités pour chaque valeur de t.

Exemple 4.2 Si on utilise les données du tableau 4.3, les indices de Paasche sont les suivants

$$P(1/0) = 100 \times \frac{\sum_{i=1}^{n} q_{1i} p_{1i}}{\sum_{i=1}^{n} q_{1i} p_{0i}} = 100 \times \frac{10 \times 150 + 12 \times 50 + 5 \times 140}{10 \times 100 + 12 \times 60 + 5 \times 160} = 111.1111,$$

$$P(2/0) = 100 \times \frac{\sum_{i=1}^{n} q_{2i} p_{2i}}{\sum_{i=1}^{n} q_{2i} p_{0i}} = 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 100 + 14 \times 60 + 5 \times 160} = 117.2131,$$

$$P(2/1) = 100 \times \frac{\sum_{i=1}^{n} q_{2i} p_{2i}}{\sum_{i=1}^{n} q_{2i} p_{1i}} = 100 \times \frac{8 \times 200 + 14 \times 40 + 5 \times 140}{8 \times 150 + 14 \times 50 + 5 \times 140} = 110.$$

4.2.5 L'indice de Fisher

L'indice de Laspeyres est en général plus grand que l'indice de Paasche, ce qui peut s'expliquer par le fait que l'indice de Laspeyres est une moyenne arithmétique d'indices élémentaires tandis que l'indice de Paasche est une moyenne harmonique. Nous avons vu qu'une moyenne harmonique est toujours inférieure

ou égale à une moyenne arithmétique (voir la remarque de la page 32). Cependant ici, ce résultat est approximatif, car on n'utilise pas les mêmes poids pour calculer l'indice de Paasche (w_{ti}) et de Laspeyres (w_{0i}).

Fisher a proposé d'utiliser un compromis entre l'indice de Paasche et de Laspeyres en calculant simplement la moyenne géométrique de ces deux indices

$$F(t/0) = \sqrt{L(t/0) \times P(t/0)}.$$

L'avantage de l'indice de Fisher est qu'il jouit de la propriété de réversibilité.

Exemple 4.3 Si on utilise toujours les données du tableau 4.3, les indices de Fisher sont les suivants :

$$F(1/0) = \sqrt{L(1/0) \times P(1/0)} = 115.3242,$$
$$F(2/0) = \sqrt{L(2/0) \times P(2/0)} = 129.2052,$$
$$F(2/1) = \sqrt{L(2/1) \times P(2/1)} = 111.7715.$$

4.2.6 L'indice de Sidgwick

L'indice de Sidgwick est la moyenne arithmétique des indices de Paasche et de Laspeyres.

$$S(t/0) = \frac{L(t/0) + P(t/0)}{2}.$$

4.2.7 Indices chaînes

Le défaut principal des indices de Laspeyres, de Paasche, de Fisher et de Sidgwick est qu'il ne possèdent pas la propriété de circularité. Un indice qui possède cette propriété est appelé indice chaîne. Pour construire un indice chaîne, avec l'indice de Laspeyres, on peut faire un produit d'indice de Laspeyres annuels.

$$CL(t/0) = 100 \times \frac{L(t/t-1)}{100} \times \frac{L(t-1/t-2)}{100} \times \cdots \times \frac{L(2/1)}{100} \times \frac{L(1/0)}{100}.$$

Pour calculer un tel indice, on doit évidemment connaître les quantités pour chaque valeur de t. L'indice suisse des prix à la consommation est un indice chaîne de Laspeyres.

Exemple 4.4 En utilisant encore les données du tableau 4.3, les indices chaînes de Laspeyres sont les suivants :

$$CL(1/0) = L(1/0) = 119.6970,$$
$$CL(2/1) = L(2/1) = 113.5714,$$
$$CL(2/0) = \frac{L(2/1) \times L(1/0)}{100} = 135.9416.$$

4.3 Mesures de l'inégalité

4.3.1 Introduction

Des indicateurs particuliers ont été développés pour mesurer les inégalités des revenus ou les inégalités de patrimoine. On considère qu'une société est parfaitement égalitaire si tous les individus reçoivent le même revenu. La situation théorique la plus inégalitaire est la situation où un individu perçoit la totalité des revenus, et les autre individus n'ont aucun revenu.

4.3.2 Courbe de Lorenz

Plusieurs indices d'inégalité sont liés à la courbe de Lorenz. On note

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

les revenus des n individus de la population étudiée. On note également

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

la statistique d'ordre, c'est-à-dire la série de revenus triés par ordre croissant.

Notons maintenant q_i la proportion de revenus par rapport au revenu total qu'ont gagné les i individus ayant les plus bas revenus, ce qui s'écrit

$$q_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} \text{ avec } q_0 = 0 \text{ et } q_n = 1.$$

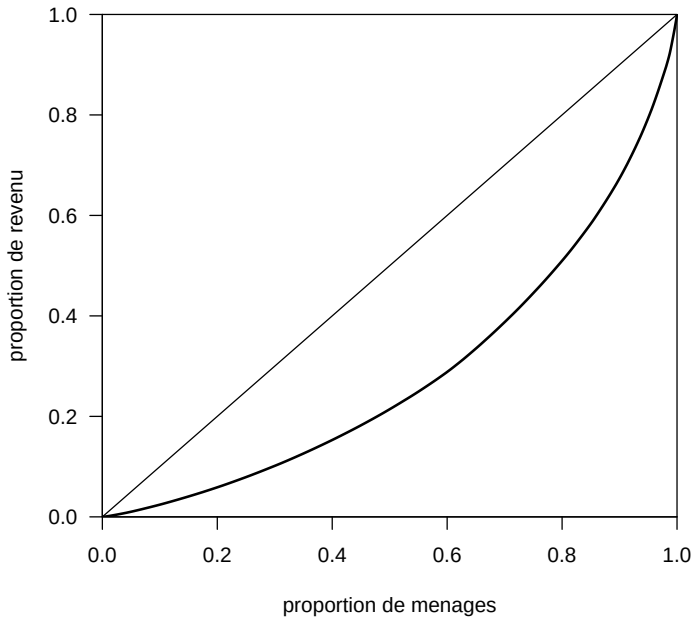
La courbe de Lorenz est la représentation graphique de la fonction qui à la part des individus les moins riches associe la part y du revenu total qu'ils perçoivent. Plus précisément, la courbe de Lorenz relie les points $(i/n, q_i)$ pour $i = 1, \ldots, n$. En abscisse, on a donc une proportion d'individus classés par ordre de revenu, et en ordonnée la proportion du revenu total reçu par ces individus.

Exemple 4.5 On utilise une enquête ménage sur le revenu dans une région des Philippines appelée Ilocos. Cette enquête de 1997 sur le revenu des ménages a été produite par l'Office philippin de Statistique. La courbe de Lorenz est présentée en Figure 4.1.

Remarque 4.1 Sur le graphique, on indique toujours la diagonale. La courbe de Lorenz est égale à la diagonale si tous les individus ont le même revenu. Plus l'écart entre la courbe de Lorenz et la diagonale est important, plus les revenus sont distribués de manière inégalitaire.

En langage R

FIGURE 4.1 – Courbe de Lorenz




```
#
# Courbe de Lorenz et indices d'inégalité
#
# Etape 1 : on installe la package ineq
utils:::menuInstallPkgs()
# choisir 'ineq' dans la liste
#
#Etape 2 : on charge le package ineq
local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
# choisir 'ineq' dans la liste
#
# Utilisation de la base de données Ilocos
# Enquête sur le revenu de l'Office de Statistique Philippin
data(Ilocos)
attach(Ilocos)
#
plot(Lc(income),xlab="proportion de menages",
ylab="proportion de revenu",main="")
```

4.3.3 Indice de Gini

L'indice de Gini, noté G est égal à deux fois la surface comprise entre la courbe de Lorenz et la diagonale. Il est possible de montrer que :

$$G = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2\bar{x}}.$$

En utilisant la statistique d'ordre $x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(n)}$, l'indice de Gini peut également s'écrire

$$G = \frac{1}{n-1} \left[\frac{2 \sum_{i=1}^{n} i x_{(i)}}{n\bar{x}} - (n+1) \right].$$

L'indice de Gini est compris entre 0 et 1. S'il est proche de 0, tous le revenus sont égaux. S'il est proche de 1, les revenus sont très inégaux.

4.3.4 Indice de Hoover

L'indice d'équirépartition de Hoover (ou *Robin Hood index*) est défini comme la proportion de revenus qu'il faudrait prendre aux individus gagnant plus que la moyenne et redistribuer aux individus gagnant moins que la moyenne pour que tout le monde ait le même revenu. Il est formellement définit par :

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

Cet indice est également compris entre 0 et 1. Il vaut 0 si tous les individus ont le même revenu.

Cet indice est également lié à la courbe de Lorenz, car il est possible de montrer qu'il correspond à la plus grande distance verticale entre la courbe de Lorenz et la diagonale.

4.3.5 Quintile et Decile share ratio

On définit d'abord :

- S_{10} revenu moyen des individus ayant un revenu inférieur au premier décile $x_{1/10}$,
- S_{20} revenu moyen des individus ayant un revenu inférieur au premier quintile ou deuxième décile $x_{1/5}$,
- S_{80} revenu moyen des individus ayant un revenu supérieur au quatrième quintile ou huitième décile $x_{4/5}$,
- S_{90} revenu moyen des individus ayant un revenu supérieur au neuvième décile $x_{9/10}$.

Le quintile share ratio est définit par

$$QSR = \frac{S_{80}}{S_{20}}.$$

Le decile share ratio est définit par

$$DSR = \frac{S_{90}}{S_{10}}.$$

Ces quantités sont toujours plus grandes que 1 et augmentent avec l'inégalité. Ces deux rapports sont facilement interprétables, par exemple si le $QSR = 5$, cela signifie que le revenu moyen de 20% des plus riches est 5 fois plus grand que le revenu moyen de 20% des plus pauvres.

4.3.6 Indice de pauvreté

Un indice simple de pauvreté consiste à calculer le pourcentage de la population gagnant moins que la moitié de la médiane.

4.3.7 Indices selon les pays

Le tableau 4.4 reprend pour tous les pays l'indice de Gini et le rapport des 20% les plus riches sur les 20% les plus pauvres. (référence : United Nations 2005 Development Programme Report, page 270).

Exercices

Exercice 4.1 Étudiez les propriétés (circularité, réversibilité, identité et transitivité) de tous les indices de prix présentés.

TABLE 4.4 – Mesures de l'inégalité par pays

Rang	Pays	Indice de Gini	DSR	QSR	Année de l'enquête
1	Denmark	24.7	8.1	4.3	1997
2	Japan	24.9	4.5	3.4	1993
3	Sweden	25	6.2	4	2000
4	Belgium	25	7.8	4.5	1996
5	Czech Republic	25.4	5.2	3.5	1996
6	Norway	25.8	6.1	3.9	2000
7	Slovakia	25.8	6.7	4	1996
8	Bosnia and Herzegovina	26.2	5.4	3.8	2001
9	Uzbekistan	26.8	6.1	4	2000
10	Finland	26.9	5.6	3.8	2000
11	Hungary	26.9	5.5	3.8	2002
12	Republic of Macedonia	28.2	6.8	4.4	1998
13	Albania	28.2	5.9	4.1	2002
14	Germany	28.3	6.9	4.3	2000
15	Slovenia	28.4	5.9	3.9	1998
16	Rwanda	28.9	5.8	4	1983
17	Croatia	29	7.3	4.8	2001
18	Ukraine	29	6.4	4.3	1999
19	Austria	30	7.6	4.7	1997
20	Ethiopia	30	6.6	4.3	1999
21	Romania	30.3	8.1	5.2	2002
22	Mongolia	30.3	17.8	9.1	1998
23	Belarus	30.4	6.9	4.6	2000
24	Netherlands	30.9	9.2	5.1	1999
25	Russia	31	7.1	4.8	2002
26	South Korea	31.6	7.8	4.7	1998
27	Bangladesh	31.8	6.8	4.6	2000
28	Lithuania	31.9	7.9	5.1	2000
29	Bulgaria	31.9	9.9	5.8	2001
30	Kazakhstan	32.3	7.5	5.1	2003
31	Spain	32.5	9	5.4	1990
32	India	32.5	7.3	4.9	1999
33	Tajikistan	32.6	7.8	5.2	2003
34	France	32.7	9.1	5.6	1995
35	Pakistan	33	7.6	4.8	1998
36	Canada	33.1	10.1	5.8	1998
37	Switzerland	33.1	9.9	5.8	1992
38	Sri Lanka	33.2	8.1	5.1	1999
39	Burundi	33.3	19.3	9.5	1998
61	Estonia	37.2	14.9	7.2	2000
65	Portugal	38.5	15	8	1997
92	United States	46.6	15.9	8.4	2000
100	Peru	49.8	49.9	18.4	2000
101	Malawi	50.3	22.7	11.6	1997
102	Mali	50.5	23.1	12.2	1994
103	Niger	50.5	46	20.7	1995
104	Nigeria	50.6	24.9	12.8	1996
105	Papua New Guinea	50.9	23.8	12.6	1996
106	Argentina	52.2	39.1	18.1	2001
107	Zambia	52.6	41.8	17.2	1998
108	El Salvador	53.2	47.4	19.8	2000
109	Mexico	54.6	45	19.3	2000
110	Honduras	55	49.1	21.5	1999
111	Panama	56.4	62.3	24.7	2000
112	Zimbabwe	56.8	22	12	1995
113	Chile	57.1	40.6	18.7	2000
114	Colombia	57.6	57.8	22.9	1999
115	Paraguay	57.8	73.4	27.8	2002
116	South Africa	57.8	33.1	17.9	2000
117	Brazil	59.3	68	26.4	2001
118	Guatemala	59.9	55.1	24.4	2000
119	Swaziland	60.9	49.7	23.8	1994
120	Central African Republic	61.3	69.2	32.7	1993
121	Sierra Leone	62.9	87.2	57.6	1989
122	Botswana	63	77.6	31.5	1993
123	Lesotho	63.2	105	44.2	1995
124	Namibia	70.7	128.8	56.1	1993

