



Université Claude Bernard Lyon 1

**NOTES DE COURS DE STATISTIQUES**

**INTRODUCTION À LA STATISTIQUE DESCRIPTIVE**

**Formation : M1IGAPAS**

**UE : STATISTIQUES/GESTION**

**2012-2013, Automne**

**Jérôme BASTIEN**

Document compilé le 4 septembre 2012



## Identification Apogée

<b>Matière</b>	Statistiques
<b>Formation</b>	Master 1 I.F.A.P.SouA (P)
<b>Formation (code)</b>	SPM108
<b>UE</b>	1MIGAPAS2 Statistique
<b>UE (code)</b>	SPT1031M



# Table des matières

Identification Apogée	i
Avant-propos	v
Chapitre 0. Questionnaire	1
Chapitre 1. Une toute petite introduction à la statistique descriptive (sans $\mathbb{R}$ )	3
1.1. Introduction	3
1.2. Les données, les variables et le principe de la statistique descriptive	3
1.3. Étude de donnée qualitatives	4
1.4. Étude de données quantitatives	4
1.5. Éléments de correction	10
Chapitre 2. Une introduction à $\mathbb{R}$	13
2.1. Introduction	13
2.2. Démarrage de $\mathbb{R}$	14
2.3. Premières instructions avec $\mathbb{R}$	14
2.4. Premiers calculs avec $\mathbb{R}$	17
2.5. Quelques commandes arrêter $\mathbb{R}$ et pour se repérer et se déplacer dans $\mathbb{R}$	22
2.6. Dangers des mauvaises affectations!!	23
Chapitre 3. Une introduction à la statistique univariée. Variables et descriptions générales avec $\mathbb{R}$	25
3.1. Introduction	25
3.2. Étude de données	25
3.3. Étude de variables qualitatives	26
3.4. Étude de variables quantitatives	28
3.5. Quelques exercices facultatifs	29
3.6. Éléments de correction	30
3.7. Ensemble des figures	32
Chapitre 4. Croisement de deux variables quantitatives	41
4.1. Introduction	41
4.2. Principe théorique	41
4.3. La significativité pratique de la liaison	42
4.4. La significativité statistique de la liaison	44
4.5. Avec $\mathbb{R}$	45
4.6. Sur les dangers de la régression linéaire abusive : exemple d'Anscombe	53
4.7. Éléments de correction	53
Chapitre 5. Croisement de deux variables qualitatives	57
5.1. Introduction	57
5.2. Principe théorique	57

5.3. La significativité pratique de la liaison	60
5.4. La significativité statistique de la liaison	60
5.5. Avec $\mathcal{R}$	60
5.6. Éléments de correction	65
Chapitre 6. Croisement d'une variable qualitative et d'une variable quantitative	67
6.1. Introduction	67
6.2. Avec R	67
6.3. Calculer tous les indicateurs	74
6.4. Quelques exercices	74
6.5. Éléments de correction	76
Chapitre 7. Créer ses propres fichiers de données	81
Chapitre 8. Une introduction à la statistique univariée. Les représentations graphiques avec $\mathcal{R}$	83
8.1. Introduction	83
8.2. Étude de variables qualitative	83
8.3. Étude de variables quantitatives	84
8.4. Pour aller plus loin	85
8.5. Éléments de correction	85
8.6. Ensemble des figures	89
Chapitre 9. Récapitulatif des notions et commandes essentielles (statistiques descriptives)	95
9.1. Analyse univariée	95
9.2. Analyse bivariée	96
Chapitre 10. Exercices de révision (statistiques descriptives)	99
10.1. Énoncés	99
10.2. Corrigés	99
Annexe A. Installation du logiciel $\mathcal{R}$	109
A.1. Installation de $\mathcal{R}$ pour Windows	109
A.2. Utilisation de $\mathcal{R}$	109
Annexe B. Prise en main à la première séance	111
B.1. Création d'un dossier de travail (ou répertoire courant)	111
B.2. Téléchargement du cours et des fichiers de données	111
B.3. Installation du logiciel $\mathcal{R}$	111
Annexe C. Exemple de d'analyse univariée : quelques statistiques sur les lettres d'un texte	113
Annexe D. Utilisation de fonctions avec $\mathcal{R}$	115
D.1. Une fonction "simple"	115
D.2. Une fonction à deux valeurs de sortie	116
D.3. D'autres fonctions	118
Annexe E. Un exemple "pédagogique" sur les danger de la régression linéaire (sous forme d'exercice corrigé)	119
Énoncé	119
Corrigé	119
Bibliographie	127

## Avant-propos

Ces notes de cours constituent un support de cours, TD et TP de Statistiques pour l'UE Statistiques/Gestion du MIIGAPAS (2012-2013, Automne). Chacun des chapitres de ce poly s'inspire de photocopiés déjà existant réalisés par des statisticiens de l'université Lyon I, qui seront cités en début de chapitre (avec les URL où leurs photocopiés sont disponibles). Il s'agira essentiellement

- Anne-Béatrice DUFOUR (en collaboration avec de nombreux auteurs). Voir [1, 2, 3] toutes disponibles sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, puis rubrique Fiches de TD, puis statistique descriptive.
- Stéphane CHAMPELY, auteur de [4] ainsi que de [5, 6] ce dernier étant disponibles sous SPIRAL.

Ce photocopié de cours et les fichiers de données sont normalement disponibles à la fois

- en ligne sur <http://utbmjb.chez-alice.fr/UFRSTAPS/index.html> à la rubrique habituelle ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  - 'Poste de travail',
  - puis sur le répertoire 'P:' (appelé aussi '\\teraetu\Enseignants'),
  - puis 'jerome.bastien',
  - enfin sur 'MIIGAPAS'.

Pour l'examen, les données se trouveront aussi, par mesure de précaution à ces deux endroits.

Vous trouverez

- en annexe A, un petit guide d'installation du logiciel  et du package Rcmdr, pour ceux qui souhaitent l'installer sur leur propre ordinateur ;
- en annexe B, une prise en main à la première séance, pour ceux ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logiciels ;
- au chapitre 9, l'essentiel (et l'exigible aux examens!) des notions et commandes avec .

Une référence à consulter : [7] (voir bibliographie en page 127).

Pour ceux qui utilisent le package Rcmdr sur les ordinateurs<sup>1</sup> de l'université Lyon I, à cause d'un problème avec le package Rcmdr de la version 2.9 de R, on prendra bien garde à utiliser la version 2.7 de  et non la version 2.9 ; on trouvera cette version, comme d'habitude, en faisant "démarrer", puis "programmes", puis "R" puis "R 2.7". Si cette version n'est pas installée sur votre ordinateur, il faut le redémarrer!

Des notes en petits caractères comme suit pourront être omises en première lecture :

Attention, passage difficile!  $\diamond$

---

1. en date du mois de décembre 2009 ; ce bug a peut-être été corrigé depuis!



## CHAPITRE 0

# Questionnaire

Répondez aux questions suivantes (avec les codes proposés en italique) entre parenthèses.

*Merci de respecter les codes proposés en italiques pour le traitement correct de vos données ! Vous m'enverrez toutes ces informations dans un mail comme indiqué après le questionnaire (page 2) ! En cas de non réponse (pour quelle raison que ce soit), entrez NA.*

- (1) Pratiquez-vous une activité sportive? (*o* ou *n*).
- (2) Indiquez le nom de l'activité (une seule, celle que vous préférez, indiquez *aucune* sinon. Nom complet en minuscule. Si celle-ci est en plusieurs mots, les séparer par des underscores; exemple : *tir\_sportif*).
- (3) Indiquez votre sexe (*M* ou *F*).
- (4) Indiquez votre taille en cm (arrondissez à l'unité près).
- (5) Indiquez votre masse en kg (arrondissez à l'unité près).
- (6) Indiquez votre rythme cardiaque au repos, devant un ordinateur ou une feuille, en pulsations par minute.
- (7) Indiquez votre âge.
- (8) Indiquez le baccalauréat que vous avez obtenu (*S, L, ES, A, B, C, D, E*, ou *autre*).
- (9) Indiquez la main avec laquelle vous écrivez (*d* ou *g*).
- (10) Indiquez la main avec laquelle vous tenez votre fourchette (*d* ou *g*).
- (11) Indiquez l'œil avec lequel vous visez (*d* ou *g*).
- (12) Vous et les statistiques
  - (a) Avez-vous déjà suivi un cours de statistique? (*o* ou *n*)
  - (b) Pensez-vous pour votre futur métier, les statistiques sont :
    - pas du tout utiles (entrez 0)
    - pas utiles (entrez 1)
    - utiles (entrez 2)
    - très utiles (entrez 3)
- (13) Vous et le hasard
  - (a) Jouez à pile ou face et entrez le résultat (*p* ou *f*)
  - (b) Entrez un nombre au hasard entre 0 et 59; vous pourrez regarder une montre et noter à l'instant que vous voulez le nombre de secondes.

Pour faciliter la saisie, vous enverrai un mail

- à `jerome.bastien@univ-lyon1.fr`,
- avec comme objet : questionnaireMIIGAPAS2012-2013
- et dans le corps du mail, *uniquement et sans autre forme de texte*, l'ensemble des réponses en une ligne, séparées par des espace.

Un exemple :

```
from : charles.leclerc@univ-lyon14.fr  
to : jerome.bastien@univ-lyon1.fr  
subject : questionnaireMIIGAPAS2012-2013
```

---

```
o tir_sportif M 178 78 55 34 S d d g o 1 p 45
```

# Une toute petite introduction à la statistique descriptive (sans $\mathbb{R}$ )

## 1.1. Introduction

Ce chapitre a pour objectifs de donner les notions de bases relatives à différents types de données. Il est conseillé de le lire sans utiliser d'ordinateurs (une petite calculatrice suffira).

## 1.2. Les données, les variables et le principe de la statistique descriptive

Taille	Poids	Sexe	Sport pratiqué
183	80	H	Basket-ball
182	75	H	Escalade
173	66	F	Basket-ball
178	78	H	Gymnastique
192	77	H	Basket-ball
158	57	F	Natation
163	50	F	Judo
172	53	F	Tennis

TABLE 1.1. Tailles, poids, sexes et sports pratiqués pour  $N = 8$  individus.

Nous allons dans toute cette séance nous intéresser aux données que l'on pourra trouver dans le tableau 1.1 ; elles ont été collectées à partir d'un échantillon de 8 personnes (réalisé pour l'année 2008 dans un groupe de M1APA).

Ces données sont des informations, de deux types : numériques (on parle aussi de données quantitatives) ou catégorielles (on parle aussi de données qualitatives). Elles ont un sens dans un contexte précis.

On pourra consulter l'article de Wikipédia intitulé Statistique descriptive (voir [http://fr.wikipedia.org/wiki/Statistique\\_descriptive](http://fr.wikipedia.org/wiki/Statistique_descriptive)).

On cherche à décrire, c'est-à-dire résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses<sup>1</sup>. Il est important de résumer les observations sans détruire l'informations qu'elles contiennent.

Ces données varient (dans le temps, chez les individus) et prennent des valeurs différentes. Cette variabilité est si importante que l'on va donner aux mesures le nom de *variables*. Ainsi, on évoquera pour la population des  $N = 8$  individus déjà évoqués, les variables taille, poids, sexe et sport pratiqué.

Les valeurs de la variable poids sont successivement 80, 75, 66, 78, 77, 57, 50, 53.

Les valeurs de la variable sexe sont successivement  $H, H, F, H, H, F, F, F$ .

---

1. ce qui n'est guère pertinent dans notre cas ici!

Nous commencerons par le cas simple où il n'y a qu'une seule variable. On parle de phénomène mono-varié. À la fin du semestre, nous étudierons des phénomènes multivariés (en fait, seul le cas de deux variables sera étudié).

On parle de variable quantitative (ou numérique) ou variable qualitative (ou catégorielle).

### 1.3. Étude de donnée qualitatives

On s'intéressera au sexe des  $N = 8$  étudiants de M1APA (voir le tableau 1.1).

#### 1.3.1. Statistiques

On détermine tout d'abord le nombre de catégorie (ou de modalités), puis pour chacune d'elles, le nombre d'effectifs (c'est-à-dire le nombre d'individu pour lesquels la variable associée est dans cette catégorie).

On divisant ces effectifs par le nombre total d'individu, on obtient les fréquences. En multipliant ces fréquences par 100, on obtient les pourcentages.

EXERCICE 1.1. Déterminer les valeurs de ces statistiques pour l'échantillon des huit étudiants étudié.

Voir éléments de correction page 10.

#### 1.3.2. Graphiques

On peut produire des graphiques du type graphe en barres : on trace autant de barres que de catégories, chacune d'elle étant de même largeur, et de hauteur proportionnelle à la fréquence.

On peut aussi tracer un camembert, où chaque catégorie est représentée par un secteur angulaire proportionnel à la fréquence ; l'ensemble des secteurs angulaire est le disque total.

EXERCICE 1.2.

- (1) Déterminer le graphe en barres et le camembert pour les sexes des huit étudiants.
- (2) Ces graphes sont-ils pertinents ?

Voir éléments de correction page 10.

EXERCICE 1.3. On s'intéresse maintenant au sport pratiqué par les huit étudiants déjà étudiés. Il faudra prendre garde au fait qu'ici, il existe des cas de non réponse possibles (si aucun sport n'est pratiqués).

- (1) Reprendre l'analyse précédente de cette variable.
- (2) Représenter graphiquement ces données
- (3) Quel ordre choisir pour les catégories ? Peut-on regrouper les catégories ou utiliser une catégorie "Autre" ?

### 1.4. Étude de données quantitatives

On s'intéressera au poids des  $N = 8$  étudiants de M1APA (voir le tableau 1.1 page précédente).

#### 1.4.1. Statistiques

Ces données constitue un ensemble de nombres relatifs à une population de  $N = 8$  individus. On les notera  $n_1, n_2, \dots, n_8$ . De façon générale, ils seront notés  $(n_i)_{1 \leq i \leq N}$ .

On a donc successivement

- $n_1 = 80$ ,
- $n_2 = 75$ ,
- $n_3 = 66$ ,
- $n_4 = 78$ ,

- $n_5 = 77$ ,
- $n_6 = 57$ ,
- $n_7 = 50$ ,
- $n_8 = 53$

#### 1.4.1.1. La centralité.

On cherche tout d'abord à définir la centralité, c'est-à-dire, la valeur autour de laquelle s'organisent les différentes données.

La notion la plus connue est *la moyenne*<sup>2</sup>.

Si l'on dispose de deux nombre, la moyenne est tout simplement le milieu, c'est-à-dire la demi-somme. De façon plus générale, la moyenne est le nombre, souvent noté  $m$ , qui se trouve à égale distance de tous les nombres  $(n_i)_{1 \leq i \leq N}$ , soit encore le nombre  $m$  tel que

$$(m - n_1) + (m - n_2) + \dots + (m - n_N) = 0$$

Cela revient à donner la définition suivante :

DÉFINITION 1.4. La moyenne  $m$  des  $N$  nombres  $(n_i)_{1 \leq i \leq N}$  est définie par

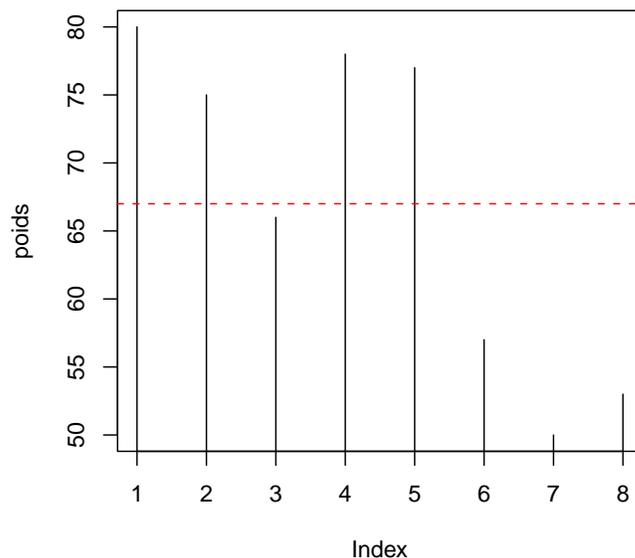
$$m = \frac{1}{N} (n_1 + n_2 + \dots + n_N) = \frac{1}{N} \sum_{i=1}^N n_i \quad (1.1)$$

EXEMPLE 1.5. La moyenne  $m$  des  $N = 8$  poids du tableau 1.1 page 3 se calcule de la façon suivante : Le détail du calcul est le suivant :

$$\begin{aligned} m &= \frac{1}{N} \sum_{i=1}^N n_i, \\ &= \frac{1}{8} (80 + 75 + 66 + 78 + 77 + 57 + 50 + 53), \\ &= \frac{536}{8}, \\ &= 67 \end{aligned}$$

---

2. l'adjectif "moyen" provient du latin *medianus*, qui signifie "du milieu"; cet adjectif a été substantivé au féminin dans "moyenne" qui a perdu son sens initial, pour exprimer ce qui est également distant des deux extrêmes et correspond au type le plus répandu [8]



Si on trace le *graphe indexé* ci-dessus avec en pointillé la moyenne, on constate que la somme des distances (algébriques) de chacun des poids à la moyenne est nulle. Un autre image peut-être donnée : on place sur une règle graduée (infiniment légère) différentes masses égales à des abscisses correspondant aux différentes données. Cette règle, posée horizontale sur une pointe, sera en équilibre si cette pointe correspond à la moyenne.

Une notions moins connue est la *la médiane*<sup>3</sup>. La médiane est un nombre qui divise en deux parties la population. On la note  $Q_2$ .

DÉFINITION 1.6. La médiane  $Q_2$  des  $N$  nombres  $(n_i)_{1 \leq i \leq N}$  est une valeur choisie pour que la moitié des données lui soit inférieure et l'autre moitié supérieure. On la notera  $Q_2$ .

De façon plus précise, pour définir la médiane, on classe les données dans l'ordre croissant. S'il y a un nombre pair de valeurs, la moyenne des deux valeurs centrales est prise. S'il y a un nombre impair de valeurs, la valeur centrale est choisie. Contrairement à la moyenne, la valeur médiane permet d'atténuer l'influence perturbatrice des valeurs extrêmes enregistrées lors de circonstances exceptionnelles. On dit que la médiane est moins sensible aux extrêmes que la moyenne.

EXEMPLE 1.7. La médiane  $Q_2$  des  $N = 8$  poids du tableau 1.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

- les données dans l'ordre croissant sont : 50, 53, 57, 66, 75, 77, 78, 80 ;
- le nombre de données est pair.
- on calcule donc la moyenne des deux valeurs centrales : 66 et de 75.
- la médiane vaut donc 70.5.

EXEMPLE 1.8. Si on avait voulu calculer la médiane des  $N - 1 = 7$  premiers poids du tableau 1.1, on aurait procédé ainsi. Le détail du calcul est le suivant :

- les données dans l'ordre croissant sont : 50, 57, 66, 75, 77, 78, 80 ;
- le nombre de données est impair.
- on prend la valeur centrale 75.
- la médiane vaut donc 75.

3. qui provient du latin *medianus*, qui signifie "du milieu" [8]

### 1.4.1.2. La dispersion ou l'hétérogénéité.

On cherche maintenant à définir si les données sont rassemblées ou non autour de la moyenne ou de la médiane.

Les extréma (minimum et maximum), notés  $\min(n_{i_1 \leq i \leq N})$  et  $\max(n_{i_1 \leq i \leq N})$ .

Les deux notions les plus importantes pour mesurer la dispersion des données autour de la moyenne sont la *variance* et l'*écart-type*.

On s'intéresse à la somme des écarts entre les données et la moyenne. Par définition la somme

$$(m - n_1) + (m - n_2) + \dots + (m - n_N) = 0$$

est nulle. On considérera une autre somme, où chaque quantité est toujours positive, en prenant par exemple le carré de chacun de ces termes :

$$(m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2$$

On divise cela par  $N$  pour donner autant d'importance à chaque terme. On obtient donc la variance

$$\frac{1}{N}((m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2)$$

Pour obtenir une quantité homogène à chacune des données, on prend la racine carrée de la variance. On obtient donc l'écart-type :

$$\sqrt{\frac{1}{N}((m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2)}$$

DÉFINITION 1.9. La *variance*, notée  $\sigma^2$ , des  $N$  nombres  $(n_i)_{1 \leq i \leq N}$  est définie par

$$\sigma^2 = \frac{1}{N}((m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2) = \frac{1}{N} \sum_{i=1}^N (m - n_i)^2 \quad (1.2)$$

DÉFINITION 1.10. L'*écart-type*, noté  $\sigma$ , des  $N$  nombres  $(n_i)_{1 \leq i \leq N}$  est défini par

$$\sigma = \sqrt{\frac{1}{N}((m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2)} = \sqrt{\frac{1}{N} \sum_{i=1}^N (m - n_i)^2} \quad (1.3)$$

REMARQUE 1.11. Attention, on parle quelque fois de l'écart-type estimé :

$$\sqrt{\frac{1}{N-1}((m - n_1)^2 + (m - n_2)^2 + \dots + (m - n_N)^2)}$$

On note aussi l'écart-type par son nom anglais, *sd*, comme standart deviation. Attention,  $\mathbb{R}$ détermine la déviation standart et non l'écart-type!

EXEMPLE 1.12. La variance  $\sigma^2$  des  $N = 8$  poids du tableau 1.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (m - n_i)^2, \\ &= \frac{1}{8} ((-13)^2 + (-8)^2 + 1^2 + (-11)^2 + (-10)^2 + 10^2 + 17^2 + 14^2), \\ &= \frac{1}{8} (169 + 64 + 1 + 121 + 100 + 100 + 289 + 196), \\ &= \frac{1040}{8}, \\ &= 130 \end{aligned}$$

EXEMPLE 1.13. L'écart-type  $\sigma$  des  $N = 8$  poids du tableau 1.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

$$\begin{aligned}\sigma &= \sqrt{\frac{1}{N} \sum_{i=1}^N (m - n_i)^2}, \\ &= \sqrt{\frac{1}{8} ((-13)^2 + (-8)^2 + 1^2 + (-11)^2 + (-10)^2 + 10^2 + 17^2 + 14^2)}, \\ &= \sqrt{\frac{1}{8} (169 + 64 + 1 + 121 + 100 + 100 + 289 + 196)}, \\ &= \sqrt{\frac{1040}{8}}, \\ &= \sqrt{130}, \\ &= 11.401754\end{aligned}$$

La notion de *quartile* permet de mesurer la dissymétrie et d'appréhender de façon différente la question de la dispersion.

DÉFINITION 1.14. De la même façon que la médiane partageait le jeu de données en deux groupes de même effectif, les quartiles vont le partager en quatre groupes d'effectifs égaux. Ainsi 25% des données seront inférieures au premier quartile ( $Q_1$ ), 50% au deuxième quartile qui n'est autre que la médiane ( $Q_2$ ) et 75% au troisième quartile ( $Q_3$ ).

Autrement dit,

- 25% des données sont inférieures à  $Q_1$ ,
- 25% des données sont comprises entre  $Q_1$  et  $Q_2$ ,
- 25% des données sont comprises entre  $Q_2$  et  $Q_3$ ,
- 25% des données sont supérieures à  $Q_3$ .

Parfois le minimum est noté  $Q_0$  et le maximum noté  $Q_4$ .

Les autres quartiles  $Q_1$  et  $Q_3$  sont donc définis comme la médiane de l'ensemble des valeurs inférieures à la médiane et la médiane de l'ensemble des valeurs supérieures à la médiane.

Cette définition n'est pas tout à fait celle utilisée par  $\mathbb{R}$ , le logiciel que vous utiliserez par la suite!  $\diamond$

EXEMPLE 1.15. Pour calculer, les trois quartiles des  $N = 8$  poids du tableau 1.1 :

- On calcule la médiane comme dans l'exemple 1.7 page 6 :  $Q_2 = 70.5$ .
- On calcule ensuite  $Q_1$ , comme la médiane des valeurs inférieures ou égales à  $Q_2 = 70.5$  : Le détail du calcul est le suivant :
  - les données dans l'ordre croissant sont : 50, 53, 57, 66 ;
  - le nombre de données est pair.
  - on calcule donc la moyenne des deux valeurs centrales : 53 et de 57.
  - la médiane vaut donc 55.
- On calcule ensuite  $Q_3$ , comme la médiane des valeurs supérieures ou égales à  $Q_2 = 70.5$  : Le détail du calcul est le suivant :
  - les données dans l'ordre croissant sont : 75, 77, 78, 80 ;
  - le nombre de données est pair.
  - on calcule donc la moyenne des deux valeurs centrales : 77 et de 78.
  - la médiane vaut donc 77.5.

☞ fournirait les valeurs suivantes pour les quantiles :

$$Q_1 = 56,$$

$$Q_2 = 70.5,$$

$$Q_3 = 77.25.$$

◇

### 1.4.1.3. Une remarque sur la moyenne et l'écart type.

Souvent, les étudiants retiennent des statistiques descriptives (quand ils retiennent quelque chose) les notions de moyenne ( $m$ ) et d'écart-type ( $\sigma$ ). C'est très souvent, en effet, les statistiques toujours présentées. On les présente souvent sous la forme de l'intervalle :

$$m \pm \sigma,$$

intervalle qui contient "la plupart" des données (en un sens à préciser ...).

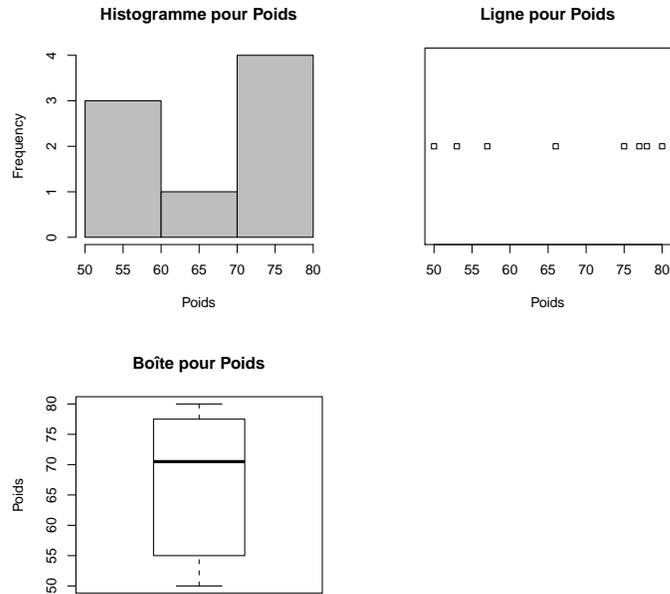
Si ces deux nombres ont autant d'importance, c'est parce que, dans un grand nombre de cas, les données étudiées suivent une loi idéale, dite en cloche ou "normale". Cette loi à la forme d'une cloche et est décrite par deux paramètres, notés moyenne et écart-type. Ces deux nombres sont proches de la moyenne et de l'écart-type des données considérées. Autrement dit, si les données suivent bien la loi normale, les deux nombres en question reflètent totalement les données considérés et les caractérisent donc.

## 1.4.2. Les graphiques

Pour représenter graphiquement des données quantitatives, on peut représenter les valeurs individuelles le long d'une échelle (ligne de points, avec empilement des points égaux). On peut aussi les regrouper par tranche et tracer un histogramme : pour chaque tranche choisie, on détermine le nombre d'individus pour lesquels la variable qualitative appartient à cette tranche. On trace ensuite des colonnes dont la base correspond à la tranche et la hauteur est proportionnelle au nombre d'individu.

On peut aussi tracer des boîtes de dispersion ou boîte à moustache mettant en évidence les extrêmes et les quartiles : pour simplifier, la boîte de dispersion comporte une boîte centrale avec des traits noirs d'ordonnées  $Q_1$ ,  $Q_2$  et  $Q_3$ , en étirant les "moustaches" jusqu'aux valeurs minimale et maximale.

On appelle  $EIQ$  l'écart interquartile  $EIQ = Q_3 - Q_1$  ; points extrêmes sont les points correspondant aux valeurs inférieures à  $Q_1 - 1.5EIQ$  ou supérieurs à  $Q_3 + 1.5EIQ$ . S'il n'y a pas de points extrêmes en dessous (resp. en dessus), on procède comme ci-dessus. Sinon, on tire la moustache jusqu'à la limite  $Q_1 - 1.5EIQ$  (resp.  $Q_3 + 1.5EIQ$ ) et on marque l'emplacement des points extrêmes par des petits cercle. ◇



Voir les trois graphiques ci-dessus pour la variable poids. Le nombre de données étant faible (8), l'histogramme et la boîtes à moustache ne sont pas très pertinents ici.

Ces trois graphiques ne sont pas toujours pertinents.

EXERCICE 1.16. Déterminer les statistiques et faire les graphiques à main levée de la variable taille. Pour l'histogramme, on prendra des classes de largeur 10 à partir de 150.

Voir éléments de correction page 11.

## 1.5. Éléments de correction

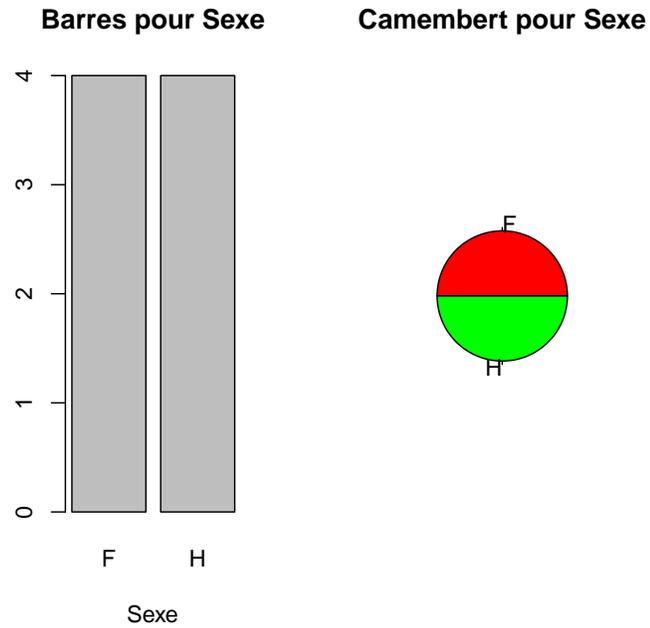
ÉLÉMENTS DE CORRECTION DE L'EXERCICE 1.1

Les effectifs et les pourcentages déterminés par  $\mathbb{R}$  sont donnés dans le tableau suivant

	effectifs	pourcentages
F	4	50.000
H	4	50.000

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 1.2

(1)



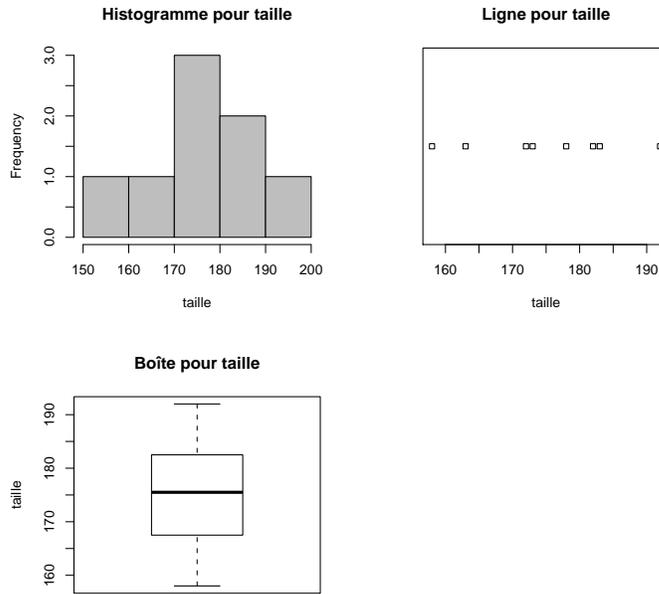
Voir les deux graphiques ci-dessus pour la variable sexe.

(2) À vous de voir ....

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 1.16

Les différents résultats déterminés par  $\mathcal{R}$  sont donnés dans le tableau suivant

noms	valeurs
moyenne	175.125
écart-type	11.063938
$Q_1$ (quartile à 25 %)	169.75
médiane	175.5
$Q_3$ (quartile à 75 %)	182.25
minimum	158
maximum	192
nombre	8



Voir les trois graphiques ci-dessus.

## Une introduction à $\mathbb{R}$

Ce chapitre s'inspire des documents [9, 10, 11], disponibles sur <http://pbil.univ-lyon1.fr/R/enseignement.html>.

Tous les fichiers des données sont des fichiers au format texte, tous disponible sur <http://utbmjb.chez-alice.fr/UFRSTAPS>. Ils proviennent essentiellement de <http://pbil.univ-lyon1.fr/R/enseignement.html>, puis rubrique **Dossiers de fichiers**.

Ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logiciels et pourront lire l'annexe B en première séance.

### 2.1. Introduction

Nous donnerons pour chacun des chapitres de ce cours, un bref rappel théorique suivi d'exemple et d'exercices à traiter avec le logiciel  $\mathbb{R}$ .

Les instructions figureront en toutes lettre sur le document distribué sous la forme suivante : on tapera les instructions suivantes dans la fenêtre de "RGui"

```
2 + 5
```

ou

```
t <- 2 + 5
t
t + 3
```

On verra alors apparaître le résultat dans la même fenêtre :

```
[1] 7
```

ou

```
[1] 7
```

```
[1] 10
```

#### REMARQUE 2.1.

- (1) Si vous disposez du document pdf de ce cours, vous pouvez normalement copier-coller une ou plusieurs lignes du document pdf vers la fenêtre de commande "Rgui", ce qui pourra vous éviter des erreurs de frappe!
- (2) Dans "Rgui", vous pouvez réutiliser les commandes déjà saisies et les modifier en les rappelant grâce aux flèches "haut" et "bas" du clavier ( $\uparrow$  et  $\downarrow$ ) et les modifier en vous aidant des flèches "gauche" et "droite" du clavier ( $\leftarrow$  et  $\rightarrow$ )
- (3) Dans ce polycopié,
  - Conformément à la présentation de  $\mathbb{R}$ , les "entrées" (les instructions à taper dans la fenêtre de commande "Rgui", derrière le "prompt"  $>$ ) sont présentées en rouge et en police "machine à écrire", comme par exemple ce qui suit :

```
cos(2)
```

- De même, conformément à la présentation de  $\mathbb{R}$ , les "sorties" (ce qui sera calculé par  $\mathbb{R}$ ) sont présentées en bleu et en police "machine à écrire", comme par exemple ce qui suit :  
[1] -0.4161468
- Enfin, si l'entrée et le résultat sont présentés simultanément, vous verrez  
`cos(2)`  
[1] -0.4161468

## 2.2. Démarrage de $\mathbb{R}$

Après avoir lancé le logiciel, il est important de spécifier un répertoire courant en cliquant sur

- fichier
- puis sur **changer le répertoire courant**
- et reprendre une ou deux fois OK (selon les versions de R utilisées).

## 2.3. Premières instructions avec $\mathbb{R}$

### 2.3.1. Importer des données

- (1) Récupérer le fichier de données `t3var.txt` et l'ouvrir avec un éditeur de bas niveau (par exemple wordpad).

Constater qu'il contient des données qualitatives et quantitatives pour un certain nombre d'individu.

- (2) Retourner dans  $\mathbb{R}$  pour le lire à nouveau avec son éditeur.

Observer d'abord ce qui se passe quand on tape

```
read.table("t3var.txt")
```

Observer ensuite ce qui se passe quand on lit les 5 premières lignes du tableau :

```
head(read.table("t3var.txt"))
```

On voit alors apparaître (réponse écrite en bleu)

```
V1 V2 V3
1 sexe poi tai
2 h 60 170
3 f 57 169
4 f 51 172
5 f 55 174
```

Le premier individu contient les entêtes !