Chapitre 5

Calcul des probabilités et variables aléatoires

5.1 Probabilités

5.1.1 Événement

Une expérience est dite aléatoire si on ne peut pas prédire *a priori* son résultat. On note ω un résultat possible de cette expérience aléatoire. L'ensemble de tous les résultats possibles est noté Ω. Par exemple, si on jette deux pièces de monnaie, on peut obtenir les résultats

$$\Omega = \{(P,P),(F,P),(P,F),(F,F)\},$$

avec F pour "face" et P pour "pile". Un événement est une assertion logique sur une expérience aléatoire comme "avoir deux fois pile" ou "avoir au moins une fois pile". Formellement, un événement est un sous-ensemble de Ω.

– L'événement "avoir deux fois pile" est le sous ensemble $\{(P,P)\}$.

– L'événement "avoir au moins une fois pile" est le sous ensemble $\{(P,P),(F,P),(P,F)\}$.

L'ensemble Ω est appelé événement certain, et l'ensemble vide $\emptyset$ est appelé événement impossible.

5.1.2 Opérations sur les événements

Sur les événements, on peut appliquer les opérations habituelles de la théorie des ensembles.

L'union

L'événement $A \cup B$ est réalisé dès que A ou B est réalisé. Dans un lancer de dé, si l'événement A est "obtenir un nombre pair" et l'événement B "obtenir un multiple de 3", l'événement $A \cup B$ est l'événement "obtenir un nombre pair OU un multiple de 3", c'est-à-dire $\{2,3,4,6\}$.

L'intersection

L'événement $A \cap B$ est réalisé dès que A et B sont réalisés conjointement dans la même expérience. Dans un lancer de dé, si l'événement A est "obtenir un nombre pair" et l'événement B "obtenir un multiple de 3", l'événement $A \cap B$ est l'événement "obtenir un nombre pair ET multiple de 3", c'est-à-dire $\{6\}$.

La différence

L'événement $A \backslash B$ est réalisé quand A est réalisé et que B ne l'est pas.

Le complémentaire

Le complémentaire de l'événement A est l'événement $\Omega \backslash A$. Le complémentaire est noté $\overline{A}$.

Exemple 5.1 L'expérience peut consister à jeter un dé, alors

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

et un événement, noté A, est "obtenir un nombre pair". On a alors

$$A = \{2, 4, 6\} \text{ et } \overline{A} = \{1, 3, 5\}.$$

5.1.3 Relations entre les événements

Événements mutuellement exclusifs

Si $A \cap B = \emptyset$ on dit que A et B sont mutuellement exclusifs, ce qui signifie que A et B ne peuvent pas se produire ensemble.

Exemple 5.2 Si on jette un dé, l'événement "obtenir un nombre pair" et l'événement "obtenir un nombre impair" ne peuvent pas être obtenus en même temps. Ils sont mutuellement exclusifs. D'autre part, si l'on jette un dé, les événements A : "obtenir un nombre pair" n'est pas mutuellement exclusif avec l'événement B : "obtenir un nombre inférieur ou égal à 3". En effet, l'intersection de A et B est non-vide et consiste en l'événement "obtenir 2".

Inclusion

Si A est inclus dans B, on écrit $A \subset B$. On dit que A implique B.

Exemple 5.3 Si on jette un dé, on considère les événements A "obtenir 2" et B "obtenir un nombre pair".

$$A = \{2\} \text{ et } B = \{2, 4, 6\}.$$

On dit que A implique B.

5.1.4 Ensemble des parties d'un ensemble et système complet

On va associer à Ω l'ensemble $\mathcal{A}$ de toutes les parties (ou sous-ensembles) de Ω.

Exemple 5.4 Si on jette une pièce de monnaie alors $\Omega = \{P, F\}$, et

$$\mathcal{A} = \{\emptyset, \{F\}, \{P\}, \{F, P\}\}.$$

Définition 5.1 *Les événements $A_1, \ldots, A_n$ forment un système complet d'événements, si ils constituent une partition de Ω, c'est-à-dire si*
- *tous les couples A_i, A_j sont mutuellement exclusifs quand $i \neq j$,*
- $\bigcup_{i=1}^{n} A_i = \Omega$.

TABLE 5.1 – Système complet d'événements

A_1	.....	A_i	.....	A_n

5.1.5 Axiomatique des Probabilités

Définition 5.2 *Une probabilité $P(.)$ est une application de $\mathcal{A}$ dans $[0, 1]$, telle que :*
- $\Pr(\Omega) = 1$,
- *Pour tout ensemble dénombrable d'événements $A_1, .., A_n$ mutuellement exclusifs (tels que $A_i \cap A_j = \emptyset$, pour tout $i \neq j$),*

$$\Pr(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \Pr(A_3) + \cdots + \Pr(A_n).$$

A partir des axiomes, on peut déduire les propriétés suivantes :

Propriété 5.1 $\Pr(\emptyset) = 0$.

Démonstration
Comme $\emptyset$ est d'intersection vide avec $\emptyset$, on a que

$$\Pr(\emptyset \cup \emptyset) = \Pr(\emptyset) + \Pr(\emptyset).$$

Donc,

$$\Pr(\emptyset) = 2\Pr(\emptyset),$$

ce qui implique que $\Pr(\emptyset) = 0$. □

Propriété 5.2 $\Pr(\overline{A}) = 1 - \Pr(A)$.

Démonstration
On sait que

$$A \cup \overline{A} = \Omega \text{ et } A \cap \overline{A} = \emptyset.$$

Ainsi, on a que

$$\Pr(\Omega) = \Pr(A \cup \overline{A}) = \Pr(A) + \Pr(\overline{A}).$$

Mais, par la définition d'une probabilité, $\Pr(\Omega) = 1$. Donc,

$$\Pr(A) + \Pr(\overline{A}) = 1$$

On en déduit que $\Pr(\overline{A}) = 1 - \Pr(A)$. □

Propriété 5.3 $\Pr(A) \leq \Pr(B)$ *si* $A \subset B$.

Démonstration
Comme $A \subset B$, on a

$$B = (B \cap \overline{A}) \cup A.$$

Mais on a que

$$(B \cap \overline{A}) \cap A = \emptyset.$$

Ainsi, on a

$$\Pr(B) = \Pr(B \cap \overline{A}) + \Pr(A).$$

Or une probabilité est à valeur dans [0,1], donc $\Pr(B \cap \overline{A}) \geq 0$. On a alors

$$\Pr(B) \geq \Pr(A).$$

□

Propriété 5.4 $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Démonstration
On a

$$A \cup B = A \cup (B \cap \overline{A}),$$

et

$$A \cap (B \cap \overline{A}) = \emptyset.$$

Donc

$$\Pr(A \cup B) = \Pr(A) + \Pr(B \cap \overline{A}).$$

Il reste à montrer que

$$\Pr(B \cap \overline{A}) = \Pr(B) - \Pr(A \cap B)$$

En effet,

$$B = (B \cap \overline{A}) \cup (B \cap A)$$

avec

$$(B \cap \overline{A}) \cap (B \cap A) = \emptyset$$

Donc

$$\Pr(B) = \Pr(B \cap \overline{A}) + \Pr(B \cap A),$$

ce qui donne

$$\Pr(B \cap \overline{A}) = \Pr(B) - \Pr(A \cap B).$$

□

Propriété 5.5 $\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$

Démonstration
Notons respectivement

$$B_1 = A_1, \quad B_2 = (A_2 \backslash A_1), \quad B_3 = (A_3 \backslash (A_1 \cup A_2)),$$

$$B_4 = (A_4 \backslash (A_1 \cup A_2 \cup A_3)), \quad \ldots, B_n = (A_n \backslash (A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_{n-1})).$$

Comme

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i,$$

et que $B_i \cap B_j = \emptyset$ pour tout $j \neq i$, alors

$$\Pr\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \Pr\left(B_i\right).$$

De plus, comme, pour tout i, $B_i \subset A_i$, on a que $\Pr(B_i) \leq \Pr(A_i)$, ce qui donne finalenent

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \Pr\left(B_i\right) \leq \sum_{i=1}^{n} \Pr\left(A_i\right).$$

□

Propriété 5.6 *Si $A_1, \ldots, A_n$ forment un système complet d'événements, alors*

$$\sum_{i=1}^{n} \Pr(B \cap A_i) = \Pr(B).$$

Démonstration
Si $A_1, \ldots, A_n$ forment un système complet d'événements, alors

$$B = \bigcup_{i=1}^{n} (B \cap A_i).$$

Mais on a, pour tout i, j tels que $i \neq j$

$$(B \cap A_i) \cap (B \cap A_j) = \emptyset.$$

Finalement, on a que

$$\Pr(B) = \Pr\left(\bigcup_{i=1}^{n} (B \cap A_i)\right) = \sum_{i=1}^{n} \Pr(B \cap A_i).$$

□

5.1.6 Probabilités conditionnelles et indépendance

Définition 5.3 *Soient deux événements A et B, si $\Pr(B) > 0$, alors*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Exemple 5.5 Si on jette un dé, et que l'on considère les deux événements suivants :
- A l'évènement 'avoir un nombre pair' et
- B l'évènement 'avoir un nombre supérieur ou égal à 4'.

On a donc
- $\Pr(A) = \Pr(\{2, 4, 6\}) = \frac{1}{2}$,
- $\Pr(B) = \Pr(\{4, 5, 6\}) = \frac{3}{6} = \frac{1}{2}$,
- $\Pr(A \cap B) = \Pr(\{4, 6\}) = \frac{2}{6} = \frac{1}{3}$,
- $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3}$.

Définition 5.4 *Deux événements A et B sont dits indépendants si*

$$\Pr(A|B) = \Pr(A).$$

On peut montrer facilement que si A et B sont indépendants, alors

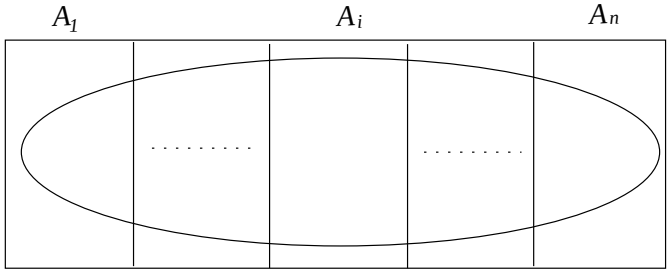
$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

5.1.7 Théorème des probabilités totales et théorème de Bayes

Théorème 5.1 *(des probabilités totales) Soit $A_1, \ldots, A_n$ un système complet d'événements, alors*

$$\Pr(B) = \sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i).$$

TABLE 5.2 – Illustration du théorème des probabilités totales



En effet,

$$\sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i) = \sum_{i=1}^{n} \Pr(B \cap A_i).$$

Comme les événements $A_i \cap B$ sont mutuellement exclusifs,

$$\sum_{i=1}^{n} \Pr(B \cap A_i) = \Pr \bigcup_{i=1}^{n} (B \cap A_i) = \Pr(B).$$

Théorème 5.2 *(de Bayes) Soit $A_1, \ldots, A_n$ un système complet d'événements, alors*

$$\Pr(A_i|B) = \frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{j=1}^{n} \Pr(A_j)\Pr(B|A_j)}.$$

En effet, par le théorème des probabilités totales,

$$\frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{j=1}^{n} \Pr(A_j)\Pr(B|A_j)} = \frac{\Pr(B \cap A_i)}{\Pr(B)} = \Pr(A_i|B).$$

Exemple 5.6 Supposons qu'une population d'adultes soit composée de 30% de fumeurs (A_1) et de 70% de non-fumeurs (A_2). Notons B l'événement "mourir d'un cancer du poumon". Supposons en outre que la probabilité de mourir d'un cancer du poumon est égale à $\Pr(B|A_1) = 20\%$ si l'on est fumeur et de $\Pr(B|A_2) = 1\%$ si l'on est non-fumeur. Le théorème de Bayes permet de calculer

les probabilités a priori, c'est-à-dire la probabilité d'avoir été fumeur si on est mort d'un cancer du poumon. En effet, cette probabilité est notée $\Pr(A_1|B)$ et peut être calculée par

$$\Pr(A_1|B) = \frac{\Pr(A_1)\Pr(B|A_1)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.3 \times 0.2}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.06}{0.06 + 0.007} \approx 0.896.$$

La probabilité de ne pas avoir été non-fumeur si on est mort d'un cancer du poumon vaut quant à elle :

$$\Pr(A_2|B) = \frac{\Pr(A_2)\Pr(B|A_2)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.7 \times 0.01}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.07}{0.06 + 0.007} \approx 0.104.$$

5.2 Analyse combinatoire

5.2.1 Introduction

L'analyse combinatoire est l'étude mathématique de la manière de ranger des objets. L'analyse combinatoire est un outil utilisé dans le calcul des probabilités.