Rajouter une valeur 'TRUE' du paramètre optionnel 'header' puis placer le résultat de cette lecture dans un objet. On dit qu'on fait une affectation ;

```
t3var <- read.table("t3var.txt", h = T)
head(t3var)
```

ou de façon équivalente

```
t3var <- read.table("t3var.txt", header = TRUE)
head(t3var)
```

On peut aussi écrire :

```
t3var <- read.table("t3var.txt", h = T)
head(t3var)
```

les deux formes d'affectation donnent le même résultat :

```
sexe poi tai
1 h 60 170
```

```
2   f  57 169
3   f  51 172
4   f  55 174
5   f  50 168
6   f  50 161
```

À vous de savoir laquelle vous préférez. Noter que la flèche en haut, au clavier, rappelle la dernière ligne frappée.

- (3) Dans le cas précédent, le fichier à charger était sauvegardé dans un répertoire local. On peut en fait le lire directement par :

```
t3var <- read.table("http://utbmjb.chez-alice.fr/UFRSTAPS/M1APA/fichiersR/t3var.txt",
  h = T)
```

Attention, cela suppose que la connexion internet avec le serveur ne soit pas défaillante!

Pour avoir l'ensemble des noms des variables, on tapera

```
names(t3var)
```

### 2.3.2. Consulter la documentation (l'aide)

```
?read.table
help("read.table")
```

Consulter la fiche de documentation est une opération fondamentale!

### 2.3.3. Conserver les résultats (Section facultative)

Ouvrir un fichier Word, définir un style listing avec la police Courier New pour enregistrer les listings.

```
summary(t3var)
```

```
sexe      poi          tai
f:25  Min.   :47.00  Min.   :150.0
h:41  1st Qu.:53.00  1st Qu.:168.0
      Median :65.50  Median :174.5
      Mean   :64.52  Mean   :174.1
      3rd Qu.:73.00  3rd Qu.:180.0
      Max.   :86.00  Max.   :200.0
```

Voir la figure 2.1 produite par

```
plot(t3var$poi, t3var$tai)
```

Copier les listings (sélection à la souris) et les figures (cliquer dessus avec le bouton droit) et coller dans un document Word. Vous pouvez faire un rapport.

On pourra aussi utiliser la fonction `Sweave` qui permet de rédiger des rapports en exportant des figures et des résultats de façon très élégante (Voir <http://pbil.univ-lyon1.fr/R/enseignement>, puis rubrique `fiche de TD`, puis `divers`, puis `tdr78`).

*Attention* pour utiliser cette méthode, il faut aussi utiliser le génialissime traitement de texte scientifique<sup>1</sup> `LATEX`.

### 2.3.4. Commencer un catalogue personnel (Section facultative)

Il n'y a plus qu'une difficulté : connaître et comprendre quelques unes des 2500 fonctions de base (sans compter 450 bibliothèques additionnelles). Au début, on fera la liste des fonctions passées en revue qu'il semble

---

1. Ce cours a été rédigé ainsi!

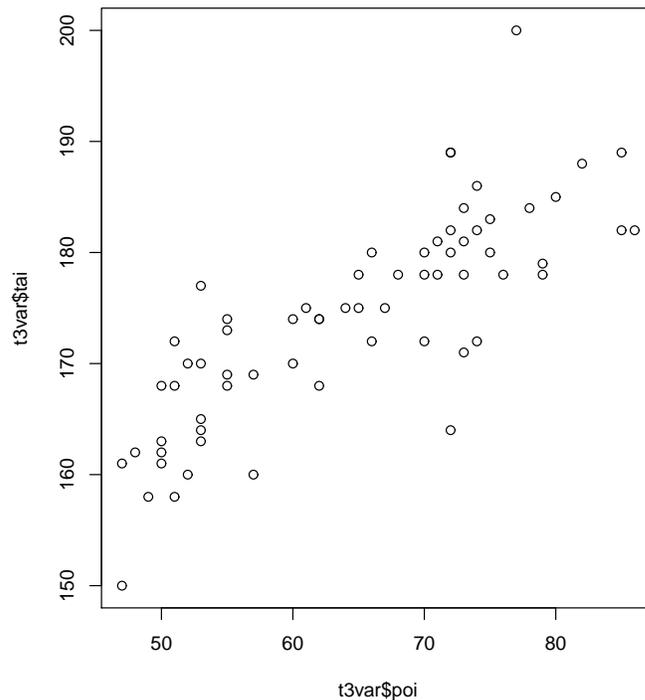


FIGURE 2.1. Un exemple de graphique

logique de mémoriser dans un petit catalogue personnel. Il convient de ne noter que des mots-clés, la documentation de chaque fonction étant toujours là pour les détails. Noter la fonction `apropos()` qui peut vous aider à retrouver un nom de fonctions. On retiendra `file` pour tout ce qui touche aux fichiers :

```
apropos("file")
```

```
[1] "bzfile"           "choose.files"     "close.srcfile"
[4] "download.file"   "env.profile"      "file"
[7] "file.access"     "file.append"      "file.choose"
[10] "file.copy"       "file.create"      "file.edit"
[13] "file.exists"    "file.info"        "file.path"
[16] "file.remove"    "file.rename"      "file.show"
[19] "file.symlink"   "file_test"        "gzfile"
[22] "list.files"     "memory.profile"   "namesdonnee.file"
[25] "nom.file"       "nom.file.blondes" "nom.file.cor"
[28] "open.srcfile"   "open.srcfilecopy" "parseNamespaceFile"
[31] "print.srcfile"  "profile"          "readCitationFile"
[34] "srcfile"        "srcfilecopy"      "system.file"
[37] "tempfile"       "win.metafile"     "xzfile"
[40] "zip.file.extract"
```

On retiendra `read` et `write` pour les entrées-sorties.

```
apropos("read")
```

```
[1] ".readRDS"          "read.csv"          "read.csv2"         "read.dcf"
[5] "read.delim"        "read.delim2"       "read.DIF"          "read.fortran"
[9] "read.ftable"       "read.fwf"          "read.socket"       "read.table"
[13] "read.table.url"    "readBin"           "readChar"          "readCitationFile"
[17] "readClipboard"     "readline"          "readLines"         "readRegistry"
[21] "spread.level.plot" "Sys.readlink"

apropos("write")

[1] "RtangleWritedoc"    "RweaveLatexWritedoc" "write"
[4] "write.csv"          "write.csv2"           "write.dcf"
[7] "write.ftable"       "write.socket"         "write.table"
[10] "write.table0"       "writeBin"             "writeChar"
[13] "writeClipboard"    "writeLines"
```

Vous entrez alors dans la caverne d'Ali Baba.

## 2.4. Premiers calculs avec $\mathbb{R}$

On consultera la fiche très complète [11]. On en donne ici quelques lignes fondamentales.

### 2.4.1. Calculs sur les vecteurs

Dans  $\mathbb{R}$ , même une simple valeur numérique est un vecteur :

```
pi
[1] 3.141593

is.vector(pi)
```

```
[1] TRUE
```

Prenons un échantillon constitué des 10 premiers entiers. Il s'écrit :

```
1:10
[1] 1 2 3 4 5 6 7 8 9 10
```

Et la moyenne de cet échantillon est obtenue par :

```
mean(1:10)
[1] 5.5
```

Les expressions sont ainsi beaucoup plus compactes et proches des notations mathématiques.

EXERCICE 2.2.

(1) Donner le carré des 10 premiers entiers :

```
[1] 1 4 9 16 25 36 49 64 81 100
```

(2) Donner le carré des 10 premiers entiers moins un :

```
[1] 0 3 8 15 24 35 48 63 80 99
```

(3) Donner le sinus des 10 premiers entiers :

```
[1] 0.8414710 0.9092974 0.1411200 -0.7568025 -0.9589243 -0.2794155 0.6569866
[8] 0.9893582 0.4121185 -0.5440211
```

(4) Donner 10 à la puissance des 10 premiers entiers :

```
[1] 1e+01 1e+02 1e+03 1e+04 1e+05 1e+06 1e+07 1e+08 1e+09 1e+10
```

(5) Donner la somme (`sum()`) des 10 premiers entiers :

```
[1] 55
```

### 2.4.2. Statistiques élémentaires

Considérons les notes de 14 étudiants d'un groupe de TP et calculons les paramètres statistiques élémentaires. L'écriture d'un vecteur est une combinaison (cf le paragraphe sur les séries de valeurs numériques ci-après).

```
notes <- c(15, 8, 14, 12, 14, 10, 18, 15, 9, 5, 12, 13, 12, 16)

sort(notes)

[1] 5 8 9 10 12 12 12 13 14 14 15 15 16 18

length(notes)

[1] 14

min(notes)

[1] 5

max(notes)

[1] 18

range(notes)

[1] 5 18

median(notes)

[1] 12.5

quantile(notes)

 0%  25%  50%  75% 100%
5.00 10.50 12.50 14.75 18.00

mean(notes)

[1] 12.35714

var(notes)

[1] 11.93956

sd(notes)

[1] 3.455367

unique(notes)

[1] 15 8 14 12 10 18 9 5 13 16

sort(unique(notes))

[1] 5 8 9 10 12 13 14 15 16 18
```

### 2.4.3. Séries de valeurs numériques

L'opérateur deux points : permet de générer des séries d'entiers :

```
1:12
```

```
[1] 1 2 3 4 5 6 7 8 9 10 11 12
```

```
-4:3
```

```
[1] -4 -3 -2 -1 0 1 2 3
```

La fonction `seq()` permet de générer une série de nombres équidistants :

```
seq(from = 1, to = 5)
```

```
[1] 1 2 3 4 5
```

```
seq(from = 1, to = 2, by = 0.1)
```

```
[1] 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
```

```
seq(from = 10, to = 50, length = 10)
```

```
[1] 10.00000 14.44444 18.88889 23.33333 27.77778 32.22222 36.66667 41.11111 45.55556
```

```
[10] 50.00000
```

### 2.4.4. Éléments des vecteurs

*2.4.4.1. Indexation par des entiers positifs.*

Soit

```
x <- c(2, 0.6, 0.8, -14, 10, 24)
```

Le quatrième élément :

```
x[4]
```

```
[1] -14
```

Du deuxième au troisième élément :

```
x[2:3]
```

```
[1] 0.6 0.8
```

On peut reprendre plusieurs fois le même :

```
x[c(2, 2, 3)]
```

```
[1] 0.6 0.6 0.8
```

Les éléments hors bornes ne sont pas disponibles (NA, not available, données manquante) :

```
x[100]
```

```
[1] NA
```

*2.4.4.2. Indexation par un vecteur logique.*

```
x[c(T, F, F, T, T, T)]
```

```
[1] 2 -14 10 24
```

Le vecteur logique d'indexation peut être issu d'un calcul logique, c'est l'utilisation la plus courante :

```
x > 10
```

```
[1] FALSE FALSE FALSE FALSE FALSE TRUE
```

```
x[x > 10]
```

```
[1] 24
```

```
x[x >= 10]
[1] 10 24
x > 0 & x < 1
[1] FALSE TRUE TRUE FALSE FALSE FALSE
x[x > 0 & x < 1]
[1] 0.6 0.8
x[x < 0 | x > 1]
[1] 2 -14 10 24
x == 10
[1] FALSE FALSE FALSE FALSE TRUE FALSE
x != 10
[1] TRUE TRUE TRUE TRUE FALSE TRUE
x[x != 10]
[1] 2.0 0.6 0.8 -14.0 24.0
x[x >= 10]
[1] 10 24
```

*Attention*, si on tape (par erreur)

```
x = 10
```

cela affecte 10 à la variable x. On n'utilisera pas le = de l'affectation, mais uniquement les deux flèches.

On donne aussi un exemple de la fonction `which` qui permet de retrouver les indices des booléens vrais :

```
x <- c(2, 0.6, 0.8, -14, 10, 24)
x >= 10
[1] FALSE FALSE FALSE FALSE TRUE TRUE
which(x >= 10)
[1] 5 6
```

## 2.4.5. Les tableaux : data frames

### 2.4.5.1. Les objets de type `data.frame`.

Les data frames sont des listes dont tous les éléments ont la même longueur. On peut mélanger dans un data.frame des variables quantitatives, qualitatives, logiques et textuelles. C'est typiquement le type d'objet que l'on récupère en lisant des données dans un fichier, par exemple :

```
t3var <- read.table("http://utbmjb.chez-alice.fr/UFRSTAPS/M1APA/fichiersR/t3var.txt",
  header = TRUE)
```

ou

```
t3var <- read.table("t3var.txt", header = TRUE)
is.data.frame(t3var)
[1] TRUE
names(t3var)
[1] "sexe" "poi" "tai"
```

```
summary(t3var)
```

```
sexe      poi      tai
f:25  Min.   :47.00  Min.   :150.0
h:41  1st Qu.:53.00  1st Qu.:168.0
      Median :65.50  Median :174.5
      Mean   :64.52  Mean   :174.1
      3rd Qu.:73.00  3rd Qu.:180.0
      Max.   :86.00  Max.   :200.0
```

2.4.5.2. *Éléments des data frames.*

\$ et

```
tab[i, j]
```

Les data frames étant des listes, les variables (en colonne) sont directement accessibles par leur nom :

```
t3var$tai
```

```
[1] 170 169 172 174 168 161 162 189 160 175 165 164 175 184 178 158 164 179 182 174
[21] 158 163 172 185 170 178 180 189 172 174 200 178 178 168 170 160 163 168 172 175
[41] 180 162 177 169 173 182 183 184 181 180 178 178 168 161 171 180 174 175 182 181
[61] 188 182 189 178 150 186
```

Pour s'affranchir de taper à chaque fois le nom de l'objet (ici `t3var`), on peut taper dès qu'on a chargé le fichier

```
attach(t3var)
```

```
The following object(s) are masked _by_ .GlobalEnv :
```

```
sexe
```

et appeler les différentes colonnes directement par leur nom

```
tai
```

```
[1] 170 169 172 174 168 161 162 189 160 175 165 164 175 184 178 158 164 179 182 174
[21] 158 163 172 185 170 178 180 189 172 174 200 178 178 168 170 160 163 168 172 175
[41] 180 162 177 169 173 182 183 184 181 180 178 178 168 161 171 180 174 175 182 181
[61] 188 182 189 178 150 186
```

```
sexe
```

```
[1] H H F H H F F F
```

```
Levels: F H
```

Attention, on ne peut attacher plusieurs objets s'ils ont des noms de colonnes en commun. Une fois le travail terminé, il faut absolument détacher l'objet :

```
detach(t3var)
```

On peut également utiliser la notation de type `tab[i,j]` où `i` représente l'index des lignes et `j` celui des colonnes. Par exemple, pour avoir les 5 premiers individus :

```
t3var[1:5, ]
```

```
sexe poi tai
1    h  60 170
2    f  57 169
3    f  51 172
```

```
4 f 55 174
5 f 50 168
```

Pour avoir les première et troisième colonnes des 5 premiers individus :

```
t3var[1:5, c(1, 3)]
```

```
sexe tai
1 h 170
2 f 169
3 f 172
4 f 174
5 f 168
```

EXERCICE 2.3. Afficher de deux façons différentes les poids de 't3var' strictement supérieurs à 50 kg et les tailles des individus de 't3var' dont les poids sont strictement supérieurs à 50 kg.

Voir éléments de correction page 22.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 2.3

Taper successivement

- `t3var$poi[t3var$poi > 50]`

```
[1] 60 57 51 55 72 52 64 53 72 61 78 68 51 53 79 74 62 74 80 53 73 70 72 70 62 77 70
[28] 76 51 52 57 53 55 66 65 75 53 55 55 72 75 73 71 66 71 79 62 73 72 60 67 85 73 82
[55] 86 85 65 74
```
- `t3var[t3var[, 2] > 50, 2]`

```
[1] 60 57 51 55 72 52 64 53 72 61 78 68 51 53 79 74 62 74 80 53 73 70 72 70 62 77 70
[28] 76 51 52 57 53 55 66 65 75 53 55 55 72 75 73 71 66 71 79 62 73 72 60 67 85 73 82
[55] 86 85 65 74
```
- `t3var$tai[t3var$poi > 50]`

```
[1] 170 169 172 174 189 160 175 165 164 175 184 178 158 164 179 182 174 172 185 170
[21] 178 180 189 172 174 200 178 178 168 170 160 163 168 172 175 180 177 169 173 182
[41] 183 184 181 180 178 178 168 171 180 174 175 182 181 188 182 189 178 186
```
- `t3var[t3var[, 2] > 50, 3]`

```
[1] 170 169 172 174 189 160 175 165 164 175 184 178 158 164 179 182 174 172 185 170
[21] 178 180 189 172 174 200 178 178 168 170 160 163 168 172 175 180 177 169 173 182
[41] 183 184 181 180 178 178 168 171 180 174 175 182 181 188 182 189 178 186
```

## 2.5. Quelques commandes arrêter R et pour se repérer et se déplacer dans R

Si un calcul ne s'arrête pas (pas de réponses) ou si vous n'avez pas parenthésé correctement une instruction (apparition de + à l'écran), stoppez le déroulement de R grâce à la touche STOP de la barre des tâches.

Tapez :

- `getwd()`  
vous verez votre répertoire courant ;
- `dir()`  
vous verez ce que contient votre répertoire courant ;
- `ls()`  
vous verez la liste des objets chargés (data fram, variables, fonctions ...);
- `setwd(dir)`  
vous irez directement dans le répertoire "dir" (sans passer par la rubrique "changer de répertoire").
- `rm(list = ls())`

vous effacez toutes les variables (et fonctions).

## 2.6. Dangers des mauvaises affectations !!

EXERCICE 2.4. *ATTENTION*, comme matlab ou d'autres logiciels du même type,  $\mathbb{R}$  présente l'inconvénient de pouvoir faire de dangereuses affectations si on utilise les mots-clés (c'est-à-dire, les noms de variables réservées, utilisées par  $\mathbb{R}$ ).

Faire les commandes suivantes et méditer aux dangers mis en évidence, par l'utilisation des mots-clés 'T' ou 'cos'.

```
T
[1] TRUE
TRUE
[1] TRUE
T & F
[1] FALSE
T <- c(1, 2, 3)
T & F
[1] FALSE FALSE FALSE
rm(T)
T
[1] TRUE
sd(c(1, 2))
[1] 0.7071068
sd <- cos
cos(c(1, 2))
[1] 0.5403023 -0.4161468
sd(c(1, 2))
[1] 0.5403023 -0.4161468
sd(1)
[1] 0.5403023
rm(sd)
sd(c(1, 2))
[1] 0.7071068
```



## Une introduction à la statistique univariée. Variables et descriptions générales avec $\mathbb{R}$

Ce chapitre s'inspire des documents [12, 13, 14] disponibles sur <http://pbil.univ-lyon1.fr/R/enseignement.html>.

Comme d'habitude, tous les fichiers des données sont des fichiers au format texte, tous disponible sur <http://utbmjb.chez-alice.fr/UFRSTAPS/index.html> Ils proviennent essentiellement de <http://pbil.univ-lyon1.fr/R/enseignement.html>, puis rubrique `Dossiers de fichiers`.

*Pour des raisons de lisibilité, toutes les figures de ce chapitre sont données à partir de la page 32.*

### 3.1. Introduction

L'objet de ce chapitre est de reprendre les calculs élémentaires de statistique déjà fait au cours du chapitre 1 en s'appuyant cette fois sur le logiciel  $\mathbb{R}$ . Les calculs seront effectués et quelques graphiques de bases seront tracés.

Le chapitre (ou l'annexe) 8 présentera les graphes de façon plus complète.

### 3.2. Étude de données

Une série statistique associée à une variable  $X$  est une liste de valeurs mesurées sur  $N$  individus. A chaque individu  $i$  est associée la valeur  $n_i$ .

Si la variable  $X$  est qualitative, les  $n_i$  sont des modalités. Si la variable  $X$  est quantitative, les  $n_i$  sont des nombres. En général, les données sont stockées dans des tableaux.

Comme au début du chapitre 2, charger le fichier `extraitM1APAP08data.txt`, qui correspond au tableau 1.1 page 3 :

```
extraitM1APAP08data <- read.table("extraitM1APAP08data.txt", h = T)
```

Taper successivement

```
summary(extraitM1APAP08data)
```

```
      Taille      Poids      Genre      Sport_pratiqué
Min.   :158.0  Min.   :50.00  F:4    Basket-ball:3
1st Qu.:169.8  1st Qu.:56.00  H:4    Escalade   :1
Median :175.5  Median :70.50           Gymnastique:1
Mean   :175.1  Mean   :67.00           Judo       :1
3rd Qu.:182.2  3rd Qu.:77.25           Natation   :1
Max.   :192.0  Max.   :80.00           Tennis    :1
```

```
names(extraitM1APAP08data)
```

```
[1] "Taille"      "Poids"      "Genre"      "Sport_pratiqué"
```

```
levels(extraitM1APAP08data$Sport_pratiqué)
```

```
[1] "Basket-ball" "Escalade"   "Gymnastique" "Judo"      "Natation"   "Tennis"
```

Une série statistique ordonnée est une série statistique où les valeurs ont été classées généralement de la plus petite à la plus grande valeur. Chaque valeur de la série est notée  $x_i$ .

Taper

```
sort(extraitM1APAP08data$Poids)
```

```
[1] 50 53 57 66 75 77 78 80
```

### 3.3. Étude de variables qualitatives

On reprend dans cette section et la suivante ce que vous avez fait "à la main" au cours de la section 1.4.1 page 4.

EXERCICE 3.1.

Dans cet exercice, on s'intéresse toujours au fichier `extraitM1APAP08data.txt`.

- (1) Faire un attachement :

```
attach(extraitM1APAP08data)
```

- (2) Afficher les différentes modalités de la variable "sport pratiqués" en tapant :

```
levels(Sport_pratiqué)
```

```
[1] "Basket-ball" "Escalade"      "Gymnastique" "Judo"          "Natation"      "Tennis"
```

- (3) Afficher les différentes fréquences de chaque sport en tapant

```
table(Sport_pratiqué)
```

```
Sport_pratiqué
```

Basket-ball	Escalade	Gymnastique	Judo	Natation	Tennis
3	1	1	1	1	1

- (4) Afficher les différentes pourcentage de chaque sport en tapant

```
u <- table(Sport_pratiqué)
```

```
100 * u/sum(u)
```

```
Sport_pratiqué
```

Basket-ball	Escalade	Gymnastique	Judo	Natation	Tennis
37.5	12.5	12.5	12.5	12.5	12.5

- (5) Que se passe-t-il si on tape

```
summary(Sport_pratiqué)
```

Basket-ball	Escalade	Gymnastique	Judo	Natation	Tennis
3	1	1	1	1	1

```
class(table(Sport_pratiqué))
```

```
[1] "table"
```

```
class(summary(Sport_pratiqué))
```

```
[1] "integer"
```

- (6) Tracer le graphe en barre et la camembert des sports pratiqués grâce aux fonctions `table`, `pie` et `barplot` en tapant

```
pie(table(Sport_pratiqué))
```

et

```
barplot(table(Sport_pratiqué))
```

Vous devriez obtenir des figures analogues aux figures 3.1 page 33.

*Pour des raisons de lisibilité, toutes les figures de ce chapitre sont données à partir de la page 32.*

- (7) Ne pas oublier de taper :

```
detach(extraitM1APAP08data)
```

### EXERCICE 3.2.

Le questionnaire fait lors du chapitre 0 a été traité et l'ensemble de vos réponses constituent le fichier 'M1IGAPASA12data.txt'.

*Seules 33 réponses ont été obtenues, soit 84.62 % des réponses ! Pour ceux qui n'ont pas fait ce questionnaire, merci de vous rendre au chapitre 0, de répondre aux questions posées et de m'envoyer les résultats par mails, en respectant le format demandé !*

- (1) Chargez le fichier M1IGAPASA12data.txt en tapant par exemple

```
M1IGAPASA12data<-read.table("M1IGAPASA12data.txt",h=T)
```

Faites un attachement en tapant

```
attach(M1IGAPASA12data)
```

- (2) Regarder l'aide la fonction `is.na` et étudier les éventuelles données manquantes.  
 (3) Retrouvez les données qui vous appartiennent dans ce tableau de données.

Vous pourrez :

- soit regarder l'ensemble du fichier et retrouver votre réponse ;
- soit retrouver votre réponse à partir de votre taille, poids et sports pratiqués ; par exemple pour le ou les individus dont
  - le sport est "equitation",
  - la taille est 164
  - le poids est 53

il faudrait taper par exemple

```
indice<-which((sport=="equitation")&(taille==164)&(masse==53))
```

```
M1IGAPASA12data[indice,]
```

Ici, vous devriez obtenir

```
pratique_sportive      sport sexe taille masse rythme_cardiaque age baccalaureat main_ecrit
1                   o equitation   F   164   53                   68 22                   S
main_fourchette oeil cours_stats interet_stats piece nombre
1                 d   d           n           1   f   11
```

- (4) Tracer le graphe en barre et la camembert des sexes et des sports pratiqués grâce aux fonctions `table`, `pie` et `barplot`.  
 (5) Quel est l'inconvénient des deux graphes qui viennent d'être tracés ? Voir ce qui se passe en tapant :

```
dotchart(table(sport))
dotchart(sort(table(sport)))
```

REMARQUE 3.3. Attention, cette commande semble ne plus fonctionner à partir de la version 2.12 de  $\mathcal{R}$  ; il faut taper à la place :

```
dotchart(summary(sport))
dotchart(sort(summary(sport)))
```

- (6) En vous rappelant qu'on a adopté le codage suivant :
- pas du tout utiles  $\rightarrow 0$  ;
  - pas utiles  $\rightarrow 1$  ;
  - utiles  $\rightarrow 2$  ;
  - très utiles  $\rightarrow 3$ .

étudier la variable "interet\_stats"

(voir question 12b du chapitre 0).

- (7) N'oubliez pas le détachement en tapant

```
detach(M1IGAPASA12data)
```

Voir les éléments de correction page 30.

EXERCICE 3.4 (facultatif). Reprendre la même étude que l'exercice 3.2 pour le fichier de données L3APA06.txt (correspondant à des étudiants de L3) qui contient un plus grand nombre de données.

### 3.4. Étude de variables quantitatives

EXERCICE 3.5.

Dans cet exercice, on s'intéresse toujours au fichier `extraitM1APAP08data.txt`.

- (1) Afficher les différentes valeurs "Taille" en tapant :

```
names(extraitM1APAP08data)
[1] "Taille"          "Poids"          "Genre"          "Sport_pratiqué"
attach(extraitM1APAP08data)
Taille
[1] 183 182 173 178 192 158 163 172
```

- (2) Calculer la moyenne, le minimum, le maximum, la médiane et les autres quartiles des tailles en servant des fonctions `mean`, `min`, `max`, `range`, `median` et `quantile`. On pourra aussi utiliser la fonction `summary` :

```
summary(Taille)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
158.0  169.8   175.5   175.1   182.2   192.0
```

- (3) Calculer l'écart type en tapant :

```
sqrt(sum((Taille - mean(Taille))^2)/length(Taille))
```

Que donne :

```
sd(Taille)
```

Attention, en fait  ne calcule pas l'écart-type donné par (1.3) page 7 mais la déviation standard définie par :

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (m - n_i)^2} \quad (3.1)$$

Par la suite, pour simplifier, on parlera d'écart-type à la place de déviation standard : quand on vous demandera de calculer l'écart-type, vous pourrez utiliser la fonction `sd`.

- (4) Tracer l'histogramme des tailles grâce à la fonction `hist`.  
 (5) Tracer la ligne de points avec empilement pour valeurs égales :

```
stripchart(Taille, method = "stack")
```

- (6) N'oubliez pas de détachez à la fin de l'exercice :

```
detach(extraitM1APAP08data)
```

Voir les éléments de correction page 31.

DÉFINITION 3.6. Il existe un graphique très intéressant pour appréhender visuellement ce problème de symétrie : la *boîte de dispersion* (ou à moustache)

Le principe en est simple, la médiane est indiquée par un trait central gras, les deux quartiles forment les extrémités de la boîte et, sortant de la boîte, deux moustaches essaient de rejoindre les valeurs minimum et

maximum. Si ces deux valeurs sont trop éloignées<sup>1</sup>, elles apparaissent comme des points que les moustaches ne rejoignent pas.

Pour obtenir ce dessin, on utilise la fonction `boxplot`.

EXERCICE 3.7. Tracer la boîte de dispersion des tailles du fichier `'extraitM1APAP08data'`. Vous devriez obtenir la figure 6(c) page 37.

EXERCICE 3.8. Étudier les variables `'masse'` et `'taille'` du fichier `'M1IGAPASA12data.txt'`.

Voir les éléments de correction page 31.

EXERCICE 3.9 (facultatif). Reprendre la même étude que les exercices 3.5 et 3.7 pour le fichier de données `L3APA06.txt`.

### 3.5. Quelques exercices facultatifs

EXERCICE 3.10.

- (1) On définit l'Indice de Masse Corporelle par

$$\text{IMC} = \frac{\text{poids}}{\text{taille}^2} \quad (3.2)$$

où la taille est en mètre et le poids en kg.

- (2) Calculer votre IMC, puis ceux de l'ensemble des individus du fichier `"M1IGAPASA12data.txt"` en tapant

```
attach(M1IGAPASA12data)
IMC <- masse/(taille/100)^2
```

- (3) Tracer les différents graphes possibles pour l'IMC et déterminer les différents indicateurs statistiques.

- (4) L'Organisation Mondiale de la Santé (OMS) a défini les critères suivants :

- maigre : inférieur à 18.5
- normal : de 18.5 à 25
- surpoids : de 25 à 30
- obèse : supérieur à 30

On rappelle que cet indice n'a qu'une valeur indicative. Pour déterminer l'existence d'une obésité réelle, il faut faire d'autres mesures destinées à établir exactement la proportion de masse grasse, car c'est l'excès de masse grasse qui représente un facteur de risque.

Introduire une variable qualitative `corpulence` en tapant

```
corpulence <- cut(IMC, breaks = c(-Inf, 18.5, 25, 30, Inf), labels = c("maigre", "normal",
"surpoids", "obèse"), right = F, left = T)
```

- (5) Tracer les différents graphes possibles pour la corpulence et déterminer les différents indicateurs statistiques.

- (6) N'oubliez pas

```
detach(M1IGAPASA12data)
```

EXERCICE 3.11.

Reprendre l'exercice 3.10 avec les données des fichiers `L3APA06.txt` et `t3var.txt`.

Voir éléments de correction en page 31.

EXERCICE 3.12.

---

1. Par rapport à un critère que nous ne développerons pas ici

- (1) Enregistrer les notes suivantes correspondant à trois groupe de TD à partir du fichier `note3TD.txt`.  
Après l'avoir stocké, on pourra taper

```
source("note3TD.txt")
```

ou même directement sans le stocker (si la connexion internet est correcte) :

```
source("http://utbmjb.chez-alice.fr/UFRSTAPS/M1APA/fichiersR/note3TD.txt")
```

On obtient donc trois variables `groupe1`, `groupe2` et `groupe3`.

- (2) Utiliser la commande `hist` pour représenter les trois histogrammes.  
(3) Pour rendre les informations comparables, taper

```
notes <- c(groupe1, groupe2, groupe3)
groupes <- rep(c(1, 2, 3), c(length(groupe1), length(groupe2), length(groupe3)))
groupes <- factor(groupes)
```

- (4) Il existe d'autres paramètres descriptifs de la variabilité comme par exemple le coefficient de variation :

$$cv = \frac{\text{écart-type}}{\text{moyenne}}. \quad (3.3)$$

qui présente l'avantage d'être sans unité. Il permet deux types d'étude :

- comparer la variabilité de plusieurs variables quantitatives mesurées au sein d'un même échantillon
- comparer la variabilité de plusieurs échantillons pour une même variable.

Calculer le `cv` de chacun des trois groupes de TD et conclure.

Voir éléments de correction en page 32.

EXERCICE 3.13. Traiter l'annexe C, sur la fréquence d'apparition de lettre dans un texte donné.

### 3.6. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.2

- (1)  
(2) Les données manquantes sont les données qui ne figurent pas dans le questionnaires et sont indiquées par des "NA" (Not Available).

On peut en avoir les indices en tapant

```
which(is.na(M1IGAPASA12data), arr.ind = TRUE)
```

ce qui donne

```
   row col
[1,]   7   6
[2,]  11   6
[3,]  22   6
[4,]  29   6
[5,]  30   6
[6,]  31   6
[7,]  32   6
[8,]  33   6
[9,]   7  10
```

Dans ces données, il y a donc des données manquantes.

- (3)  
(4) Voir figures 3.2 et 3.3 page 35.

- (5) Quant le nombre de catégories est trop important, les graphes ne sont pas toujours lisibles. Voir la figure 3.4.
- (6) • Les effectifs et les pourcentages déterminés par  $\mathcal{R}$  sont donnés dans le tableau suivant

	effectifs	pourcentages
0	3	9.091
1	7	21.212
2	23	69.697

- Voir les trois graphiques de la figure 3.5.

(7)

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.5

(1)

(2) on obtient

```

mean(Taille)
[1] 175.125
min(Taille)
[1] 158
max(Taille)
[1] 192
range(Taille)
[1] 158 192
median(Taille)
[1] 175.5
quantile(Taille)
  0%   25%   50%   75%  100%
158.00 169.75 175.50 182.25 192.00
summary(Taille)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 158.0  169.8   175.5   175.1   182.2   192.0

```

(3) On obtient :

```

sqrt(sum((Taille - mean(Taille))^2)/length(Taille))
[1] 10.34937
sd(Taille)
[1] 11.06394

```

(4)

(5) Voir les figures 6(a) et 6(b) page 37.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.11

Voir figure 3.7.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.8

- (1) • On étudie la variable quantitative (ou numérique) 'taille'.

- Les différents résultats déterminés par  $\mathbb{R}$  sont donnés dans le tableau suivant

noms	valeurs
moyenne	168.787879
écart-type	8.695392
$Q_1$ (quartile à 25 %)	163
médiane	167
$Q_3$ (quartile à 75 %)	174
minimum	153
maximum	186
nombre	33

- Voir les trois graphiques sur la figure 3.8 pour la variable "taille".
- (2)
- On étudie la variable quantitative (ou numérique) 'masse'.
  - Les différents résultats déterminés par  $\mathbb{R}$  sont donnés dans le tableau suivant

noms	valeurs
moyenne	62.757576
écart-type	9.546041
$Q_1$ (quartile à 25 %)	56
médiane	62
$Q_3$ (quartile à 75 %)	68
minimum	47
maximum	85
nombre	33

- Voir les trois graphiques sur la figure 3.9 pour la variable "masse".

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.12

En tapant

```
cla <- seq(from = 4, to = 18, by = 1)
par(mfrow = c(3, 1))
hist(groupe1, breaks = cla, main = "Histogramme du groupe 1")
hist(groupe2, breaks = cla, main = "Histogramme du groupe 2")
hist(groupe3, breaks = cla, main = "Histogramme du groupe 3")
```

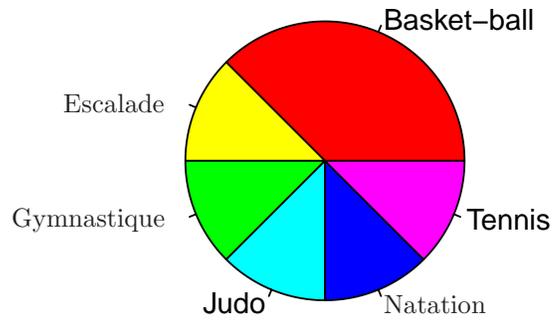
ou

```
cla <- seq(from = 4, to = 18, by = 1)
par(mfrow = c(3, 1))
for (i in 1:3) hist(notes[as.numeric(groupe) == i], breaks = cla, main = paste("Histogramme du groupe",
  as.character(i)))
```

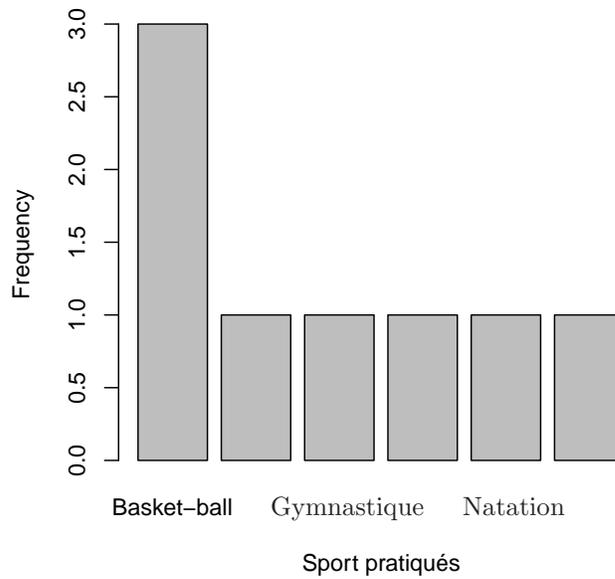
en obtient l'histogramme de la figure 3.10 page 39.