5.2.2 Permutations (sans répétition)

Une permutation sans répétition est un classement ordonné de n objets distincts. Considérons par exemple l'ensemble $\{1, 2, 3\}$. Il existe 6 manières d'ordonner ces trois chiffres :

$$\{1, 2, 3\}, \{1, 3, 2\}, \{2, 1, 3\}, \{2, 3, 1\}, \{3, 1, 2\}, \{3, 2, 1\}.$$

Si on dispose de n objets, chacun des n objets peut être placé à la première place. Il reste ensuite $n - 1$ objets qui peuvent être placés à la deuxième place, puis $n-2$ objets pour la troisième place, et ainsi de suite. Le nombre de permutations possibles de n objets distincts vaut donc

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = n!.$$

La notation $n!$ se lit factorielle de n (voir tableau 5.3).

TABLE 5.3 – Factorielle des nombres de 1 à 10

n	0	1	2	3	4	5	6	7	8	9	10
$n!$	1	1	2	6	24	120	720	5040	40320	362880	3628800

5.2.3 Permutations avec répétition

On peut également se poser la question du nombre de manières de ranger des objets qui ne sont pas tous distincts. Supposons que nous ayons 2 boules rouges (notées R) et 3 boules blanches (notées B). Il existe 10 permutations possibles qui sont :

$$\{R,R,B,B,B\},\{R,B,R,B,B\},\{R,B,B,R,B\},\{R,B,B,B,R\},\{B,R,R,B,B\},$$

$$\{B,R,B,R,B\},\{B,R,B,B,R\},\{B,B,R,R,B\},\{B,B,R,B,R\},\{B,B,B,R,R\}.$$

Si l'on dispose de n objets appartenant à deux groupes de tailles n_1 et n_2, le nombre de permutations avec répétition est

$$\frac{n!}{n_1!n_2!}.$$

Par exemple si l'on a 3 boules blanches et 2 boules rouges, on obtient

$$\frac{n!}{n_1!n_2!} = \frac{5!}{2!3!} = \frac{120}{2\times 6} = 10.$$

Si l'on dispose de n objets appartenant à p groupes de tailles $n_1, n_2, \ldots, n_p$, le nombre de permutations avec répétition est

$$\frac{n!}{n_1!n_2!\times\cdots\times n_p!}.$$

5.2.4 Arrangements (sans répétition)

Soit n objets distincts. On appelle un arrangement une manière de sélectionner k objets parmi les n et de les ranger dans des boîtes numérotées de 1 à k.

Dans la première boîte, on peut mettre chacun des n objets. Dans la seconde boîte, on peut mettre chacun des $n-1$ objets restants, dans la troisième boîte, on peut mettre chacun des $n-2$ objets restants et ainsi de suite. Le nombre d'arrangements possibles est donc égal à :

$$A_n^k = n\times(n-1)\times(n-2)\times\cdots\times(n-k+1) = \frac{n!}{(n-k)!}.$$

5.2.5 Combinaisons

Soit n objets distincts. On appelle une combinaison une manière de sélectionner k objets parmi les n sans tenir compte de leur ordre. Le nombre de combinaisons est le nombre de sous-ensembles de taille k dans un ensemble de taille n. Soit l'ensemble $\{1,2,3,4,5\}$. Il existe 10 sous-ensembles de taille 3 qui sont :

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{1,4,5\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}.$$

De manière générale, quel est le nombre de combinaisons de k objets parmi n ? Commençons par calculer le nombre de manières différentes de sélectionner

k objets parmi n en tenant compte de l'ordre : c'est le nombre d'arrangements sans répétition A_n^k. Comme il existe $k!$ manières d'ordonner ces k éléments, si l'on ne veut pas tenir compte de l'ordre on divise A_n^k par $k!$. Le nombre de combinaisons de k objets parmi n vaut donc

$$\frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}.$$

Le nombre de combinaisons de k objets parmi n s'écrit parfois $\binom{n}{k}$ et parfois C_n^k :

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}.$$

Par exemple, si on cherche à déterminer le nombre de combinaisons de 3 objets parmi 5, on a

$$\binom{5}{3} = C_5^3 = \frac{5!}{3!(5-3)!} = \frac{120}{6 \times 2} = 10.$$

5.3 Variables aléatoires

5.3.1 Définition

La notion de variable aléatoire formalise l'association d'une valeur au résultat d'une expérience aléatoire.

Définition 5.5 *Une variable aléatoire X est une application de l'ensemble fondamental Ω dans $\mathbb{R}$.*

Exemple 5.7 On considère une expérience aléatoire consistant à lancer deux pièces de monnaie. L'ensemble des résultats possibles est

$$\Omega = \{(F,F),(F,P),(P,F),(P,P)\}.$$

Chacun des éléments de Ω a une probabilité $1/4$. Une variable aléatoire va associer une valeur à chacun des éléments de Ω. Considérons la variable aléatoire représentant le nombre de "Faces" obtenus :

$$X = \begin{cases} 0 & \text{avec une probabilité } 1/4 \\ 1 & \text{avec une probabilité } 1/2 \\ 2 & \text{avec une probabilité } 1/4. \end{cases}$$

C'est une variable aléatoire discrète dont la distribution de probabilités est présentée en Figure 5.1.

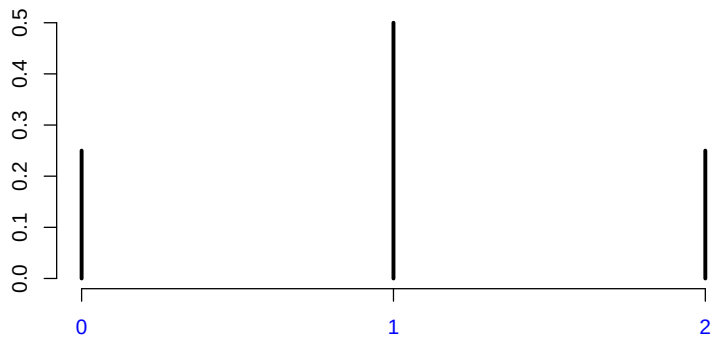


FIGURE 5.1 – Distribution de "faces" obtenus.

5.4 Variables aléatoires discrètes

5.4.1 Définition, espérance et variance

Une variable aléatoire discrète prend uniquement des valeurs entières (de $\mathbb{Z}$).

Une distribution de probabilités $p_X(x)$ est une fonction qui associe à chaque valeur entière une probabilité.

$$p_X(x) = \Pr(X = x), x \in \mathbb{Z}.$$

La fonction de répartition est définie par

$$F_X(x) = \Pr(X \leq x) = \sum_{z \leq x} p_X(z).$$

L'espérance mathématique d'une variable aléatoire discrète est définie de la manière suivante :

$$\mu = \mathrm{E}(X) = \sum_{x \in \mathbb{Z}} x p_X(x),$$

et sa variance

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}\left[\{X - \mathrm{E}(X)\}^2\right] = \sum_{x \in \mathbb{Z}} p_X(x)(x - \mu)^2 = \sum_{x \in \mathbb{Z}} p_X(x)x^2 - \mu^2.$$

On peut aussi calculer les moments et tous les autres paramètres.

5.4.2 Variable indicatrice ou bernoullienne

La variable indicatrice X de paramètre $p \in [0, 1]$ a la distribution de probabilités suivante :

$$X = \begin{cases} 1 \text{ avec une probabilité } p \\ 0 \text{ avec une probabilité } 1 - p. \end{cases}$$

L'espérance vaut

$$\mu = \mathrm{E}(X) = 0 \times (1-p) + 1 \times p = p,$$

et la variance vaut

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}(X-p)^2 = (1-p)(0-p)^2 + p(1-p)^2 = p(1-p).$$

Exemple 5.8 On tire au hasard une boule dans une urne contenant 18 boules rouges et 12 boules blanches. Si X vaut 1 si la boule est rouge et 0 sinon, alors X a une loi bernoullienne de paramètre $p = 18/(18+12) = 0.6$.

5.4.3 Variable binomiale

La variable aléatoire binomiale de paramètres n et p correspond à l'expérience suivante. On renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p, où p est la probabilité de succès pour une expérience élémentaire. Ensuite, on note X le nombre de succès obtenus. Le nombre de succès est une variable aléatoire prenant des valeurs entières de 0 à n et ayant une distribution binomiale.

Une variable X suit une loi binomiale de paramètre $0 < p < 1$ et d'exposant n, si

$$\Pr(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \ldots, n-1, n,$$

où $q = 1 - p$, et

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

De manière synthétique, si X a une distribution binomiale, on note :

$$X \sim \mathcal{B}(n, p).$$

Rappel Cette variable est appelée binomiale car sa distribution de probabilités est un terme du développement du binôme de Newton $(p+q)^n$.

$$\begin{aligned}
(p+q)^0 &= 1 \\
(p+q)^1 &= p + q = 1 \\
(p+q)^2 &= p^2 + 2pq + q^2 = 1 \\
(p+q)^3 &= p^3 + 3p^2q + 3pq^2 + q^3 = 1 \\
(p+q)^4 &= p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4 = 1 \\
&\vdots \\
(p+q)^n &= \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = 1.
\end{aligned}$$

La somme de ces probabilités vaut 1. En effet

$$\sum_{x=0}^{n} \Pr(X = x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1.$$

L'espérance se calcule de la manière suivante :

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{n} x \Pr(X = x) \\
&= \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=1}^{n} x \binom{n}{x} p^x q^{n-x} \text{ (on peut enlever le terme } x = 0) \\
&= \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x q^{n-x} \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{(n-1)-(x-1)} \\
&= np \sum_{z=0}^{n-1} \binom{n-1}{z} p^z q^{(n-1)-z} \quad \text{(en posant } z = x-1) \\
&= np(p+q)^{n-1} \\
&= np.
\end{aligned}$$

Théorème 5.3 *La variance est donnée par*

$$\mathrm{var}(X) = npq.$$

Démonstration
Pour calculer cette variance, nous allons d'abbord calculer $\mathrm{E}[X(X-1)]$. Ce

résultat préliminaire nous permettra de déterminer ensuite la variance.

$$\begin{aligned}
\mathrm{E}[X(X-1)] &= \sum_{x=0}^{n} x(x-1)\Pr(X=x) \\
&= \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x q^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\binom{n}{x}p^x q^{n-x} \text{ (on peut enlever les termes } x=0 \text{ et } x=1) \\
&= \sum_{x=2}^{n} n(n-1)\binom{n-2}{x-2}p^x q^{n-x} \\
&= n(n-1)p^2 \sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}q^{(n-2)-(x-2)} \\
&= n(n-1)p^2 \sum_{z=0}^{n-2}\binom{n-2}{z}p^z q^{(n-2)-z} \quad \text{(en posant } z=x-2) \\
&= n(n-1)p^2(p+q)^{n-2} \\
&= n(n-1)p^2.
\end{aligned}$$

Comme

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$$

et que

$$\mathrm{E}[X(X-1)] = \mathrm{E}(X^2) - \mathrm{E}(X),$$

on obtient

$$\mathrm{var}(X) = \mathrm{E}[X(X-1)] + \mathrm{E}(X) - \mathrm{E}^2(X) = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq.$$

□

Exemple 5.9 On tire au hasard avec remise et de manière indépendante 5 boules dans une urne contenant 18 boules rouges et 12 boules blanches. Si X est le nombre de boules rouges obtenues, alors X a une loi binomiale de paramètre $p = 18/(18+12) = 0.6$, et d'exposant $n = 5$. Donc,

$$\Pr(X=x) = \binom{5}{x}0.6^x 0.4^{5-x}, x = 0, 1, \ldots, 4, 5,$$

ce qui donne

$$\begin{aligned}
\Pr(X=0) &= \frac{5!}{0!(5-0)!}0.6^0 \times 0.4^{5-0} = 1 \times 0.4^5 = 0.01024 \\
\Pr(X=1) &= \frac{5!}{1!(5-1)!}0.6^1 \times 0.4^{5-1} = 5 \times 0.6^1 \times 0.4^4 = 0.0768 \\
\Pr(X=2) &= \frac{5!}{2!(5-2)!}0.6^2 \times 0.4^{5-2} = 10 \times 0.6^2 \times 0.4^3 = 0.2304 \\
\Pr(X=3) &= \frac{5!}{3!(5-3)!}0.6^3 \times 0.4^{5-3} = 10 \times 0.6^3 \times 0.4^2 = 0.3456 \\
\Pr(X=4) &= \frac{5!}{4!(5-4)!}0.6^4 \times 0.4^{5-4} = 5 \times 0.6^4 \times 0.4^1 = 0.2592 \\
\Pr(X=5) &= \frac{5!}{5!(5-5)!}0.6^5 \times 0.4^{5-5} = 1 \times 0.6^5 = 0.07776.
\end{aligned}$$

La distribution de probabilités de la variable X est présentée dans la Figure 5.2.