### 3.7. Ensemble des figures

### Sport pratiqués



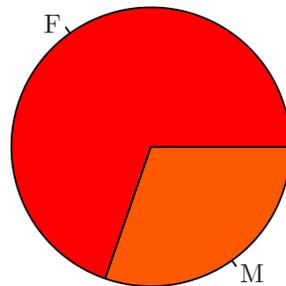
(a) : camembert



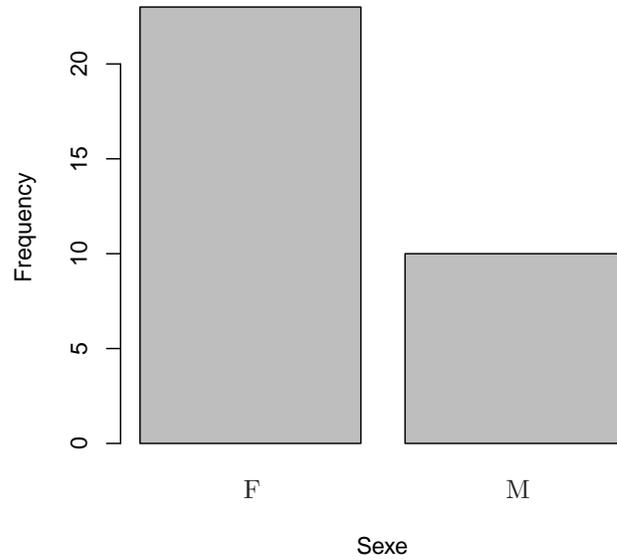
(b) : le graphique en barre

FIGURE 3.1. Deux graphiques pour la variable sport (données extraitM1APAP08data).

### Sexe



(a) : camembert



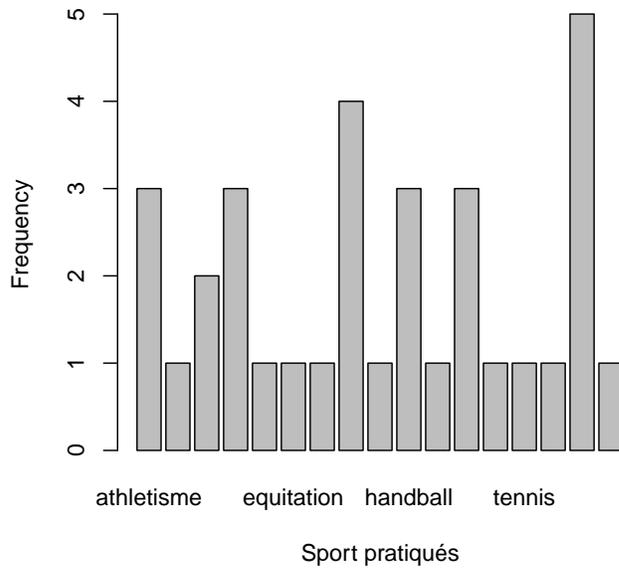
(b) : le graphique en barre

FIGURE 3.2. Deux graphiques pour la variable sexe (données du fichier "MIIGAPASA12data.txt").

### Sport pratiqués



(a) : camembert



(b) : le graphique en barre

FIGURE 3.3. Deux graphiques pour la variable sport (données du fichier "M1IGAPASA12data.txt").

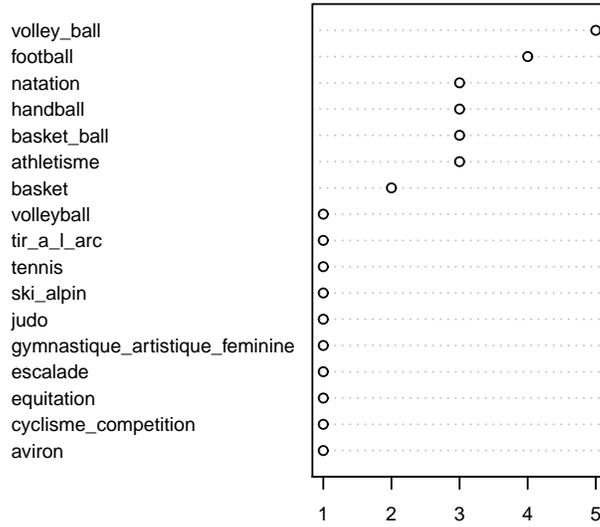


FIGURE 3.4. le diagramme de clévlend des sports (données du fichier "M1IGAPASA12data.txt").

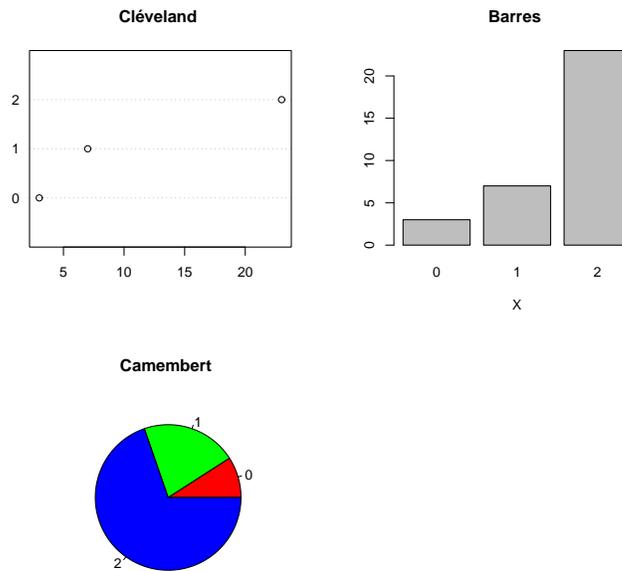
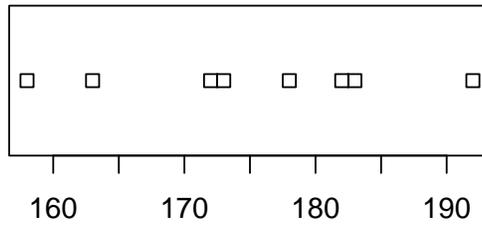
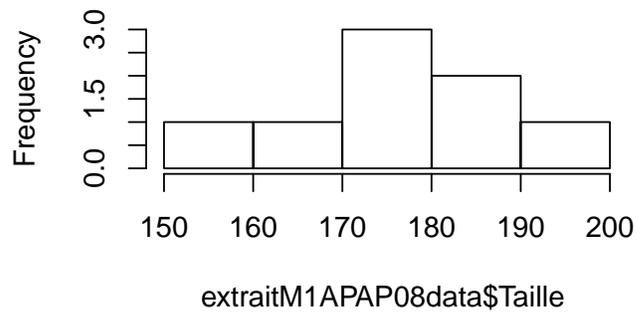


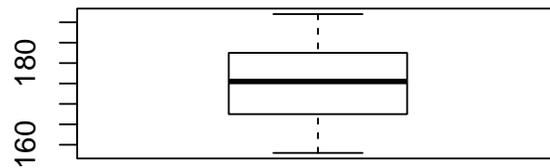
FIGURE 3.5. Trois graphiques pour la variable interet\_stats (données du fichier "M1IGAPASA12data.txt").



(a) : ligne de points



(b) : histogramme



(c) : boîte de dispersion

FIGURE 3.6. Trois graphiques pour la variable taille (données extraitM1APAP08data).

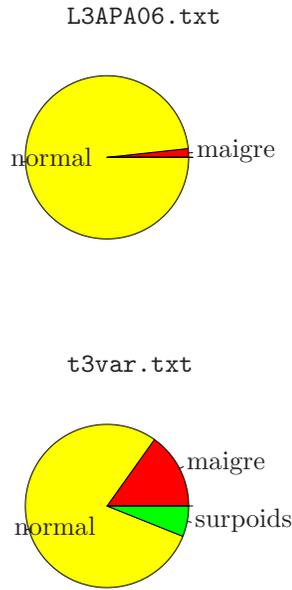


FIGURE 3.7. Le camembert des corpulence pour les fichiers L3APA06.txt et t3var.txt

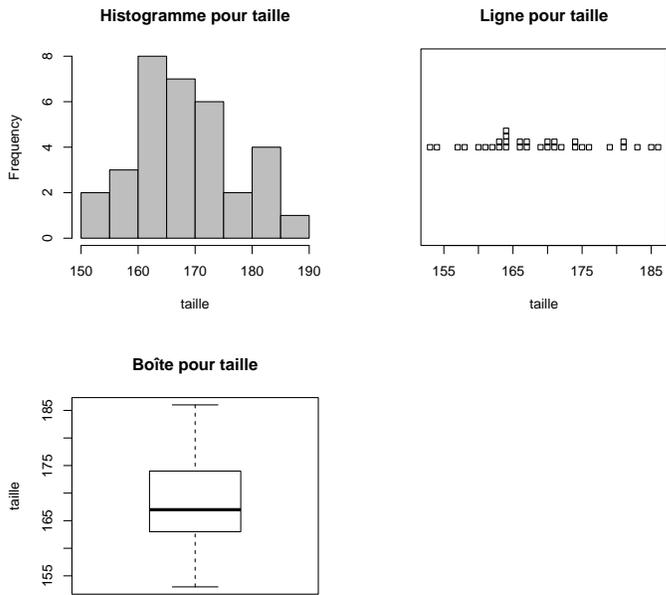


FIGURE 3.8. Trois graphiques pour la variable taille du fichier M1IGAPASA12data.txt.

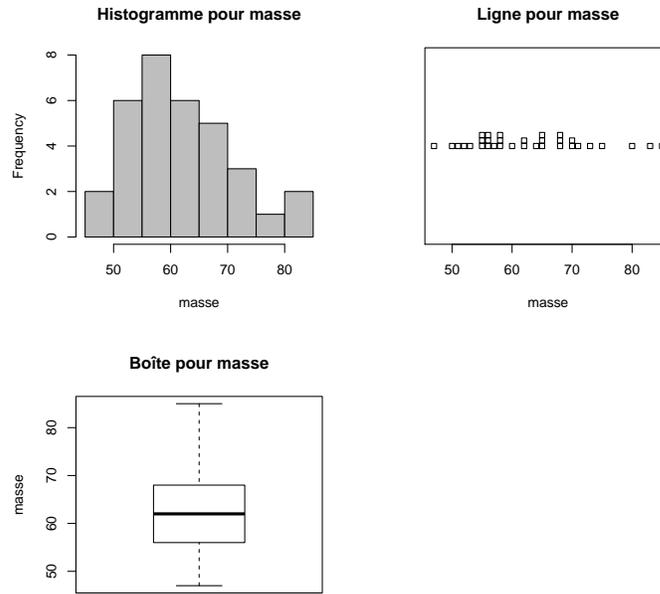


FIGURE 3.9. Trois graphiques pour la variable masse du fichier M1IGAPASA12data.txt.

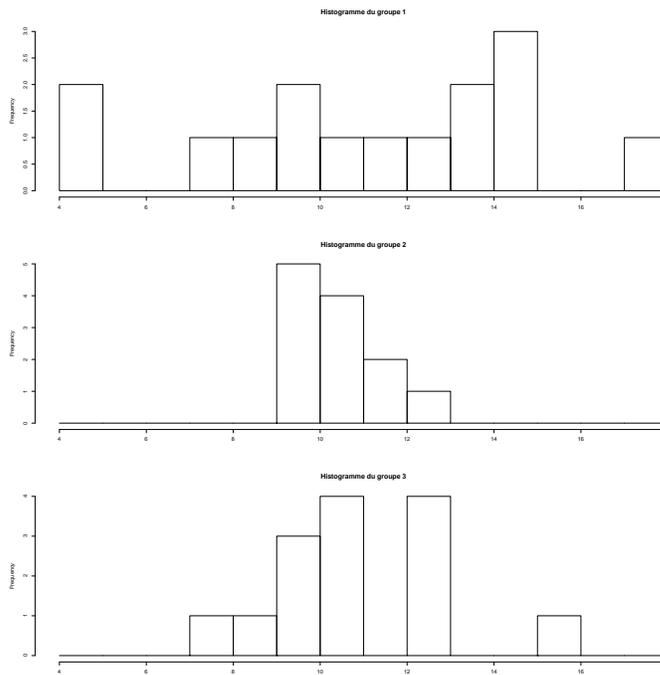


FIGURE 3.10. l'histogramme des trois groupes de TD



## Croisement de deux variables quantitatives

Ce chapitre s'inspire fortement du document [1] et du chapitre 9 de [5].

### 4.1. Introduction

Étudier une variable à la fois n'est généralement qu'un début lors de l'analyse d'un problème réel. Il va de soi que les travaux les plus intéressants consistent à relier plusieurs variables afin de comprendre les liaisons qu'elles entretiennent ou, pour le dire autrement, si l'on peut en "expliquer" certaines par d'autres<sup>1</sup>. Ainsi, même la description d'un phénomène aussi "simple" que la taille d'un individu serait bien limitée si l'on ne prenait pas en compte son âge et son sexe. Nous allons donc voir à présent des méthodes qui permettent de croiser deux informations. À nouveau, ces méthodes statistiques diffèrent selon la nature des variables. Suivant qu'elles sont numérique ou catégorielle, on s'orientera soit vers la régression (deux variables numériques, ce chapitre), vers l'analyse de tableau croisée (deux variables catégorielles, voir chapitre 5 page 57) ou vers l'analyse de variance (une numérique et une catégorielle, voir chapitre 6 page 67).

La situation statistique visée est extrêmement courante : il s'agit du cas où deux mesures numériques sont prises sur un même échantillon d'unités statistiques. On peut ainsi étudier s'il existe une relation entre la taille et le poids d'un groupe d'hommes, entre le prix au mètre carré et les impôts locaux pour des logements ...

On pourra consulter la fiche très complète [15].

### 4.2. Principe théorique

On s'intéresse donc à deux variables quantitatives  $X$  et  $Y$ . À chaque individu  $i$  pour  $1 \leq i \leq n$  sont donc associées deux valeurs  $x_i$  et  $y_i$ . Si une relation mathématique existe entre  $X$  et  $Y$ , il existe donc une fonction  $f$  telle que en théorie

$$Y = f(X) \quad (4.1)$$

ce qui se traduira donc par

$$\forall i \in \{1, \dots, n\}, \quad y_i \approx f(x_i). \quad (4.2)$$

Nous reviendrons sur ce signe  $\approx$  et comment caractériser la "précision" de cette approximation.

Il existe des tas de façon possible de déterminer  $f$  : on peut la chercher sous la forme d'un polynôme, d'une exponentielle, d'une somme de lignes trigonométriques en sinus et cosinus, d'une combinaison d'un grand nombre de fonctions connues.

Dans ce chapitre, nous n'étudierons que les *relations de type affine*, c'est-à-dire que la fonction  $f$  sera *affine* :

$$f(X) = aX + b \quad (4.3)$$

où on rappelle que  $a$  est la pente et  $b$  l'ordonnée à l'origine de la droite associée. De façon générale, on cherche à résoudre (4.2) au *sens des moindres carrés*, c'est-à-dire trouver une fonction  $f$  parmi un ensemble de fonctions données qui minimise l'expression :

$$\sum_{i=1}^n (f(x_i) - y_i)^2. \quad (4.4)$$

---

1. Restons toutefois prudent, le problème de la causalité est extrêmement délicat en statistique.

Plus petite sera cette quantité, meilleure sera l'approximation (4.2). Dans le cas de ce chapitre, on est dans l'hypothèse (4.3) ; les coordonnées  $(x_i, y_i)_{1 \leq i \leq n}$  sont connues (de façon expérimentale par mesure) et on cherche donc à résoudre le problème suivant :

$$\text{trouver } (a, b) \text{ qui minimise } S = \sum_{i=1}^n (ax_i + b - y_i)^2. \quad (4.5)$$

La quantité  $S$  est appelé l'écart entre les données et la droite d'équation  $Y = aX + b$ . Ce problème s'écrit aussi : trouver le couple  $(a_0, b_0)$  tel que

$$\forall (a, b) \in \mathbb{R}^2, \quad \sum_{i=1}^n (a_0x_i + b_0 - y_i)^2 \leq \sum_{i=1}^n (ax_i + b - y_i)^2 \quad (4.6)$$

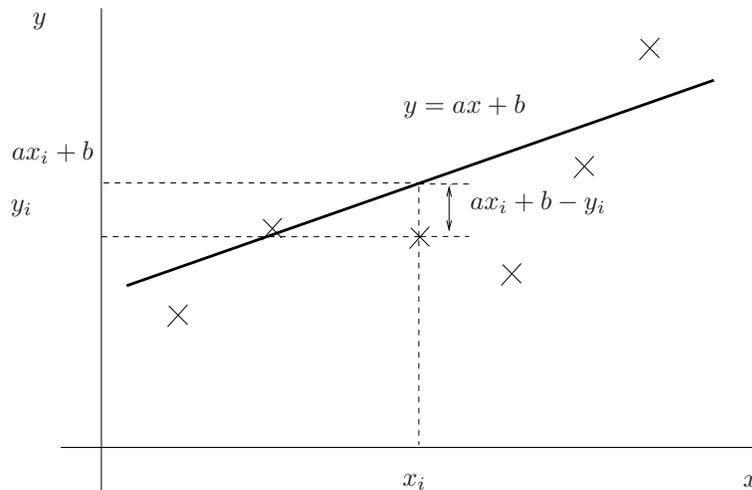


FIGURE 4.1. le principe de la droite de régression linéaire

Voir la figure 4.1.

On peut expliciter les coefficients  $a_0$  et  $b_0$  en fonction de  $(x_i, y_i)_{1 \leq i \leq n}$  ; voir par exemple la rubrique "régression linéaire" de Wikipédia ([http://fr.wikipedia.org/wiki/R%C3%A9gression\\_lin%C3%A9aire](http://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire)). Mais  $\mathbb{R}$  sait déterminer ces coefficients, par la suite notés  $a$  et  $b$ .

Voir par exemple la figure 4.2 page ci-contre, où sont tracés les points expérimentaux  $(x_i, y_i)_{1 \leq i \leq n}$ , deux droites différentes correspondant à deux couples  $(a, b)$  avec les écart associés et la "meilleure droite". Sur cette figure,

- les points de coordonnées  $(x_i, y_i)_{1 \leq i \leq n}$  sont représentés par des carrés noirs ;
- les points de coordonnées  $(x_i, ax_i + b)_{1 \leq i \leq n}$  sont représentés par des ronds bleu ;
- deux droites sont tracées en noir et la "meilleure" en rouge. Cette droite a une pente  $a$  positive.

Cette courbe et le script R permettant de la réaliser proviennent de [15].

### 4.3. La significativité pratique de la liaison

EXEMPLE 4.1. Avant de commencer à quantifier, il faut d'abord comprendre dans quelles situations on considère qu'une liaison est intense, c'est-à-dire que les points sont "bien" alignés. Le graphique 4.3 montre quatre situations possibles avec deux groupes représentés par des collections de lignes de points empilés.

On peut observer que les points sont "de moins en moins bien alignés" sur ces quatre graphiques.

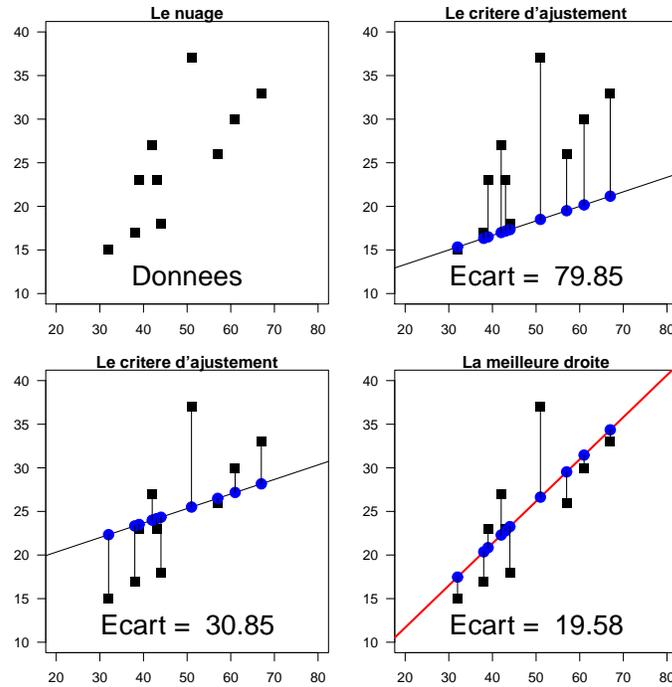


FIGURE 4.2. la droite de régression linéaire

DÉFINITION 4.2. On définit le *coefficient de corrélation linéaire*<sup>2</sup> comme une mesure de la liaison linéaire, c'est-à-dire de la capacité de prédire une variable  $X$  par une autre  $Y$  à l'aide d'une équation linéaire du type (4.1)-(4.3).

Nous ne donnons pas l'expression de ce coefficient, noté  $r$  (on pourra consulter l'URL de wikipédia donnée précédemment par exemple).

Notons que  $r$  est toujours compris entre -1 et 1. Il du signe de la pente  $a$  de la droite. Plus la valeur absolue  $|r|$  de ce nombre est proche de 1, "plus les points sont alignés". Dans ce cas l'approximation (4.2) sera d'autant meilleure. Autrement dit, plus  $|r|$  est proche de 1, plus l'écart  $S$  défini par (4.5) est proche de 0. Si  $|r| = 1$ , alors  $S = 0$  et le points sont alignés.

Cohen dans [16] a introduit les seuils  $r_1 = 0.1$ ,  $r_2 = 0.3$  et  $r_3 = 0.5$  permettant de quantifier la significativité pratique de la liaison

$$\text{si } |r| \begin{cases} < r_1, & \text{la significativité pratique de la liaison linéaire est faible,} \\ \in [r_1, r_2[, & \text{la significativité pratique de la liaison linéaire est moyenne,} \\ \in [r_2, r_3[, & \text{la significativité pratique de la liaison linéaire est forte,} \\ > r_3, & \text{la significativité pratique de la liaison linéaire est très forte} \end{cases} \quad (4.7)$$

---

2. il sera plus exacte de dire affine.

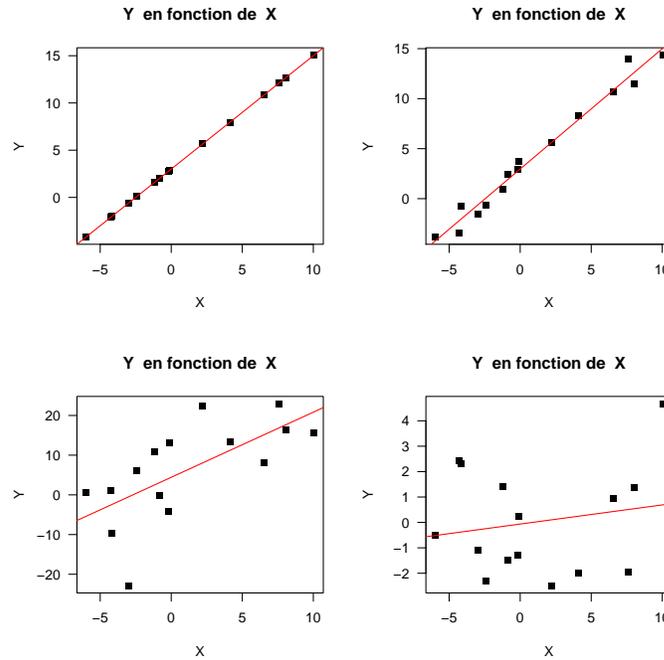


FIGURE 4.3. Quatre situations concernant quatre nuages de points

EXEMPLE 4.3. Donnons les différentes valeurs des  $r$  pour l'exemple 4.1 page 42 qui sont dans l'ordre des graphiques de la figure 4.3

$$\begin{aligned} r_1 &= 1, \\ r_2 &= 0.989064, \\ r_3 &= 0.669144, \\ r_4 &= 0.180478. \end{aligned}$$

Ainsi, les quatre graphiques montrent l'exemple de variables allant successivement d'une situation très fortement liée à une situation non liée.

#### 4.4. La significativité statistique de la liaison

Une autre approche est de considérer la significativité statistique : on essaie de voir *si un résultat tel que celui qu'on a obtenu aurait pu se produire par hasard* ? Plus précisément, si des groupes (de même taille, avec les mêmes données numériques) avaient été formés complètement au hasard, quelle serait la valeur du rapport de corrélation ?

On peut calculer une proportion de fois, où un coefficient de corrélation linéaire simulé au hasard dépasse celui réellement observé sur l'échantillon. On appelle cette quantité *probabilité critique* ou *p value*<sup>3</sup>.

Plus précisément, décrivons cela : Imaginons (sans le faire ...) qu'informatiquement, nous puissions simuler au hasard des données de mêmes taille et issues d'une même population. On obtiendrait alors une autre valeur de la corrélation linéaire. Si on

3. En réalité la construction théorique est un peu différente, basée sur la théorie des probabilité plutôt que sur des simulations informatiques, mais on montre que les deux méthodes, simulation informatique et théorie probabiliste, convergent vers le même résultat.

faisait un grand nombre de fois cette simulation, on pourrait calculer le nombre de fois où la corrélation observée est supérieure à celle obtenue sur nos données. On en déduit alors une proportion notée  $p_c$ . Sur un grand nombre de telles simulations, on va voir si la situation observée dans le jeu de données est exceptionnelle - et dans ce cas, il y a une relation statistiquement significative et  $p_c$  est petit- ou bien si la situation aurait pu se produire par hasard - et la relation n'est donc pas statistiquement significative et dans ce cas  $p_c$  est grand.

La formulation complète de cette façon de procéder fait partie de la théorie des tests d'hypothèses. Un certain nombre d'hypothèses (notamment la normalité des données) devraient être vérifiées, en toute rigueur, avant de conclure.

Ce nombre  $p_c$  est probabilité critique ou p value en anglais.  $\diamond$

DÉFINITION 4.4. La probabilité critique  $p_c$  est comprise entre 0 et 1. Proche de zéro (inférieure ou égale à  $0.05 = 5\%$ , valeur traditionnellement choisie) elle indique une relation statistiquement significative, c'est-à-dire qui a peu de chance d'être due au hasard. En revanche, strictement supérieure à 0.05, elle indique que la relation n'est pas statistiquement significative donc qu'elle peut-être due au hasard.

Nous indiquerons son calcul sous R en section 4.5.

## 4.5. Avec $\mathbb{R}$

L'exemple 4.5 et l'exercice 4.11 de la cette section sont issus de [4].

### EXEMPLE 4.5.

Le jeu de données 'coureurs.txt' disponible à l'URL habituelle contient pour 13 coureurs de niveau moyen leur âge et leur fréquence cardiaque maximum. Ces deux mesures sont bien entendu numériques.

Que la fréquence cardiaque maximum (fcm) soit reliée à l'âge est un résultat classique de la littérature scientifique sportive qui a même été vulgarisé dans les ouvrages d'entraînement sous la forme d'une équation (appelé formule d'Astrand) :

$$\text{fcm} = 220 - \text{âge}. \quad (4.8)$$

Nous allons tenter de le confirmer sur notre (très) petit échantillon de sujets.

REMARQUE 4.6. Pour toute la suite de ce cours, nous allons utiliser des fonctions. Lire (ou relire) pour cela l'annexe D page 115.

- (1) La description d'une liaison entre deux variables numériques commence par la représentation du nuage de points  $(x_i, y_i)_{1 \leq i \leq n}$ . Il n'y a qu'une difficulté, comment choisir la variable qui sera présentée sur l'axe des X? Lorsque l'une des deux variables doit servir à "expliquer" l'autre, c'est la variable explicative qui est placée en X et la variable à expliquer en Y. Ici, on a décidé d'étudier l'évolution de la fréquence cardiaque en fonction de l'âge, X sera donc l'âge et Y la fréquence cardiaque. Si on souhaite simplement étudier si les deux variables sont reliées, de façon réellement symétrique, peu importe alors le choix de X et de Y. Autrement dit, les valeurs de  $a$  et de  $b$  dépendent de l'ordre (X,Y) ou (Y,X); en revanche, les valeurs de  $r$  et de  $p_c$  n'en dépendent pas.

La commande

```
plot(coureurs$age, coureurs$fcm)
```

donnera une figure ressemblant à celle de la figure du bas de 4.4 page suivante. On peut préciser la valeur de certains paramètres facultatifs qui jouent sur l'aspect du graphique : la commande

```
plot(coureurs$age, coureurs$fcm, pch = 15, las = 1, main = "fcm en fonction de l'age")
```

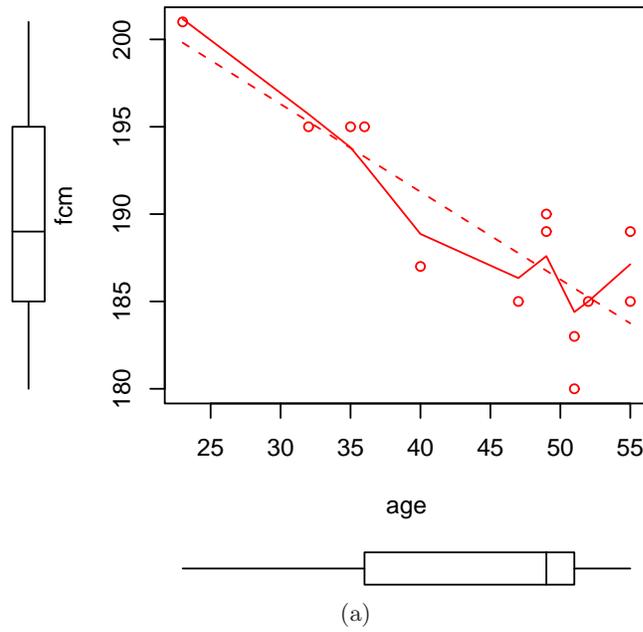
donnera la figure du bas de 4.4 page suivante.

Si on veut obtenir la figure du haut, il faudra charger le package `car` en tapant par exemple

```
library(car)
```

puis

```
scatterplot(coureurs$age, coureurs$fcm)
```



**fcm en fonction de l'age**

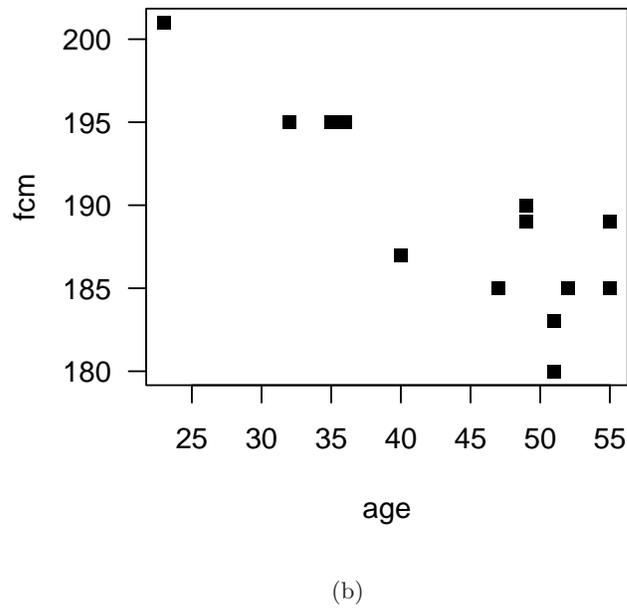


FIGURE 4.4. Nuages de points  $(x_i, y_i)_{1 \leq i \leq n}$  pour les données 'coureurs.txt'

Ou plus simplement, on tracera la droite de régression linéaire seule et le nuage de point en tapant

```
lmd <- lm(fcm ~ age, data = coureurs)
plot(coureurs$age, coureurs$fcm, pch = 15, las = 1, main = "fcm en fonction de l'age")
abline(lmd, col = "red")
```

- (2) Il faut maintenant déterminer les coefficients  $a$  et  $b$  de la droite, le coefficient de corrélation linéaire  $r$  et la probabilité critique  $p_c$ .

On tape

```
coefficients(lm(coureurs$fcm ~ coureurs$age))
```

et on voit apparaître

```
(Intercept)      age
211.3537051 -0.5019099
```

REMARQUE 4.7. On peut aussi taper la commande équivalente

```
coefficients(lm(fcm ~ age, data = coureurs))
```

La pente  $a$  se lit en dessous de `age` et l'ordonnée à l'origine  $b$  se lit en dessous de `(Intercept)`. On a donc ici

$$a = -0.5019, \quad (4.9a)$$

$$b = 211.3537. \quad (4.9b)$$

On peut aussi écrire

```
lmd <- lm(fcm ~ age, data = coureurs)
coeff <- coefficients(lmd)
b <- as.numeric(coeff[1])
a <- as.numeric(coeff[2])
```

Le coefficient directeur de  $a = -0.5$ , signifie que lorsque l'âge augmente d'une unité, c'est-à-dire 1 an, la fréquence cardiaque diminue de 0.5 unités (bpm). On obtient comme ordonnée à l'origine  $b = 211.3537$ , qui est très délicat à interpréter. Formellement, il signifie que si un individu à un âge nul, sa fréquence cardiaque maximum est de 211.3537. Ceci n'est pas absurde mais aucun nouveau né n'a été mesuré dans notre échantillon, le plus jeune ayant 20 ans, il ne faut donc pas se risquer à extrapoler cette valeur. On n'accordera en général pas beaucoup d'importance à l'interprétation de ce paramètre sauf pour le tracé de la droite (ordonnée à l'origine).