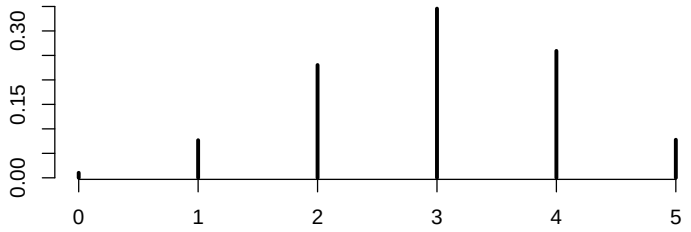


FIGURE 5.2 – Distribution d'une variable aléatoire binomiale avec $n = 5$ et $p = 0.6$.

Exemple 5.10 Supposons que, dans une population d'électeurs, 60% des électeurs s'apprêtent à voter pour le candidat A et 40% pour le candidat B et que l'on sélectionne un échantillon aléatoire de 10 électeurs avec remise dans cette population. Soit X le nombre de personnes s'apprêtant à voter pour le candidat A dans l'échantillon. La variable X a une distribution binomiale de paramètres $n = 10$ et $p = 0.6$ et donc

$$\Pr(X=x) = \binom{10}{x} 0.6^x (0.4)^{10-x}, x = 0, 1, \ldots, n-1, n.$$

5.4.4 Variable de Poisson

La variable X suit une loi de Poisson, ou loi des événements rares, de paramètre $\lambda \in \mathbb{R}^+$ si

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, 3, \ldots.$$

On note alors $X \sim \mathcal{P}(\lambda)$. La somme des probabilités est bien égale à 1, en effet

$$\sum_{x=0}^{\infty} \Pr(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Cette loi exprime la probabilité de l'occurence d'un nombre d'événements dans un laps de temps fixe si ces événements se produisent avec un taux moyen connu (λ) et indépendamment du temps d'occurrence du dernier événement.

L'espérance et la variance d'une loi de Poisson sont égales au paramètre λ. En effet

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{\infty} x\Pr(X = x) \\
&= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\lambda\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \text{ en posant } z = x - 1 \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda.
\end{aligned}$$

En outre, il est possible de montrer que

$$\mathrm{var}(X) = \lambda.$$

La distribution de probabilités d'une variable de Poisson $\mathcal{P}(\lambda = 1)$ est présentée dans la Figure 5.3.

En langage R

```
#
# distributions de probabilités discrètes
#
```

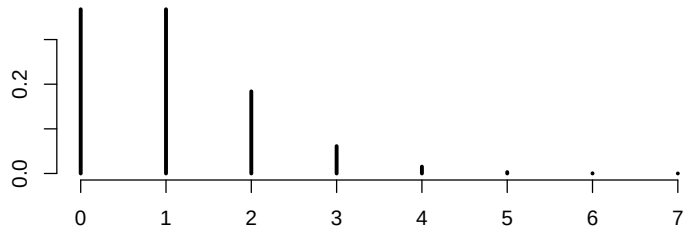


FIGURE 5.3 – Distribution d'une variable de Poisson avec $\lambda = 1$.

```
# nombre de faces obtenues en lançant deux pièces
plot(0:2,dbinom(0:2, 2,0.5),type = "h", lwd=3,
     ylim=c(0,0.5),xlab="",ylab="",xaxt = "n",frame = FALSE)
axis(1, 0:2, 0:2, col.axis = "blue")
# binomiale B(5,0.6)
plot(dbinom(0:5, 5,0.6),type = "h",
     lwd=3,xlab="",ylab="",main="",frame=FALSE)
# Poisson P(1)
plot(dpois(0:7, 1),type = "h",
     lwd=3,xlab="",ylab="",main="",frame=FALSE)
```

5.5 Variable aléatoire continue

5.5.1 Définition, espérance et variance

Une variable aléatoire continue prend des valeurs dans $\mathbb{R}$ ou dans un intervalle de $\mathbb{R}$.

La probabilité qu'une variable aléatoire continue soit inférieure à une valeur particulière est donnée par sa fonction de répartition.

$$\Pr(X \leq x) = F(x).$$

La fonction de répartition d'une variable aléatoire continue est toujours :
- dérivable,
- positive : $F(x) \geq 0$, pour tout x,
- croissante,
- $\lim_{x\to\infty} F(x) = 1$,
- $\lim_{x\to-\infty} F(x) = 0$.

On a

$$\Pr(a \leq X \leq b) = F(b) - F(a).$$

La fonction de densité d'une variable aléatoire continue est la dérivée de la fonction de répartition en un point

$$f(x) = \frac{dF(x)}{dx}.$$

Une fonction de densité est toujours :
- positive : $f(x) \geq 0$, pour tout x,
- d'aire égale à un : $\int_{-\infty}^{\infty} f(x)dx = 1$.

On a évidemment la relation :

$$F(b) = \int_{-\infty}^{b} f(x)dx.$$

La probabilité que la variable aléatoire soit inférieure à une valeur quelconque vaut :

$$\Pr(X \leq a) = \int_{-\infty}^{a} f(x)dx = F(a).$$

Dans la Figure 5.4, la probabilité $\Pr[X \leq a]$ est l'aire sous la densité de $-\infty$ à a.

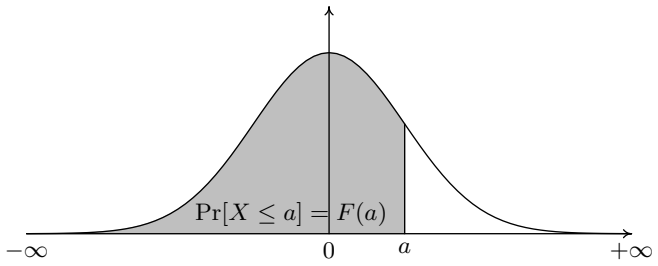


FIGURE 5.4 – Probabilité que la variable aléatoire soit inférieure à a

La probabilité que la variable aléatoire prenne une valeur comprise entre a et b vaut

$$\Pr(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a).$$

Si la variable aléatoire est continue, la probabilité qu'elle prenne exactement une valeur quelconque est nulle :

$$\Pr(X = a) = 0.$$

L'espérance d'une variable aléatoire continue est définie par :

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

et la variance

$$\sigma^2 = \mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

5.5.2 Variable uniforme

Une variable aléatoire X est dite uniforme dans un intervalle [a,b] (avec $a < b$), si sa répartition est :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ (x-a)/(b-a) & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b. \end{cases}$$

Sa densité est alors

$$f(x) = \begin{cases} 0 & \text{si } x < a \\ 1/(b-a) & \text{si } a \leq x \leq b \\ 0 & \text{si } x > b. \end{cases}$$

De manière synthétique, on écrit :

$$X \sim U(a, b).$$

Les logiciels génèrent en général des variables aléatoires uniformes dans [0,1]. Les Figures 5.5 et 5.6 représentent respectivement les fonctions de densité et de répartition d'une variable uniforme.

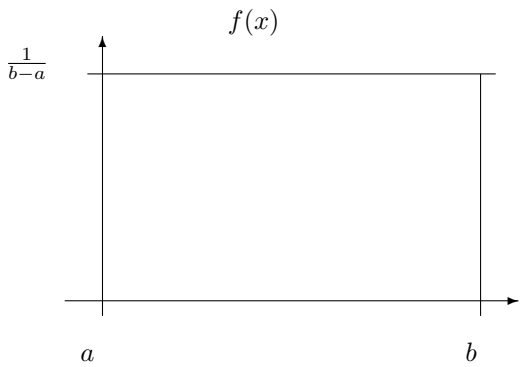


FIGURE 5.5 – Fonction de densité d'une variable uniforme

On peut calculer l'espérance et la variance :

Résultat 5.1

$$\mu = \mathrm{E}(X) = \frac{b+a}{2}$$

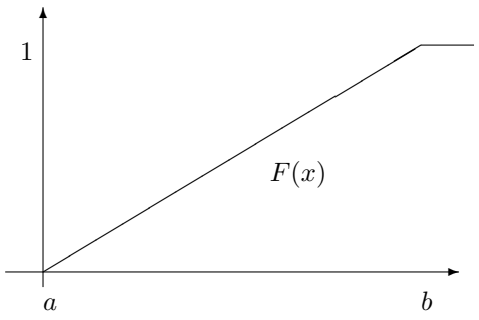


FIGURE 5.6 – Fonction de répartition d'une variable uniforme

Démonstration

$$\begin{aligned}
\mu &= \mathrm{E}(X) \\
&= \int_a^b x f(x) dx \\
&= \int_a^b x \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \int_a^b x dx \\
&= \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b \\
&= \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) \\
&= \frac{1}{b-a} \frac{1}{2} (b+a)(b-a) \\
&= \frac{a+b}{2}.
\end{aligned}$$

□

Résultat 5.2

$$\sigma^2 = \mathrm{var}(X) = \frac{(b-a)^2}{12}.$$

Démonstration
De manière générale, une variance peut toujours s'écrire comme un moment à

l'origine d'ordre 2 moins le carré de la moyenne. En effet,

$$\begin{aligned}
\sigma^2 &= \text{var}(X) \\
&= \int_a^b (x-\mu)^2 f(x)dx \\
&= \int_a^b (x^2+\mu^2-2x\mu) f(x)dx \\
&= \int_a^b x^2 f(x)dx + \int_a^b \mu^2 f(x)dx - 2\mu \int_a^b x f(x)dx \\
&= \int_a^b x^2 f(x)dx + \mu^2 - 2\mu^2 \\
&= \int_a^b x^2 f(x)dx - \mu^2.
\end{aligned}$$

On calcule ensuite un moment à l'origine d'ordre 2 :

$$\begin{aligned}
\int_a^b x^2 f(x)dx &= \int_a^b x^2 \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \int_a^b x^2 dx \\
&= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b \\
&= \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right) \\
&= \frac{1}{b-a} \frac{1}{3} (b^2+ab+a^2)(b-a) \\
&= \frac{b^2+ab+a^2}{3}.
\end{aligned}$$

On obtient enfin la variance par différence :

$$\begin{aligned}
\sigma^2 &= \int_a^b x^2 f(x)dx - \mu^2 \\
&= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} \\
&= \frac{4b^2+4ab+4a^2}{12} - \frac{3a^2+6ab+3b^2}{12} \\
&= \frac{b^2-2ab+a^2}{12} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

□

5.5.3 Variable normale

Une variable aléatoire X est dite normale si sa densité vaut

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \tag{5.1}$$

où $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$ sont les paramètres de la distribution. Le paramètre μ est appelé la moyenne et le paramètre σ l'écart-type de la distribution.

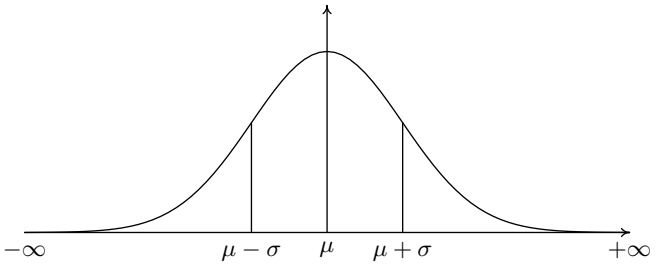


FIGURE 5.7 – Fonction de densité d'une variable normale

De manière synthétique, pour noter que X suit une loi normale (ou gaussienne, d'après Carl Friedrich Gauss) de moyenne μ et de variance σ^2 on écrit :

$$X \sim N(\mu, \sigma^2).$$

La loi normale est une des principales distributions de probabilité. Elle a de nombreuses applications en statistique. Sa fonction de densité dessine une courbe dite courbe de Gauss. On peut montrer (sans démonstration) que

$$\mathrm{E}(X) = \mu,$$

et

$$\mathrm{var}(X) = \sigma^2.$$

La fonction de répartition vaut

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2 du.$$

5.5.4 Variable normale centrée réduite

La variable aléatoire normale centrée réduite est une variable normale, d'espérance nulle, $\mu = 0$, et de variance $\sigma^2 = 1$. Sa fonction de densité vaut

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}.$$

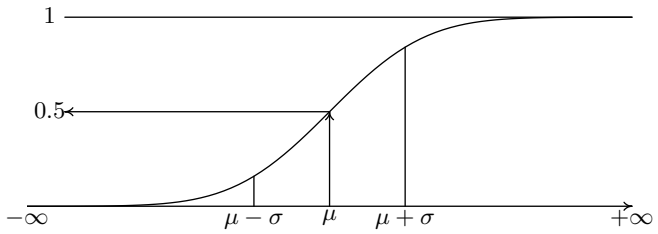


FIGURE 5.8 – Fonction de répartition d'une variable normale

et sa répartition vaut

$$\Phi(x) = F_{0,1}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp - \left(\frac{u^2}{2}\right) du.$$

Du fait de la symétrie de la densité, on a la relation

$$\Phi(-x) = 1 - \Phi(x),$$

qui se comprend facilement en examinant la Figure 5.9.

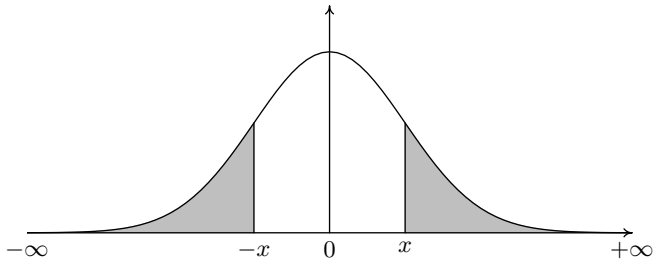


FIGURE 5.9 – Densité d'une normale centrée réduite, symétrie

De plus, le calcul de la répartition d'une variable normale de moyenne μ et de variance σ^2 peut toujours être ramené à une normale centrée réduite.

Résultat 5.3

$$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Démonstration
On a

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp - \left\{\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\} du.$$

En posant

$$z = \frac{u-\mu}{\sigma},$$

on obtient $u = z\sigma + \mu$, et donc $du = \sigma dz$. Donc,

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}} \exp - \left(\frac{z^2}{2}\right) \sigma dz = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

□

Les tables de la variable normale ne sont données que pour la normale centrée réduite. Les tables ne donnent $\Phi(x)$ que pour les valeurs positives de x, car les valeurs négatives peuvent être trouvées par la relation de symétrie.

5.5.5 Distribution exponentielle

Soit une variable aléatoire X qui définit la durée de vie d'un phénomène ou d'on objet. Si la durée de vie est *sans vieillissement*, c'est-à-dire la durée de vie au dela d'un instant T est indépendante de l'instant T, alors sa fonction de densité est donnée par :

$$f(x) = \begin{cases} \lambda \exp -(\lambda x), & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

On dit que X suit une loi exponentielle de paramètre λ positif. De manière synthétique, on écrit :

$$X \sim \varepsilon(\lambda).$$

Quand $x > 0$, sa fonction de répartition vaut :

$$F(x) = \int_0^x f(u)du = \int_0^x \lambda e^{-\lambda u} du = \left[-e^{-\lambda u}\right]_0^x = 1 - e^{-\lambda x}.$$

On peut alors calculer la moyenne :

Résultat 5.4 $\mathrm{E}(X) = \dfrac{1}{\lambda}$

Démonstration

$$\mathrm{E}(X) = \int_0^\infty x f(x)dx = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-\frac{1+x\lambda}{\lambda} e^{-\lambda x}\right]_0^\infty = \left(0 + \frac{1}{\lambda}\right) = \frac{1}{\lambda}.$$

□

Il est également possible de montrer que la variance vaut :

$$\mathrm{var}(X) = \frac{1}{\lambda^2}.$$

5.6 Distribution bivariée

Deux variables aléatoires peuvent avoir une distribution jointe.

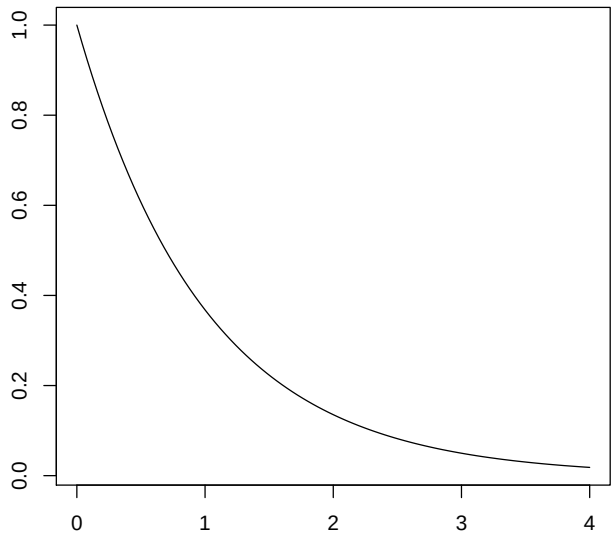


FIGURE 5.10 – Fonction de densité d'une variable exponentielle avec $\lambda = 1$.

5.6.1 Cas continu

Soit deux variables aléatoires X et Y continues, leur distribution de densité $f(x, y)$ est une fonction continue, positive, et telle que

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$$

La fonction de répartition jointe est définie par

$$F(x,y) = \Pr(X \leq x \text{ et } Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dvdu.$$

On appelle densités marginales les fonctions

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \text{ et } f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Avec les distributions marginales, on peut définir les moyennes marginales, et les variances marginales :

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x)dx, \text{ et } \mu_Y = \int_{-\infty}^{\infty} y f_Y(y)dy,$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx, \text{ et } \sigma_Y^2 = \int_{-\infty}^{\infty} (y-\mu_Y)^2 f_Y(y)dy.$$

On appelle densités conditionnelles, les fonctions

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \text{ et } f(y|x) = \frac{f(x,y)}{f_X(x)}.$$

Avec les distributions conditionnelles, on peut définir les moyennes conditionnelles, et les variances conditionnelles :

$$\mu_X(y) = \mathrm{E}(X|Y=y) = \int_{-\infty}^{\infty} xf(x|y)dx, \text{ et } \mu_Y(x) = \mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} yf(y|x)dy,$$

$$\sigma_X^2(y) = \mathrm{var}(X|Y=y) = \int_{-\infty}^{\infty} \{x-\mu_X(y)\}^2 f(x|y)dx, \text{ et } \sigma_Y^2(x) = \mathrm{var}(Y|X=x) = \int_{-\infty}^{\infty} \{y-\mu_Y(x)$$

Enfin, la covariance entre X et Y est définie par

$$\sigma_{xy} \quad = \quad \mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)f(x,y)dxdy.$$

5.6.2 Cas discret

Soit deux variables aléatoires X et Y discrètes, leur distribution de probabilité jointe $p(x,y)$ est telle que

$$\sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} p(x,y) = 1.$$

La fonction de répartition jointe est définie par

$$F(x,y) = \Pr(X \le x \text{ et } Y \le y) = \sum_{u\le x}\sum_{v\le v} p(u,v).$$

On appelle distributions de probabilté marginales les fonctions

$$p_X(x) = \sum_{y\in\mathbb{Z}} p(x,y), \text{ et } p_Y(y) = \sum_{x\in\mathbb{Z}} p(x,y).$$

Avec les distributions marginales, on peut définir les moyennes marginales, et les variances marginales :

$$\mu_X = \sum_{x\in\mathbb{Z}} xp_X(x), \text{ et } \mu_Y = \sum_{y\in\mathbb{Z}} yp_Y(y),$$

$$\sigma_X^2 = \sum_{x\in\mathbb{Z}} (x-\mu_X)^2 p_X(x), \text{ et } \sigma_Y^2 = \sum_{y\in\mathbb{Z}} (y-\mu_Y)^2 p_Y(y).$$

On appelle densités conditionnelles, les fonctions

$$p(x|y) = \frac{p(x,y)}{p_Y(y)} \text{ et } p(y|x) = \frac{p(x,y)}{p_X(x)}.$$

Avec les distributions conditionnelles, on peut définir les moyennes conditionnelles, et les variances conditionnelles :

$$\mu_X(y) = \sum_{x\in\mathbb{Z}} xp(x|y), \text{ et } \mu_Y(x) = \sum_{y\in\mathbb{Z}} yp(y|x),$$

$$\sigma^2_X(y) = \sum_{x\in\mathbb{Z}} \{x-\mu_X(y)\}^2 p(x|y), \text{ et } \sigma^2_Y(x) = \sum_{x\in\mathbb{Z}} \{y-\mu_Y(x)\}^2 p(y|x).$$

Enfin, la covariance entre X et Y est définie par

$$\sigma_{xy} = \text{cov}(X,Y) = \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} (x-\mu_X)(y-\mu_Y)p(x,y).$$

5.6.3 Remarques

Dans les deux cas discrets et continus, on peut toujours écrire

$$\begin{aligned}\text{var}(X) &= \text{E}[X-\text{E}(X)]^2 = \text{E}\left[X^2 - 2X\text{E}(X) + \text{E}^2(X)\right] \\ &= \text{E}(X^2) - 2\text{E}(X)\text{E}(X) + \text{E}^2(X) = \text{E}(X^2) - \text{E}^2(X).\end{aligned}$$

De même,

$$\text{var}(X|Y=y) = \text{E}\left\{[X-\text{E}(X|Y=y)]^2|Y=y\right\} = \text{E}(X^2|Y=y) - \text{E}^2(X|Y=y).$$

On a également

$$\begin{aligned}\text{cov}(X,Y) &= \text{E}[X-\text{E}(X)][Y-\text{E}(Y)] = \text{E}\left[XY - Y\text{E}(X) - X\text{E}(Y) + \text{E}(X)\text{E}(Y)\right] \\ &= \text{E}(XY) - \text{E}(X)\text{E}(Y) - \text{E}(X)\text{E}(Y) + \text{E}(X)\text{E}(Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y).\end{aligned}$$

L'opérateur espérance permet donc de définir la variance et la covariance.

5.6.4 Indépendance de deux variables aléatoires

Deux variables aléatoires X et Y sont dites indépendantes, si

$$\Pr(X \le x \text{ et } Y \le y) = \Pr(X \le x)\Pr(Y \le y), \text{ pour tout } x, y \in \mathbb{R}.$$

– Si X et Y sont discrètes, cela implique que

$$\Pr(X = x \text{ et } Y = y) = \Pr(X = x)\Pr(Y = y), \text{ pour tout } x, y \in \mathbb{Z}.$$

– Si X et Y sont continues, en notant $f_X(.)$ et $f_Y(.)$ les fonctions de densité respectives de X et Y, et en notant $f_{XY}(x,y)$ la densité jointe des deux variables, alors X et Y sont indépendants si

$$f_{XY}(x,y) = f_X(x)f_Y(y), x, y \in \mathbb{R}.$$

5.7 Propriétés des espérances et des variances

De manière générale, pour des variables aléatoires X et Y, et avec a et b constants, on a les résultats suivants qui sont démontrées pour le cas continu. Ces résultats sont également valables pour le cas discret pour lequel les démonstrations sont similaires.