Pour obtenir  $p_c$ , on tape

```
summary(lm(coureurs$fcm ~ coureurs$age))
```

On obtient :

Call:

```
lm(formula = coureurs$fcm ~ coureurs$age)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-5.756 -2.756  1.190  1.715  5.251
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  211.35371     4.25279   49.698 2.69e-14 ***
coureurs$age  -0.50191     0.09394   -5.343 0.000236 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.266 on 11 degrees of freedom  
Multiple R-squared: 0.7218, Adjusted R-squared: 0.6965  
F-statistic: 28.54 on 1 and 11 DF, p-value: 0.0002365

La probabilité critique se lit en face de p-value; on a donc

$$p_c = 0.000236 \quad (4.10)$$

Pour obtenir  $r$ , on tape

```
cor(coueurs$age, coueurs$fcm)
[1] -0.8496042
```

REMARQUE 4.8. *Attention*, les données manquantes ne sont pas prises en compte par cette ligne de commande; pour palier cette difficulté, il faudra taper par exemple

```
indc <- !is.na(coueurs$age) & !is.na(coueurs$fcm)
r <- cor(coueurs$age[indc], coueurs$fcm[indc])
```

On a donc

$$r = -0.8496 \quad (4.11)$$

Ici, au vu des seuils de Cohen, on a une très forte liaison pratique et la liaison est statistiquement significative.

On peut utiliser la fonction `determin.quantiquanti` (voir annexe D page 115) disponible sur le site et qui fournit directement les valeurs de  $a$ ,  $b$ ,  $r$  et  $p_c$  : en tapant (ici, on a indiqué X puis Y)

```
determin.quantiquanti(coueurs$age, coueurs$fcm)
```

ce qui donne les 4 valeurs (naturellement identiques à celle de (4.9), (4.10),(4.11) ), sous la forme d'une liste (ce qui permet d'avoir plusieurs arguments de sortie)

```
$a
[1] -0.5019099
```

```
$b
[1] 211.3537
```

```
$r
[1] -0.8496042
```

```
$pc
[1] 0.0002364563
```

On pourra pour comprendre comment fonctionne une liste en tapant par exemple

```
res <- determin.quantiquanti(coueurs$age, coueurs$fcm)
class(res)
[1] "list"
names(res)
[1] "a" "b" "r" "pc"
res$a
[1] -0.5019099
res$pc
```

```
[1] 0.0002364563
```

On pourra aussi tracer le graphique de la droite de régression en utilisant les arguments optionnels de cette fonction qui sont `echo` et `fig`. Si `'echo'` est vrai, les résultats sont données et si `'fig'` est vrai la figure est créée. Comparer ce que donne

```
determin.quantiquanti(coueurs$age, coueurs$fcm)
determin.quantiquanti(coueurs$age, coueurs$fcm, fig = T)
determin.quantiquanti(coueurs$age, coueurs$fcm, fig = T, echo = F)
res <- determin.quantiquanti(coueurs$age, coueurs$fcm, echo = T)
res
```

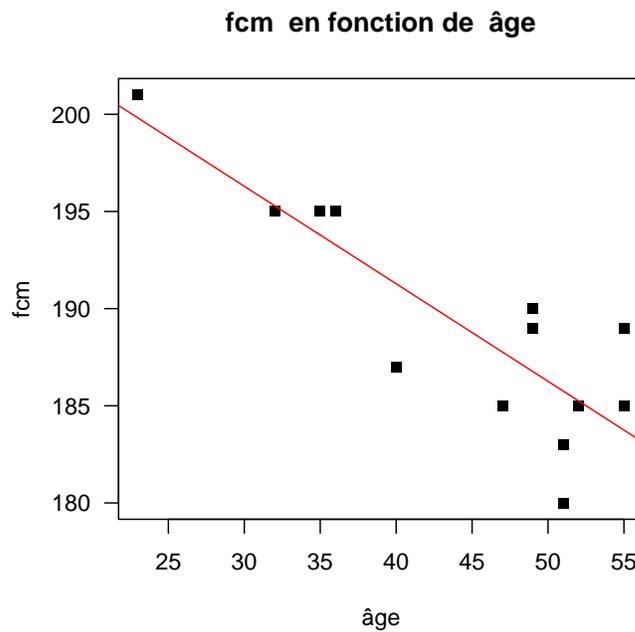


FIGURE 4.5. Le nuage de point et la droite de régression pour les données `'coueurs.txt'`

Pour obtenir une figure analogue à celle de la figure 4.5, il suffit de taper directement

```
determin.quantiquanti(coueurs$age, coueurs$fcm, fig = T, echo = F)
```

On pourra aussi préciser les labels  $X$  et  $Y$  des axes (choisis par défaut égaux à "X" et "Y") en tapant

```
determin.quantiquanti(coueurs$age, coueurs$fcm, fig = T, echo = F,
  labelX = "âge", labelY = "fcm")
```

REMARQUE 4.9. Constater en tapant

```
determin.quantiquanti(coueurs$fcm, coueurs$age)
```

que l'ordre de  $x$  ou  $y$  a une influence sur les valeurs de  $a$  et de  $b$  mais pas de  $r$  et de  $p_c$ .

REMARQUE 4.10. En accord avec la remarque 4.9, si on tape

```
lmd <- lm(fcm ~ age, data = coueurs)
lmdinv <- lm(age ~ fcm, data = coueurs)
```

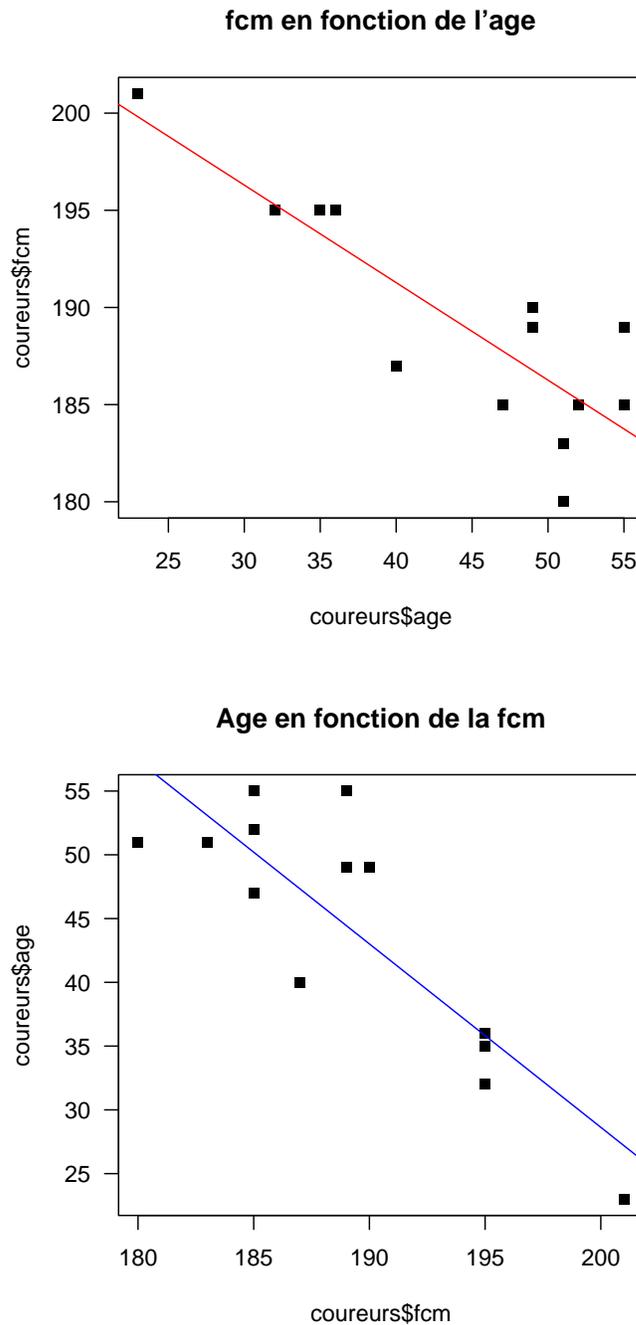


FIGURE 4.6. Le nuage de point et les deux droites de régression pour les données 'coureurs.txt'

```

par(mfrow = c(2, 1))
plot(coureurs$age, coureurs$fcv, pch = 15, las = 1, main = "fcv en fonction de l'age")
abline(lmd, col = "red")

```

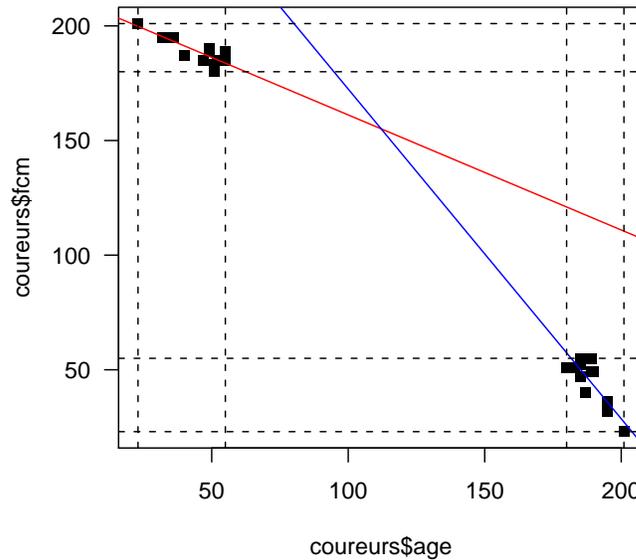


FIGURE 4.7. Le nuage de point et les deux droites de régression sur le même graphique pour les données 'coureurs.txt'

```
plot(coureurs$fcm, coureurs$age, pch = 15, las = 1, main = "Age en fonction de la fcm")
abline(lmdinv, col = "blue")
```

on constate que l'ordre des variables compte pour la droite de régression. Voir figures 4.6 et 4.7.

- (3) On peut utiliser la droite de régression déterminée pour déterminer "sa" fcm en tapant dans la fenêtre de Rgui :

```
res <- determin.quantiquanti(coureurs$age, coureurs$fcm)
monage <- 24
mafcm <- res$a * monage + res$b
```

On peut rajouter ce point sur le graphique déjà tracé en tapant par exemple

```
plot(coureurs$age, coureurs$fcm, pch = 15, las = 1, main = "fcm en fonction de l'age")
abline(lmd, col = "red")
points(monage, mafcm, pch = 19, col = "blue")
```

Voir figure 4.8.

- (4) Enfin, on peut rajouter la droite d'astran en tapant

```
plot(coureurs$age, coureurs$fcm, pch = 15, las = 1, main = "fcm en fonction de l'age")
abline(lmd, col = "red")
abline(220, -1, col = "green")
points(monage, mafcm, pch = 19, col = "blue")
```

Voir figure 4.8.

- (5) On peut aussi rajouter la fcm "prédite" par la regression linéaire en tapant par exemple

```
lmd <- lm(fcm ~ age, data = coureurs)
fcmtheo <- predict(lmd)
```

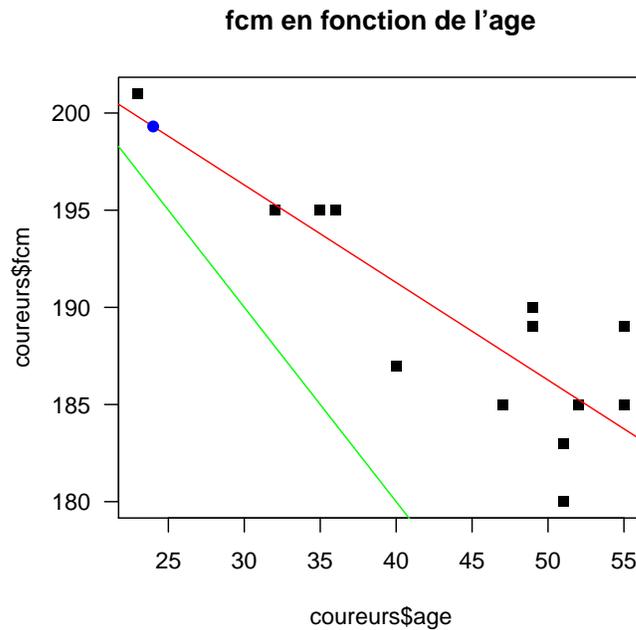


FIGURE 4.8. Le nuage de point, la droite de régression avec le point de coordonnées (24, 199.3079) et la droite d'Astran ( $fcem=220-\text{âge}$ ) en vert pour les données 'coureurs.txt'