Résultat 5.5

$$\mathrm{E}(a+bX) = a + b\mathrm{E}(X)$$

Démonstration

$$\mathrm{E}(a+bX) = \int_{\mathbb{R}} (a+bx)f(x)dx = a\int_{\mathbb{R}} f(x)dx + b\int_{\mathbb{R}} xf(x)dx = a + b\mathrm{E}(X).$$

□

Résultat 5.6

$$\mathrm{E}(aY+bX) = a\mathrm{E}(Y) + b\mathrm{E}(X).$$

Démonstration

$$\begin{aligned}
\mathrm{E}(aY+bX) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (ay+bx)f(x,y)dxdy \\
&= a\int_{\mathbb{R}}\int_{\mathbb{R}} yf(x,y)dxdy + b\int_{\mathbb{R}}\int_{\mathbb{R}} xf(x,y)dxdy \\
&= a\int_{\mathbb{R}} y\int_{\mathbb{R}} f(x,y)dxdy + b\int_{\mathbb{R}} x\int_{\mathbb{R}} f(x,y)dydx \\
&= a\int_{\mathbb{R}} yf(y)dy + b\int_{\mathbb{R}} xf(x)dx \\
&= a\mathrm{E}(Y) + b\mathrm{E}(X)
\end{aligned}$$

□

Quand a et b valent 1, on obtient que l'espérance de la somme de deux variables aléatoires est égale à la somme de leur espérances :

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Résultat 5.7

$$\mathrm{var}(a+bX) = b^2\mathrm{var}(X).$$

Démonstration

$$\begin{aligned}
\mathrm{var}(a+bX) &= \int_{\mathbb{R}} [a+bx-\mathrm{E}(a+bX)]^2 f(x)dx \\
&= \int_{\mathbb{R}} [a+bx-(a+b\mathrm{E}(X))]^2 f(x)dx \\
&= \int_{\mathbb{R}} [bx-b\mathrm{E}(X)]^2 f(x)dx \\
&= b^2 \int_{\mathbb{R}} [x-\mathrm{E}(X)]^2 f(x)dx \\
&= b^2 \mathrm{var}(X).
\end{aligned}$$

□ La variance n'est donc pas sensible à un changement d'origine, mais est affectée par le carré d'un changement d'unité.

Résultat 5.8

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y).$$

Démonstration

$$\begin{aligned}
\mathrm{var}(X+Y) &= \int_{\mathbb{R}}\int_{\mathbb{R}} [x+y-\mathrm{E}(X+Y)]^2 f(x,y)dxdy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} [x-\mathrm{E}(X)+y-\mathrm{E}(Y)]^2 f(x,y)dxdy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \left\{[x-\mathrm{E}(X)]^2+[y-\mathrm{E}(Y)]^2+2[x-\mathrm{E}(X)][y-\mathrm{E}(Y)]\right\} f(x,y)dxdy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} [x-\mathrm{E}(X)]^2 f(x,y)dxdy + \int_{\mathbb{R}}\int_{\mathbb{R}} [y-\mathrm{E}(Y)]^2] f(x,y)dxdy \\
&\quad +2\int_{\mathbb{R}}\int_{\mathbb{R}} [x-\mathrm{E}(X)][y-\mathrm{E}(Y)] f(x,y)dxdy \\
&= \int_{\mathbb{R}} [x-\mathrm{E}(X)]^2 \int_{\mathbb{R}} f(x,y)dydx + \int_{\mathbb{R}} [y-\mathrm{E}(Y)]^2] \int_{\mathbb{R}} f(x,y)dxdy + 2\mathrm{cov}(X,Y) \\
&= \int_{\mathbb{R}} [x-\mathrm{E}(X)]^2 f_X(x)dx + \int_{\mathbb{R}} [y-\mathrm{E}(Y)]^2] f_y(Y)dy + 2\mathrm{cov}(X,Y) \\
&= \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y)
\end{aligned}$$

□

Résultat 5.9 *De plus, si X et Y sont indépendantes, on a $f(x,y) = f_X(x)_Y f(y)$ pour tout x, y*

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y).$$

Démonstration

$$\begin{aligned} \mathrm{E}(XY) &= \int_{\mathbb{R}}\int_{\mathbb{R}} xyf_X(x)f_Y(y)dxdy \\ &= \int_{\mathbb{R}} xf_X(x)dx \int_{\mathbb{R}} yf_Y(y)dy \\ &= \mathrm{E}(X)\mathrm{E}(Y). \end{aligned}$$

□

Comme, de manière générale $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$, on déduit directement du Résultat 5.9 que, si X et Y sont indépendantes, on a $\mathrm{cov}(X,Y) = 0$, et donc

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Attention, la réciproque n'est pas vraie. Une covariance nulle n'implique pas que les deux variables sont indépendantes.

Enfin, il est possible de calculer l'espérance et la variance d'une somme de variables aléatoires indépendantes, et identiquement distribuées.

Théorème 5.4 *Soit $X_1, \ldots, X_n$ une suite de variables aléatoires, indépendantes et identiquement distribuées et dont la moyenne μ et la variance σ^2 existent et sont finies, alors si*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

on a

$$\mathrm{E}(\bar{X}) = \mu, \text{ et } \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Démonstration

$$\mathrm{E}\left(\bar{X}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

et

$$\mathrm{var}\left(\bar{X}\right) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

□

5.8 Autres variables aléatoires

5.8.1 Variable khi-carrée

Soit une suite de variables aléatoires indépendantes, normales centrées réduites, $X_1, \ldots, X_p$, (c'est-à-dire de moyenne nulle et de variance égale à 1), alors la variable aléatoire

$$\chi^2_p = \sum_{i=1}^{p} X_i^2,$$

est appelée variable aléatoire khi-carré à p degrés de liberté.

Il est possible de montrer que

$$\mathrm{E}(\chi_p^2) = p,$$

et que

$$\mathrm{var}(\chi_p^2) = 2p.$$

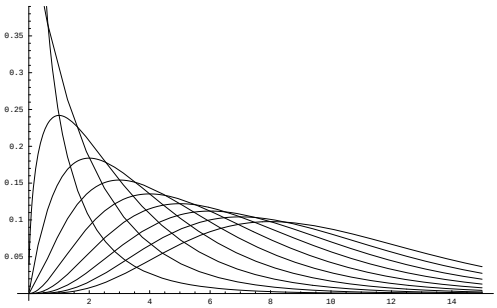


FIGURE 5.11 – Densité d'une variable de chi-carré avec $p = 1, 2, \ldots, 10$

5.8.2 Variable de Student

Soit une variable aléatoire X normale centrée réduite, et une variable aléatoire khi-carré χ_p^2 à p degrés de liberté, indépendante de X, alors la variable aléatoire

$$t_p = \frac{X}{\sqrt{\chi_p^2/p}}$$

est appelée variable aléatoire de Student à p degrés de liberté.

5.8.3 Variable de Fisher

Soient deux variables aléatoires khi-carrés indépendantes χ_p^2, χ_q^2, respectivement à p et q degrés de liberté, alors la variable aléatoire

$$F_{p,q} = \frac{\chi_p^2/p}{\chi_q^2/q}$$

est appelée variable aléatoire de Fisher à p et q degrés de liberté.

Remarque 5.1 Il est facile de montrer que le carré d'une variable de Student à q degrés de liberté est une variable de Fisher à 1 et q degrés de liberté.

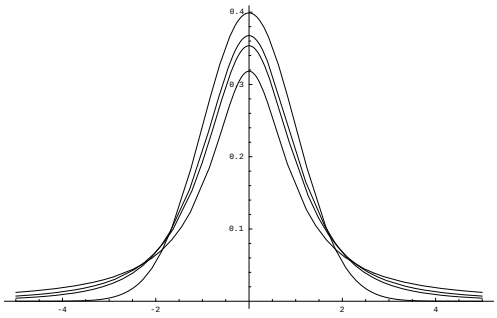


FIGURE 5.12 – Densités de variables de Student avec $p = 1, 2$ et 3 et d'une variable normale

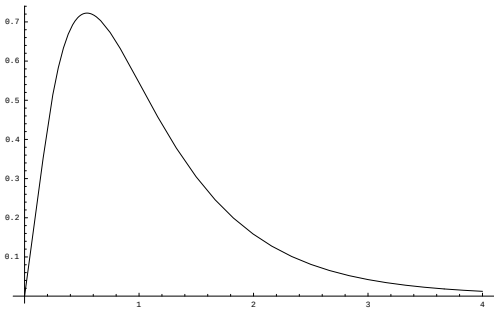


FIGURE 5.13 – Densité d'une variable de Fisher

5.8.4 Loi normale bivariée

Les variables X et Y suivent une loi normale bivariée si leur densité jointe est donnée par

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}. \tag{5.2}$$

La fonction de densité dépend de 5 paramètres

- les deux moyennes marginales $\mu_x \in \mathbb{R}$ et $\mu_y \in \mathbb{R}$,
- les deux variances marginales $\sigma_x^2 > 0$ et $\sigma_y^2 > 0$,
- le coefficient de corrélation $-1 < \rho < 1$.

Un exemple de normale bivariée est présentée dans la Figure 5.14.

La Figure 5.15 montre le nuage de points de 1000 réalisations d'une normale bivariée avec les paramètres suivants : $\mu_x = 8$, $\mu_y = 20$, $\sigma_x^2 = 9$, $\sigma_y^2 = 25$, $\rho = 0.6$.

En langage R

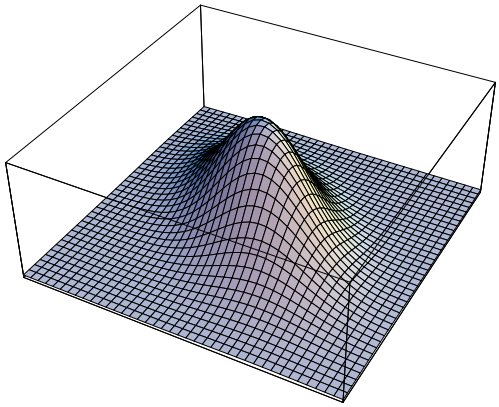


FIGURE 5.14 – Densité d'une normale bivariée

```
a=8; b=3 ;c=12 ; d=4
X=a+ b*rnorm(2000)
Y=c+X+d*rnorm(2000)
plot(X,Y,type="p")
```

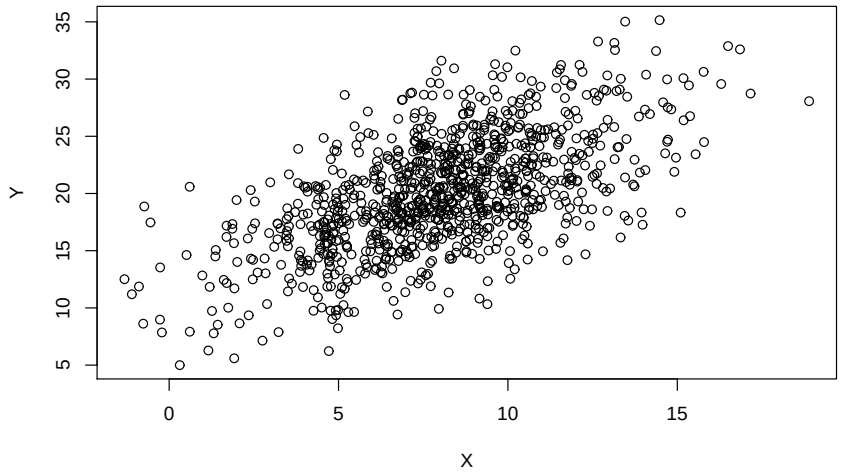


FIGURE 5.15 – Nuage de points de réalisations d'une normale bivariée

Théorème 5.5 *Les deux distributions marginales d'une distribution normale*

bivariée ont une distribution normale donnée par :

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sigma_x\sqrt{2\pi}} \exp - \frac{(x-\mu_x)^2}{2\sigma_x^2}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{\sigma_y\sqrt{2\pi}} \exp - \frac{(y-\mu_y)^2}{2\sigma_y^2}$$

Démonstration (pour $f_X(x)$)
On peut vérifier que la densité jointe peut également s'écrire :

$$f(x,y) = \left(\frac{1}{\sigma_x\sqrt{2\pi}} \exp - \frac{(x-\mu_x)^2}{2\sigma_x^2}\right) \frac{1}{\sigma_y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_y(x)}{\sigma_y(x)}\right)^2\right\},$$

où

$$\mu_y(x) = \mu_y + \frac{\sigma_y\rho}{\sigma_x}(x-\mu_x) \text{ et } \sigma_y^2(x) = \sigma_y^2(1-\rho^2).$$

On a

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \left(\frac{1}{\sigma_x\sqrt{2\pi}} \exp - \frac{(x-\mu_x)^2}{2\sigma_x^2}\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sigma_y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_y(x)}{\sigma_y(x)}\right)^2\right\} dy}_{=1}. \end{aligned}$$

□

Le Théorème 5.5 montre que les deux distributions marginales sont normales, que μ_x et μ_y sont les moyennes marginales, et que σ_x^2 et σ_x^2 sont les deux variance marginales de la distribution jointes. On peut également montrer à partir du Théorème 5.5 que le volume sous la courbe vaut bien 1. En effet

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} f_Y(y)dy = 1.$$

Attention, la réciproque du Théorème 5.5 n'est pas nécessairement vraie. Une distribution bivariée dont les deux distributions marginales sont normales, n'est pas nécessairement normale.

Théorème 5.6 *Toutes les distributions conditionnelles d'une distribution normale bivariée ont une distribution normale donnée par :*

$$f(y|x) = \frac{1}{\sigma_y(x)\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{y-\mu_y(x)}{\sigma_y(x)}\right)^2\right\}$$

où

$$\mu_y(x) = \mu_y + \frac{\sigma_y\rho}{\sigma_x}(x-\mu_x) \text{ et } \sigma_y^2(x) = \sigma_y^2(1-\rho^2).$$

et

$$f(x|y) = \frac{1}{\sigma_x(y)\sqrt{2\pi}} \exp\left\{ \frac{-1}{2} \left(\frac{x - \mu_x(y)}{\sigma_x(y)} \right)^2 \right\}$$

où

$$\mu_x(y) = \mu_x + \frac{\sigma_x \rho}{\sigma_y}(y - \mu_y) \text{ et } \sigma_x^2(y) = \sigma_x^2(1 - \rho^2).$$

Démonstration (pour $f(y|x)$)

$$\begin{aligned}
f(y|x) &= \frac{f(x,y)}{f_X(x)} \\
&= \frac{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\}}{\frac{1}{\sigma_x\sqrt{2\pi}} \exp - \frac{(x-\mu_x)^2}{2\sigma_x^2}} \\
&= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] + \frac{(x-\mu_x)^2}{2\sigma_x^2} \right\} \\
&= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{\rho^2(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\} \\
&= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \exp\left\{ \frac{-1}{2\sqrt{1-\rho^2}} \left(\frac{y-\mu_y}{\sigma_y} - \frac{\rho(x-\mu_x)}{\sigma_x} \right)^2 \right\} \\
&= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \exp\left\{ \frac{-1}{2\sqrt{1-\rho^2}} \left(\frac{y-\mu_y - \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)}{\sigma_y} \right)^2 \right\} \\
&= \frac{1}{\sigma_y(x)\sqrt{2\pi}} \exp\left\{ \frac{-1}{2} \left(\frac{y-\mu_y(x)}{\sigma_y(x)} \right)^2 \right\}.
\end{aligned}$$

□

Le Theorème 5.6 montre que toutes les distributions conditionnelles sont également normales. La variance conditionnelle de Y pour une valeur fixée de x de la variable X vaut :

$$\mathrm{E}(Y|X = x) = \mu_y(x) = \mu_y + \frac{\sigma_y \rho}{\sigma_x}(x - \mu_x).$$

De même, l'espérance conditionnelle de X pour une valeur fixée de y de la variable Y vaut :

$$\mathrm{E}(X|Y = y) = \mu_x(y) = \mu_x + \frac{\sigma_x \rho}{\sigma_y}(y - \mu_y).$$

La variance conditionnelle de Y pour une valeur fixée de x de la variable X vaut :

$$\mathrm{var}(Y|X = x) = \sigma_y^2(x) = \sigma_y^2(1 - \rho^2).$$

Cette variance conditionnelle ne dépend pas de x. La variance conditionnelle de X pour une valeur fixée de y de la variable Y vaut :

$$\text{var}(X|Y=y) = \sigma_x^2(y) = \sigma_x^2(1-\rho^2),$$

et ne dépend pas de y. Cette variance conditionnelle ne dépend pas de y. Les variances conditionnelles sont donc homoscédastiques (même variance).

Théorème 5.7

$$\text{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)dydx = \sigma_x\sigma_y\rho.$$

Démonstration
La covariance peut également s'écrire

$$\text{cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dydx - \mu_x\mu_y.$$

On a :

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_X(x)f(y|x)dydx = \int_{-\infty}^{\infty}xf_X(x)\int_{-\infty}^{\infty}yf(y|x)dydx \\
&= \int_{-\infty}^{\infty}xf_X(x)\left[\mu_y + \frac{\sigma_y\rho}{\sigma_x}(x-\mu_x)\right]dx = \mu_y\int_{-\infty}^{\infty}xf_X(x)dx + \frac{\sigma_y\rho}{\sigma_x}\int_{-\infty}^{\infty}xf_X(x)(x-\mu_x)dx \\
&= \mu_y\mu_x + \frac{\sigma_y\rho}{\sigma_x}\sigma_x^2 = \mu_y\mu_x + \sigma_x\sigma_y\rho.
\end{aligned}$$

Donc

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)dxdy = \sigma_x\sigma_y\rho.$$

□

Le paramètre ρ est bien un coefficient de corrélation entre les variables X et X car il peut s'écrire :

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_x\sigma_y\rho}{\sigma_x\sigma_y} = \rho.$$

Théorème 5.8 *Si les deux variables X et Y ont une distribution normale bivariée et que leur coefficient de corrélation est nul, alors X et Y sont indépendantes.*

Démonstration
Si $\rho = 0$, alors de l'Expression 5.2, la distribution jointe vaut :

$$\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left\{-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right\}\right) \\
&= f_X(x)f_Y(y).
\end{aligned}$$

Dans ce cas, la densité jointe peut s'écrire comme le produit des deux densités marginales. Les deux variables sont donc indépendantes. □

Attention, si les deux variables n'ont pas une distribution normale bivariée, une covariance nulle n'implique plus que les variables sont indépendantes.

Exercices

Exercice 5.1 Soit $Z \sim N(0,1)$. Déterminez :

1. $\Pr[Z \leq 1,23]$;
2. $\Pr[Z \leq -1,23]$;
3. $\Pr[Z \in [0,36; 1,23]]$;
4. $\Pr[Z \in [-0,88; 1,23]]$;
5. $\Pr[Z > 2,65 \text{ ou } Z \leq -1,49]$.

Solution

1. $\Pr[Z \leq 1,23] = F(1,23) = 0,8907$
2. $\Pr[Z \leq -1,23] = 1 - F(1,23) = 0.1093$
3. $\Pr[Z \in [0,36; 1,23]] = F(1,23) - F(0,36) = 0,8907 - 0,6406 = 0,2501$
4. $\Pr[Z \in [-0,88; 1,23] = F(1,23) - F(-0,88) = 0,8907 - (1 - F(0,88)) = 0,8907 - 0,1894 = 0,7013$
5. $\Pr[Z > 2,65 \text{ ou } Z \leq -1,49] = \Pr[Z > 2,65] + \Pr[Z \leq -1,49] = 1 - F(2,65) + F(-1,49) = 1 - F(2,65) + 1 - F(1,49) = 2 - 0,9960 - 0,9319 = 0,0721$

Exercice 5.2 Déterminez les valeurs j de la variable normale centrée réduite Z telles que :

1. $\Pr[Z \leq j] = 0,9332$;
2. $\Pr[-j \leq Z \leq j] = 0,3438$;
3. $\Pr[Z \leq j] = 0,0125$;
4. $\Pr[Z \geq j] = 0,0125$;
5. $\Pr[j \leq Z \leq 3] = 0,7907$.

Solution
Lecture inverse de la table.

1. $\Pr[Z \leq j] = 0,9332 \Rightarrow F(j) = 0,9332 \Rightarrow j = 1,5$

2. $\Pr[-j \leq Z \leq j] = 0,3438 \Rightarrow F(j) - F(-j) = F(j) - 1 + F(j) = 2F(j) - 1 = 0,3438 \Rightarrow F(j) = 0,6719 \Rightarrow j = 0,45$
3. $\Pr[Z \leq j] = 0,0125 \Rightarrow F(j) = 0,0125 (j \text{ est négatif}) \Rightarrow 1 - F(-j) = 0,0125 \Rightarrow F(-j) = 0,9875 \Rightarrow j = -2,24$
4. $\Pr[Z \geq j] = 0,0125 = 1 - F(j) \Rightarrow F(j) = 0,9875 \Rightarrow j = 2,24$
5. $\Pr[j \leq Z \leq 3] = 0,7907 = F(3) - F(j) \Rightarrow 0,7907 = 0,9987 - F(j) \Rightarrow F(j) = 0,2080$ (négatif) $\Rightarrow F(-j) = 0,7920 \Rightarrow -j = 0,81 \Rightarrow j = -0,81.$

Exercice 5.3 Soit une variable aléatoire $X \sim N(53; \sigma^2 = 100)$ représentant le résultat d'un examen pour un étudiant d'une section. Déterminez la probabilité pour que le résultat soit compris entre 33,4 et 72,6.

Solution

$$\text{Soit } X \sim N(53, 100) \Rightarrow Z = \frac{X - 53}{10} \sim N(0, 1)$$

$$\begin{aligned} &\Pr[33,4 \leq X \leq 72,6] \\ &= \Pr\left[\frac{33,4 - 53}{10} \leq \frac{X - 53}{10} \leq \frac{72,6 - 53}{10}\right] \\ &= \Pr[-1,96 \leq Z \leq 1,96] \\ &= 2\mathrm{F}(1,96) - 1 = 2\ 0,975 - 1 \\ &= 0,95 \end{aligned}$$

Exercice 5.4 Soit une variable aléatoire X $N(50; \sigma^2 = 100)$. Déterminez le premier quartile de cette distribution.

Solution
Si $X \sim N(50, 10)$, alors $Z = (X - 50)/10 \sim N(0, 1)$. Par définition le premier quartile $x_{1/4}$ est tel que

$$\Pr\left[X \leq x_{1/4}\right] = 1/4.$$

Donc

$$\begin{aligned} \Pr\left[X \leq x_{1/4}\right] &= P\left[\frac{X - 50}{10} \leq \frac{\chi_{1/4} - 50}{10}\right] \\ &= P[Z \leq z_{1/4}] = 0,25, \end{aligned}$$

où $z_{1/4}$ est le premier quartile d'une variable aléatoire normale centrée réduite. Si $F(.)$ est la fonction de répartition d'une variable aléatoire normale centrée réduite, on a par la définition du quartile que

$$F(z_{1/4}) = 0,25.$$

Le premier quartile $z_{1/4}$ est donc négatif. On a cependant. par la symétrie de la distribution, que

$$F(z_{1/4}) = 1 - F(-z_{1/4}) = 0,25,$$

ce qui donne

$$F(-z_{1/4}) = 0,75.$$

La table nous donne que $-z_{1/4} = 0,67$ et donc $z_{1/4} = -0,67$. Enfin, comme

$$\frac{x_{1/4} - 50}{10} = z_{1/4} = -0,67,$$

on a une équation en $x_{1/4}$ qu'il suffit de résoudre

$$x_{1/4} = 50 - 0,67 \times 10 = 43,3.$$

Exercice 5.5 En supposant que les tailles en cm des étudiants d'un pays admettent la distribution normale $N(172; \sigma^2 = 9)$. On demande de déterminer le pourcentage théorique :
a) d'étudiants mesurant au moins 180 cm.
b) d'étudiants dont la taille est comprise entre 168 et 180.

Solution
a) 0,0038 ; b) 0,9044.

Exercice 5.6 Sur une route principale où la vitesse est limitée à 80 km/h, un radar a mesuré la vitesse de toutes les automobiles pendant une journée. En supposant que les vitesses recueillies soient distribuées normalement avec une moyenne de 72 km/h et un écart-type de 8 km/h, quelle est approximativement la proportion d'automobiles ayant commis un excès de vitesse ?

Solution
La proportion d'automobiles ayant commis un excès de vitesse vaut

$$P[X > 80] = 1 - P[X \leq 80] = 1 - P\left[\frac{X - \bar{x}}{s} \leq \frac{80 - 72}{8}\right] = 1 - P[Z \leq 1] = 0,159,$$

où X représente la vitesse.

Exercice 5.7 Pour l'assemblage d'une machine, on produit des cylindres dont le diamètre varie d'après une loi normale de moyenne 10 cm et d'écart-type 0,2 cm. On groupe les cylindres en 3 catégories :
A : défectueux et inutilisable si le diamètre est $\leq$ 9.95, le cylindre est alors détruit.
B : utilisable et vendu au prix réduit de Fr. 5.-, si 9,95 le diamètre $\leq$ 9,99.
C : correspond aux normes et est vendu Fr. 15.-, si le diamètre est ¿ 9,99.
a) Calculer les proportions de cylindres produits de chaque type A, B et C.
b) La production d'un cylindre coûte Fr. 7.-. Quel est le profit moyen par cylindre produit ?

Solution
a) Soit X le diamètre, ainsi $X \sim N(10, 0.2^2)$

$$P[X \leq 9.95] = P\left[\frac{X-10}{0.2} \leq 0.25\right] = 0.401$$

$$P[9.95 < X \leq 9.99] = P\left[-0.25 < \frac{X-10}{0.2} \leq -0.05\right] = 0.079,$$

$$P[X > 9.99] = 1 - (P[X \leq 9.95] + P[9.95 < X \leq 9.99] = 0.52.$$

b) profit= 5*0.079 + 15*0.52 − 7 = 1.195 fr.

Exercice 5.8 Donnez les quantiles d'ordre 99%, 97.5% et 95% :

1. d'une variable normale centrée réduite ;
2. d'une variable Khi-carrée à 17 degrés de liberté ;
3. d'une variable de Student à 8 degrés de liberté ;
4. d'une variable de Fisher (uniquement d'ordre 95%) à 5 et 7 degrés de liberté.

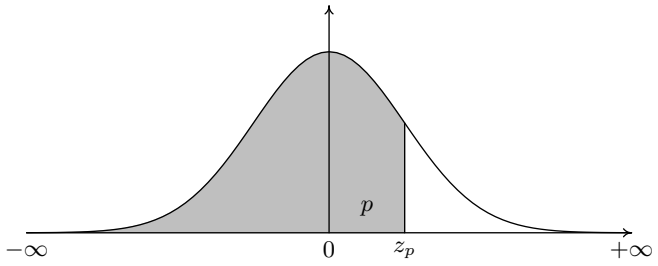
Solution

1. à 99% : 2.3263, à 97.5% : 1.9600, à 95% : 1.6449 ;
2. à 99% : 33.41, à 97.5% : 30.19, à 95% : 27.59 ;
3. à 99% : 2.896, à 97.5% : 2.306, à 95% : 1.860 ;
4. à 95% : 3.972.

Chapitre 7

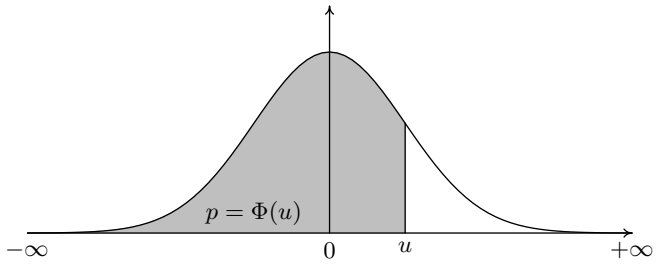
Tables statistiques

TABLE 7.1 – Table des quantiles $z_p = \Phi^{-1}(p)$ d'une variable normale centrée réduite



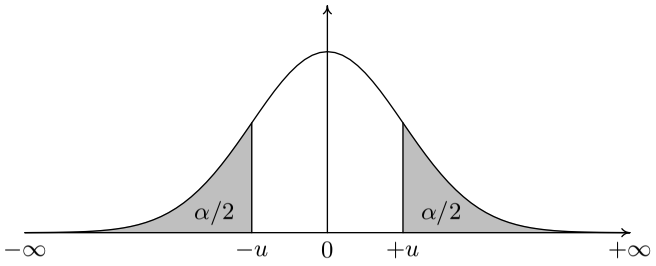
Ordre du quantile (p)	quantile (z_p)	Ordre du quantile (p)	Quantile (z_p)
0.500	0.0000	0.975	1.9600
0.550	0.1257	0.976	1.9774
0.600	0.2533	0.977	1.9954
0.650	0.3853	0.978	2.0141
0.700	0.5244	0.979	2.0335
0.750	0.6745	0.990	2.3263
0.800	0.8416	0.991	2.3656
0.850	1.0364	0.992	2.4089
0.900	1.2816	0.993	2.4573
0.950	1.6449	0.994	2.5121
0.970	1.8808	0.995	2.5758
0.971	1.8957	0.996	2.6521
0.972	1.9110	0.997	2.7478
0.973	1.9268	0.998	2.8782
0.974	1.9431	0.999	3.0902

TABLE 7.2 – Fonction de répartition de la loi normale centrée réduite
(*Probabilité de trouver une valeur inférieur à u*)



u	0.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

TABLE 7.3 – quantiles de la loi normale centrée réduite
(*u : valeur ayant la probabilité α d'être dépassé en valeur absolue*)



α	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4958	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

TABLE 7.4 – Table des quantiles d'une variable χ^2 à n degrés de liberté

	ordre du quantile					
	0.01	0.025	0.05	0.95	0.975	0.99
n=1	0.000157	0.000982	0.003932	3.841	5.024	6.635
2	0.02010	0.05064	0.103	5.991	7.378	9.210
3	0.115	0.216	0.352	7.815	9.348	11.34
4	0.297	0.484	0.711	9.488	11.14	13.28
5	0.554	0.831	1.145	11.07	12.83	15.09
6	0.872	1.237	1.635	12.59	14.45	16.81
7	1.239	1.690	2.167	14.07	16.01	18.48
8	1.646	2.180	2.733	15.51	17.53	20.09
9	2.088	2.700	3.325	16.92	19.02	21.67
10	2.558	3.247	3.940	18.31	20.48	23.21
11	3.053	3.816	4.575	19.68	21.92	24.72
12	3.571	4.404	5.226	21.03	23.34	26.22
13	4.107	5.009	5.892	22.36	24.74	27.69
14	4.660	5.629	6.571	23.68	26.12	29.14
15	5.229	6.262	7.261	25.00	27.49	30.58
16	5.812	6.908	7.962	26.30	28.85	32.00
17	6.408	7.564	8.672	27.59	30.19	33.41
18	7.015	8.231	9.390	28.87	31.53	34.81
19	7.633	8.907	10.12	30.14	32.85	36.19
20	8.260	9.591	10.85	31.41	34.17	37.57
21	8.897	10.28	11.59	32.67	35.48	38.93
22	9.542	10.98	12.34	33.92	36.78	40.29
23	10.20	11.69	13.09	35.17	38.08	41.64
24	10.86	12.40	13.85	36.42	39.36	42.98
25	11.52	13.12	14.61	37.65	40.65	44.31
26	12.20	13.84	15.38	38.89	41.92	45.64
27	12.88	14.57	16.15	40.11	43.19	46.96
28	13.56	15.31	16.93	41.34	44.46	48.28
29	14.26	16.05	17.71	42.56	45.72	49.59
30	14.95	16.79	18.49	43.77	46.98	50.89
31	15.66	17.54	19.28	44.99	48.23	52.19
32	16.36	18.29	20.07	46.19	49.48	53.49
33	17.07	19.05	20.87	47.40	50.73	54.78
34	17.79	19.81	21.66	48.60	51.97	56.06
35	18.51	20.57	22.47	49.80	53.20	57.34
36	19.23	21.34	23.27	51.00	54.44	58.62
37	19.96	22.11	24.07	52.19	55.67	59.89
38	20.69	22.88	24.88	53.38	56.90	61.16
39	21.43	23.65	25.70	54.57	58.12	62.43
40	22.16	24.43	26.51	55.76	59.34	63.69
42	23.65	26.00	28.14	58.12	61.78	66.21
44	25.15	27.57	29.79	60.48	64.20	68.71
46	26.66	29.16	31.44	62.83	66.62	71.20
48	28.18	30.75	33.10	65.17	69.02	73.68
50	29.71	32.36	34.76	67.50	71.42	76.15
60	37.48	40.48	43.19	79.08	83.30	88.38
70	45.44	48.76	51.74	90.53	95.02	100.43
80	53.54	57.15	60.39	101.88	106.63	112.33
90	61.75	65.65	69.13	113.15	118.14	124.12
100	70.06	74.22	77.93	124.34	129.56	135.81
110	78.46	82.87	86.79	135.48	140.92	147.41
120	86.92	91.57	95.70	146.57	152.21	158.95

TABLE 7.5 – Table des quantiles d'une variable de Student à n degrés de liberté

	ordre du quantile			
	0.95	0.975	0.99	0.995
n=1	6.314	12.71	31.82	63.66
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
31	1.696	2.040	2.453	2.744
32	1.694	2.037	2.449	2.738
33	1.692	2.035	2.445	2.733
34	1.691	2.032	2.441	2.728
35	1.690	2.030	2.438	2.724
36	1.688	2.028	2.434	2.719
37	1.687	2.026	2.431	2.715
38	1.686	2.024	2.429	2.712
39	1.685	2.023	2.426	2.708
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
60	1.671	2.000	2.390	2.660
70	1.667	1.994	2.381	2.648
80	1.664	1.990	2.374	2.639
90	1.662	1.987	2.368	2.632
100	1.660	1.984	2.364	2.626
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.327	2.576

TABLE 7.6 – Table des quantiles d'ordre 0.95 d'une variable de Fisher à n_1 et n_2 degrés de liberté

	n_1=1	2	3	4	5	6	7	8	9	10	12	14	16	20	30	∞
n_2=1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.4	246.5	248.0	250.1	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.45	19.46	19.50
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.715	8.692	8.660	8.617	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.873	5.844	5.803	5.746	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.636	4.604	4.558	4.496	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.956	3.922	3.874	3.808	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.529	3.494	3.445	3.376	3.230
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.237	3.202	3.150	3.079	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.025	2.989	2.936	2.864	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.865	2.828	2.774	2.700	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.739	2.701	2.646	2.570	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.637	2.599	2.544	2.466	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.554	2.515	2.459	2.380	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.484	2.445	2.388	2.308	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.424	2.385	2.328	2.247	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.373	2.333	2.276	2.194	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.329	2.289	2.230	2.148	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.290	2.250	2.191	2.107	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.256	2.215	2.155	2.071	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.225	2.184	2.124	2.039	1.843
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.197	2.156	2.096	2.010	1.812
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.173	2.131	2.071	1.984	1.783
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.150	2.109	2.048	1.961	1.757
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.130	2.088	2.027	1.939	1.733
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.111	2.069	2.007	1.919	1.711
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	2.094	2.052	1.990	1.901	1.691
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	2.078	2.036	1.974	1.884	1.672
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	2.064	2.021	1.959	1.869	1.654
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	2.050	2.007	1.945	1.854	1.638
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.037	1.995	1.932	1.841	1.622
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142	2.070	2.015	1.972	1.908	1.817	1.594
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123	2.050	1.995	1.952	1.888	1.795	1.569
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106	2.033	1.977	1.934	1.870	1.776	1.547
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091	2.017	1.962	1.918	1.853	1.760	1.527
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.948	1.904	1.839	1.744	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.952	1.895	1.850	1.784	1.687	1.438
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.860	1.815	1.748	1.649	1.389
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.775	1.728	1.659	1.554	1.254
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831	1.752	1.692	1.644	1.571	1.459	1.000

TABLE 7.7 – Table des quantiles d'ordre 0.99 d'une variable de Fisher à n_1 et n_2 degrés de liberté

	n_1=1	2	3	4	5	6	7	8	9	10	12	14	16	20	30	∞
n_2=1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6143	6170	6209	6261	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.45	99.47	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.69	26.51	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.02	13.84	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.888	9.770	9.680	9.553	9.379	9.020
6	13.75	10.93	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.605	7.519	7.396	7.229	6.880
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.359	6.275	6.155	5.992	5.650
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.559	5.477	5.359	5.198	4.859
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	5.005	4.924	4.808	4.649	4.311
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.601	4.520	4.405	4.247	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.293	4.213	4.099	3.941	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.052	3.972	3.858	3.701	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.857	3.778	3.665	3.507	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.698	3.619	3.505	3.348	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.564	3.485	3.372	3.214	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.451	3.372	3.259	3.101	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.353	3.275	3.162	3.003	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.269	3.190	3.077	2.919	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.195	3.116	3.003	2.844	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.130	3.051	2.938	2.778	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.072	2.993	2.880	2.720	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	3.019	2.941	2.827	2.667	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.973	2.894	2.781	2.620	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.930	2.852	2.738	2.577	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.892	2.813	2.699	2.538	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.857	2.778	2.664	2.503	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.824	2.746	2.632	2.470	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.795	2.716	2.602	2.440	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.767	2.689	2.574	2.412	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.742	2.663	2.549	2.386	2.006
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934	2.798	2.696	2.618	2.503	2.340	1.956
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894	2.758	2.657	2.578	2.463	2.299	1.911
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859	2.723	2.622	2.543	2.428	2.263	1.872
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828	2.692	2.591	2.512	2.397	2.232	1.837
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.563	2.484	2.369	2.203	1.805
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.562	2.461	2.382	2.265	2.098	1.683
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.394	2.315	2.198	2.028	1.601
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.234	2.154	2.035	1.860	1.381
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.185	2.082	2.000	1.878	1.696	1.000