```
plot(coureurs$age, coureurs$fcem, pch = 15, las = 1, main = "fcem en fonction de l'age")
abline(lmd, col = "red")
points(age, fcemtheo, pch = 20, col = "blue")
segments(age, fcemtheo, age, fcem)
```

Voir figure 4.9.

EXERCICE 4.11. Charger le fichier L3APA06.txt à l'URL habituelle.

- (1) Définissez une nouvelle variable égale à l'IMC : Indice de masse Corporelle, voir (3.2) page 29.
- (2) Étudier les relations ('poids','taille'), ('poids','IMC') et ('taille','IMC').

Voir éléments de correction page 53

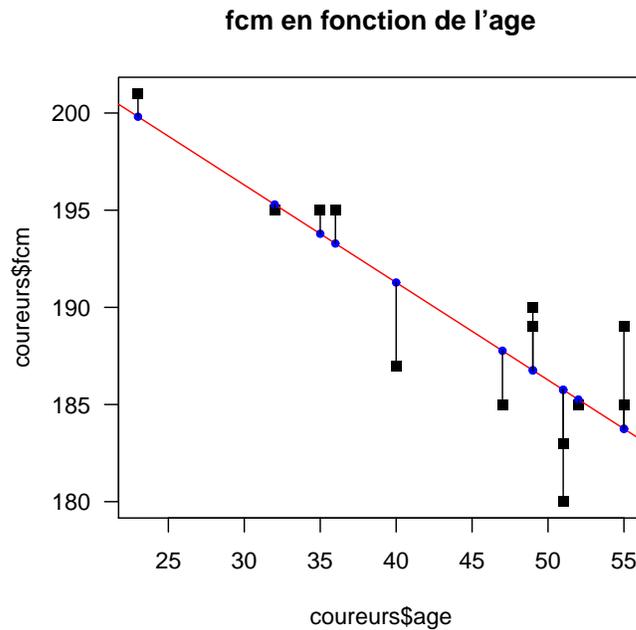


FIGURE 4.9. Le nuage de point, la droite de régression et les fmv prédites pour les données 'coureurs.txt'

#### 4.6. Sur les dangers de la régression linéaire abusive : exemple d'Anscombe

On consultera l'annexe E.

#### 4.7. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.11

On prendra garde au fait que la taille du fichier de donnée est en cm. On utilisera donc la formule suivante de l'IMC :

$$\text{IMC} = \frac{\text{poids}}{(\text{taille}/100)^2}, \quad (4.12)$$

(1) Étude de la relation ('poids', 'taille')

- On étudie le croisement de la variable quantitative (ou numérique) 'poids' et de la variable quantitative (ou numérique) 'taille'.
- Voir la figure 4.10 page suivante. Sur cette figure, les points semblent assez bien alignés.
- Confirmons cela grâce à  $\mathcal{R}$ .

Les résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
pente $a$	0.832284
ordonnée à l'origine $b$	118.708569
corrélation linéaire $r$	0.891533
probabilité critique $p_c$	6.50023e-21

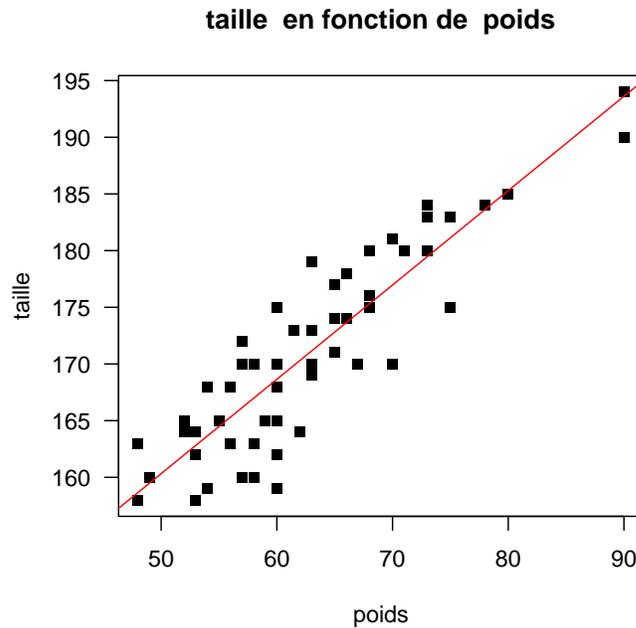


FIGURE 4.10. Le nuage de point et la droite de régression

On compare la valeur absolue de la corrélation linéaire  $r = 0.891533$  aux seuils de Cohen (0.1, 0.3, 0.5) (voir [16]) et la probabilité critique  $p_c = 6.50023e-21$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	<b>très forte</b>
significativité statistique	<b>oui</b>

- On peut donc affirmer il existe une relation entre les variables 'poids' et 'taille'.

(2) Étude de la relation ('poids', 'IMC')

- On étudie le croisement de la variable quantitative (ou numérique) 'poids' et de la variable quantitative (ou numérique) 'IMC'.
- Voir la figure 4.11 page suivante. Sur cette figure, les points semblent très bien alignés.
- Confirmons cela grâce à  $\mathcal{R}$ .

Les résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
pente $a$	0.121453
ordonnée à l'origine $b$	13.751405
corrélation linéaire $r$	0.742738
probabilité critique $p_c$	2.47844e-11

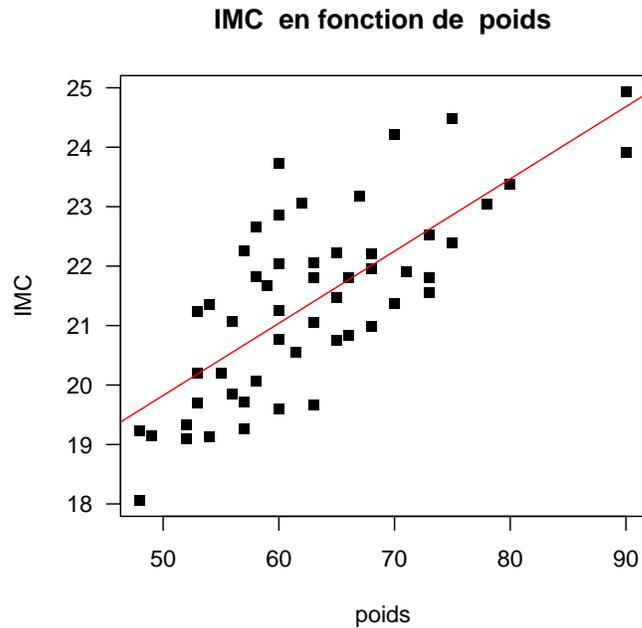


FIGURE 4.11. Le nuage de point et la droite de régression

On compare la valeur absolue de la corrélation linéaire  $r = 0.742738$  aux seuils de Cohen (0.1, 0.3, 0.5) (voir [16]) et la probabilité critique  $p_c = 2.47844e-11$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	<b>très forte</b>
significativité statistique	<b>oui</b>

- On peut donc affirmer il existe une relation entre les variables 'poids' et 'IMC'.

(3) Étude de la relation ('taille', 'IMC')

- On étudie le croisement de la variable quantitative (ou numérique) 'taille' et de la variable quantitative (ou numérique) 'IMC'.
- Voir la figure 4.12 page suivante. Sur cette figure, les points semblent très bien alignés.
- Confirmons cela grâce à  $\mathcal{R}$ .

Les résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
pende $a$	0.063981
ordonnée à l'origine $b$	10.465404
corrélation linéaire $r$	0.365269
probabilité critique $p_c$	0.00481162

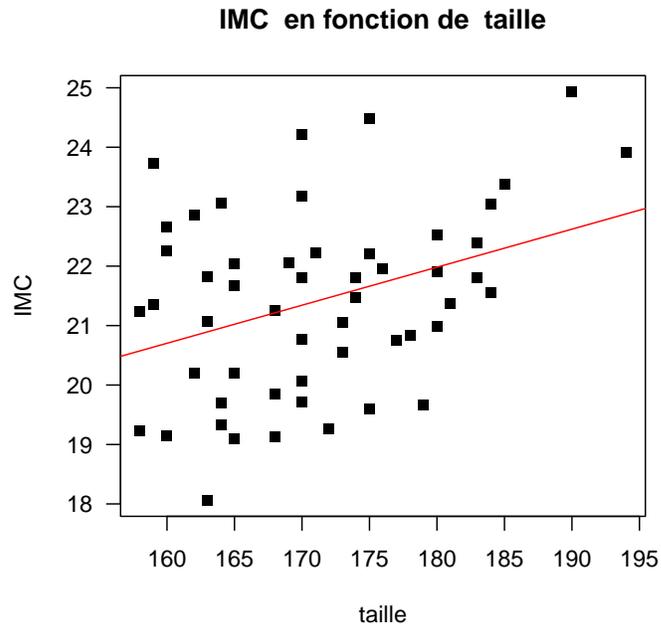


FIGURE 4.12. Le nuage de point et la droite de régression

On compare la valeur absolue de la corrélation linéaire  $r = 0.365269$  aux seuils de Cohen (0.1, 0.3, 0.5) (voir [16]) et la probabilité critique  $p_c = 0.00481162$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	<b>forte</b>
significativité statistique	<b>oui</b>

- On peut donc affirmer il existe une relation entre les variables 'taille' et 'IMC'.

## Croisement de deux variables qualitatives

Ce chapitre s'inspire fortement du document [2] et du chapitre 7 de [5].

### 5.1. Introduction

Il est possible de comparer deux groupes de mesures, mais que la mesure en question soit catégorielle (qualitative). On présente alors généralement les données sous une forme particulière dite du *tableau de contingence*. À nouveau, on commencera par décrire numériquement et graphiquement la liaison, puis on calculera un indicateur de liaison lié à un modèle statistique (celui d'*indépendance*) dont on définira la significativité pratique et statistique.

### 5.2. Principe théorique

Soient  $A$  et  $B$ , deux variables qualitatives ayant respectivement  $p$  et  $q$  modalités. Soit  $n$  le nombre d'individus sur lesquels  $A$  et  $B$  ont été observées. La table de contingence observée est un tableau croisé où les colonnes correspondent aux  $q$  modalités de la variable  $B$  et les lignes aux  $p$  modalités de la variable  $A$ . On note  $n_{ij}$  le nombre d'individus possédant à la fois la modalité  $i$  de la variable  $A$  et la modalité  $j$  de la variable  $B$ .

	$B_1$	...	$B_j$	...	$B_q$		total
$A_1$	$n_{11}$	...	$n_{1j}$	...	$n_{1q}$		$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
$A_i$	$n_{i1}$	...	$n_{ij}$	...	$n_{iq}$		$n_{i.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
$A_p$	$n_{p1}$	...	$n_{pj}$	...	$n_{pq}$		$n_{p.}$
total	$n_{.1}$	...	$n_{.j}$	...	$n_{.q}$		$n$

TABLE 5.1. Tableau de contingence

On obtient le tableau de contingence 5.1. Ici,  $n_{i.}$  est la somme des éléments de la  $i$ -ième ligne,  $n_{.j}$  est la somme des éléments de la  $j$ -ième colonne et  $n$  est la somme de tous les éléments du tableau (c'est aussi la somme des  $n_{i.}$  ou des  $n_{.j}$ , c'est aussi le nombre d'individus). Les différents coefficients  $n_{i.}$  et  $n_{.j}$  sont appelés les marges.

Prenons par exemple les deux variables sexe (M, F) et examen (R réussi et E échoué), observé sur 20 étudiants :

	E	R		total
F	4	6		10
M	6	4		10
total	10	10		20

Il y a donc

- 4 femmes qui échouent,
- 6 femmes qui réussissent,
- 6 hommes qui échouent,
- 4 homme qui réussissent,
- au total (partiel) 10 femmes,
- au total (partiel) 10 hommes,
- au total (partiel) 10 réussites,
- au total (partiel) 10 échecs.
- au total 20 individus interrogés.

On introduit alors la table de contingence théorique : répartition des 20 étudiants entre les différentes cases de la table s'il n'y a aucun lien entre les deux variables sexe et examen ; on ne tient compte que des marges qui indiquent ici que chacun des quatre catégories présentes doivent contenir un quart de la population totale, ce qui donne ici :

	E	R		total
F	5	5		10
M	5	5		10
total	10	10		20

DÉFINITION 5.1. De façon plus générale, on construit la table de contingence théorique sous l'hypothèse de l'indépendance des deux variables de façon que les marges soient égales. On traduit cette indépendance de la façon suivante :

- pour tout  $i \in \{1, \dots, q\}$ , la ligne  $L_i$  est proportionnelle à la ligne des marges  $(n_{.1}, \dots, n_{.j}, \dots, n_{.p})$ .
- pour tout  $j \in \{1, \dots, p\}$ , la colonne  $C_j$  est proportionnelle à la colonne des marges  $(n_{1.}, \dots, n_{i.}, \dots, n_{q.})$ .

On obtient la table de contingence théorique 5.2 où

$$\hat{n}_{ij} = \frac{n_{i.} n_{.j}}{n} \quad (5.1)$$

Remarquons *a posteriori* que les marges de la table des contingence théorique (5.1) sont bien celle de la table expérimentale : On a en effet les sommes en lignes

$$\begin{aligned} \sum_i \hat{n}_{ij} &= \sum_i \frac{n_{i.} n_{.j}}{n}, \\ &= \frac{n_{.j}}{n} \sum_i n_{i.}, \\ &= n_{.j} \end{aligned}$$

et les sommes en colonnes

$$\begin{aligned}\sum_j \widehat{n}_{ij} &= \sum_j \frac{n_{i \cdot} n_{\cdot j}}{n}, \\ &= \frac{n_{i \cdot}}{n} \sum_j n_{\cdot j}, \\ &= n_{i \cdot}.\end{aligned}$$

PREUVE FACULTATIVE. la première hypothèse se traduit par l'existence d'un nombre  $\alpha_i$  (qui ne dépend que la ligne  $i$ ) tel que

$$\forall(i, j), \quad \widehat{n}_{ij} = \alpha_i n_{\cdot j} \quad (5.2)$$

Si on somme ces équations par rapport à  $j$ , on a

$$n_{i \cdot} = \sum_j \alpha_i n_{\cdot j} = \alpha_i \sum_j n_{\cdot j} = \alpha_i n,$$

et donc

$$\alpha_i = \frac{n_{i \cdot}}{n}.$$

En reportant dans (5.2), on obtient

$$\forall(i, j), \quad \widehat{n}_{ij} = \frac{n_{i \cdot}}{n} n_{\cdot j}$$

ce qui le résultat annoncé. De même, on aboutirait au même résultat en utilisant la seconde hypothèse.  $\square$

	$B_1$	...	$B_j$	...	$B_q$		total
$A_1$	$\widehat{n}_{11}$	...	$\widehat{n}_{1j}$	...	$\widehat{n}_{1q}$		$n_{1 \cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
$A_i$	$\widehat{n}_{i1}$	...	$\widehat{n}_{ij}$	...	$\widehat{n}_{iq}$		$n_{i \cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
$A_p$	$\widehat{n}_{p1}$	...	$\widehat{n}_{pj}$	...	$\widehat{n}_{pq}$		$n_{p \cdot}$
total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot q}$		$n$

TABLE 5.2. Tableau de contingence théorique

DÉFINITION 5.2. On introduit enfin le  $\chi^2$  qui mesure l'écart entre la table théorique et la table observée défini par

$$\chi^2 = \sum_{ij} \frac{(\widehat{n}_{ij} - n_{ij})^2}{\widehat{n}_{ij}} \quad (5.3)$$

Si  $\chi^2 = 0$ , les effectifs observés sont identiques aux effectifs théoriques et il y a indépendance entre les deux variables. Si  $\chi^2$  est petit, les effectifs observés sont presque identiques aux effectifs théoriques. Les deux variables sont peu liées entre elles. Si  $\chi^2$  est grand, les effectifs observés sont différents des effectifs théoriques. Les deux variables sont liées entre elles.

### 5.3. La significativité pratique de la liaison

DÉFINITION 5.3. Afin d'évaluer le degré de relation entre les deux variables qualitatives, divers indices ont été proposés. On rappelle que

- $n$  est le nombre total d'individu ;
- $p$  est le nombre de lignes du tableau de contingence (le nombre de modalités de la première variable) ;
- $q$  est le nombre de colonnes du tableau de contingence (le nombre de modalités de la seconde variable).

Nous avons en avons retenu deux indices :

- (1) L'indice de Cramer défini par

$$V = \sqrt{\frac{\chi^2}{n \min(p-1, q-1)}}. \quad (5.4)$$

Il varie entre 0 et 1. Si le coefficient est proche de 0, les variables ne sont pas liées. Si le coefficient est proche de 1, les variables sont liées.

- (2) La taille d'effet  $w$  introduite par Cohen [16] définie par

$$w = \sqrt{\frac{\chi^2}{n}}. \quad (5.5)$$

avec les seuils correspondant  $w_1 = 0.1$ ,  $w_2 = 0.3$  et  $w_3 = 0.5$  tels que

$$\text{si } w \begin{cases} < w_1, & \text{la significativité pratique de la liaison est faible,} \\ \in [w_1, w_2[, & \text{la significativité pratique de la liaison est moyenne,} \\ \in [w_2, w_3[, & \text{la significativité pratique de la liaison est forte,} \\ > w_3, & \text{la significativité pratique de la liaison est très forte} \end{cases} \quad (5.6)$$

### 5.4. La significativité statistique de la liaison

Comme dans la section 4.4, on introduit une probabilité critique  $p_c$ , comprise entre 0 et 1. Proche de zéro (inférieure ou égale à  $0.05 = 5\%$ , valeur traditionnellement choisie) elle indique une relation statistiquement significative, c'est-à-dire qui a peu de chance d'être due au hasard. En revanche, strictement supérieure à 0.05, elle indique que la relation n'est pas statistiquement significative donc qu'elle peut-être due au hasard.

Nous indiquerons son calcul sous R en section 5.5.

### 5.5. Avec

Nous allons travailler sur le fichier `hebergement.txt` disponible à l'URL habituelle. Les résultats des questionnaires sont rentrés sous la forme d'un tableau contenant en lignes les enquêtés et en colonnes leurs réponses aux différentes questions. Nous avons conservé dans un tableau les résultats aux deux questions : classe d'âge (attention, la variable `age` est quantitative, mais mise sous forme de classes, elle devient qualitative) et type d'hébergement pendant les vacances.

Il comprend 591 lignes et 2 colonnes. On va donc croiser les deux variables qualitatives '`age`' et '`logement`'.

- (1) Construction de la table de contingence

Il faut taper la commande suivante :

```
table(hebergement$age, hebergement$logement)
```

ou

```
xtabs(~hebergement$age + hebergement$logement)
```

On obtient le tableau 5.3.

	camping	non_camping
16-19	7	4
20-24	43	21
25-29	37	37
30-39	99	78
40-49	79	62
50-54	19	25
55-59	15	15
60-65	9	15
plus_de_65	2	24

TABLE 5.3. table de contingence du type d'hébergement en fonction des classes d'âges

- (2) Pour créer la table de contingence théorique sous l'hypothèse de l'indépendances des deux variables,  
Tapez

```
tab <- table(hebergement$age, hebergement$logement)
res <- chisq.test(tab)
res$expected
```

Que font les commandes :

```
margin.table(tab, 1)
margin.table(tab, 2)
prop.table(tab, 1)
prop.table(tab, 2)
```

	camping	non_camping
16-19	5.770	5.230
20-24	33.570	30.430
25-29	38.816	35.184
30-39	92.843	84.157
40-49	73.959	67.041
50-54	23.080	20.920
55-59	15.736	14.264
60-65	12.589	11.411
plus_de_65	13.638	12.362

TABLE 5.4. table de contingence théorique du type d'hébergement en fonction des classes d'âges

On obtient le tableau 5.4. Remarquons qu'il comprend  $p = 9$  lignes et  $q = 2$  colonnes.

- (3) Pour obtenir le  $\chi^2$ ,

– soit on tape

```
tab <- table(hebergement$age, hebergement$logement)
res <- chisq.test(tab)
res
```

et on voit apparaître sa valeur en face de X-square :

**Pearson's Chi-squared test**

```

data: tab
X-squared = 32.5107, df = 8, p-value = 7.543e-05
- soit on tape
  tab <- table(hebergement$age, hebergement$logement)
  res <- chisq.test(tab)
  res$statistic

```

On obtient

$$\chi^2 = 32.510687 \quad (5.7)$$

REMARQUE 5.4. Seule cette dernière étape est totalement nécessaire!

- (4) Pour calculer le coefficient de Cramer défini par (5.4), il faut évaluer

$$V = \sqrt{\frac{\chi^2}{n \min(p-1, q-1)}}.$$

Il faut d'abord déterminer  $n$ , le nombre total d'individus,  $p$ , le nombre de lignes du tableau de contingence (le nombre de modalités de la première variable) et  $q$ , le nombre de colonnes du tableau de contingence (le nombre de modalités de la seconde variable). Pour cela :

On tape dans "Rgui"

```

dim(hebergement)
[1] 591  2

```

ou

```

dim(hebergement)[1]
[1] 591

```

dont on déduit  $n = 591$ .

On tape ensuite dans "Rgui"

```

dim(tab)
[1] 9 2

```

dont on déduit que  $p = 9$ ,  $q = 2$ .

Bref, on obtient

$$\begin{aligned} n &= 591, \\ p &= 9, \\ q &= 2 \end{aligned}$$

et on a alors

$$V = \sqrt{\frac{32.510687}{591 \min(9-1, 2-1)}}.$$

On tape donc dans la fenêtre de Rgui :

```

sqrt(32.510687/(591*1))

```

On obtient donc

$$V = 0.2345413 \quad (5.8)$$

- (5) On procède de même pour la taille d'effet (5.5) :

$$\begin{aligned} w &= \sqrt{\frac{\chi^2}{n}}, \\ &= \sqrt{\frac{32.510687}{591}} \end{aligned}$$

et donc

$$w = 0.2345413 \quad (5.9)$$

Ici,  $w = V$  ! Pourquoi ?

Le  $V$  est proche de zéro donc les deux variables ne sont pas liées. Les seuils de Cohen données par (5.6) nous indique que la la significativité pratique de la liaison est moyenne.

(6) La probabilité critique ici se lit ici en face de "p-value". On obtient ici

$$p_c = 7.5e - 05 \quad (5.10)$$

Elle est inférieur au seuil de 5 % et donc on observe une significativité statistique de la liaison, donc non due au hasard.

(7) Création d'un graphique

Par la suite, nous n'utiliserons guère ce graphique mais nous citons tout de même son interprétation et sa création avec  $\mathbb{R}$ .

Il faut d'abord charger le package 'ade4', puis utiliser la fonction 'table.cont' en tapant :

```
tab <- table(hebergement$age, hebergement$logement)
table.cont(tab)
```

Sur ce graphique, apparaissent à la place de chaque nombre de la table de contingence, un carré dont la surface lui est proportionnel.

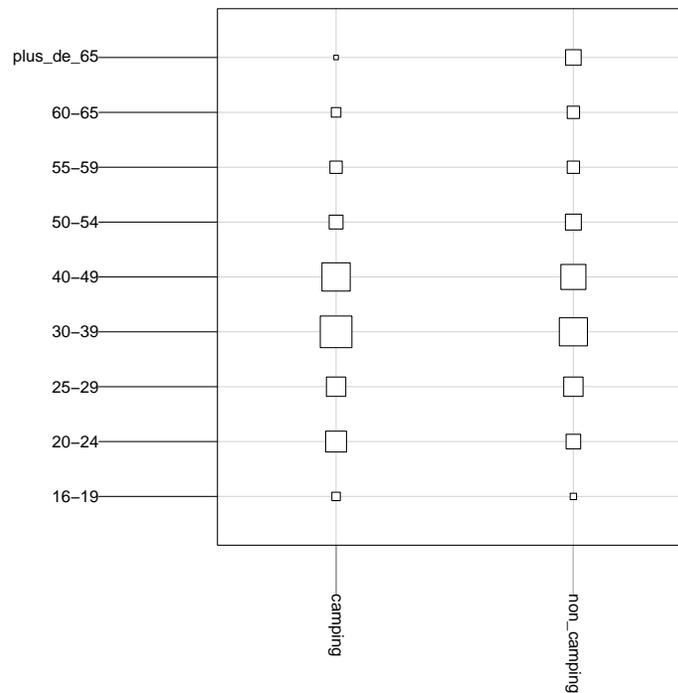


FIGURE 5.1. Le graphique illustrant la table de contingence des données du fichier 'hebergement.txt'.

Voir la figure 5.1.

Des arguments optionnels 'csize' et 'col.labels' permettent de donner un échelle pour la taille des carrés et un label pour les noms de colonne. Par exemple le graphique de la figure 5.1 a été créé en tapant

```
tab <- table(hebergement$age, hebergement$logement)
table.cont(tab, csize = 2, col.labels = colnames(tab))
```

(8) Calcul des indicateurs statistiques en une seule étape.

On peut utiliser la fonction `determin.qualiquali` (voir annexe D) disponible sur le site et qui fournit directement les valeurs de  $\chi^2$ ,  $V$ ,  $w$  et  $p_c$  :

```
determin.qualiquali(hebergement$age, hebergement$logement)
$chi2
[1] 32.51069

$V
[1] 0.2345413

$w
[1] 0.2345413

$pc
[1] 7.54284e-05
```

renvoie bien les valeurs indiquées par (5.7), (5.8), (5.9) et (5.10).

Notez que cette fonction a un argument facultatif `'tabcontingence'`, égal à faux par défaut. S'il est vrai, cette fonction renvoie en outre la table de contingence :

```
determin.qualiquali(hebergement$age, hebergement$logement, tabcontingence = T)
$table.cont
      y
x      camping non_camping
16-19      7           4
20-24     43          21
25-29     37          37
30-39     99          78
40-49     79          62
50-54     19          25
55-59     15          15
60-65      9          15
plus_de_65  2          24

$chi2
[1] 32.51069

$V
[1] 0.2345413

$w
[1] 0.2345413

$pc
[1] 7.54284e-05
```

EXERCICE 5.5. Considérons le jeu de données portant sur 592 étudiants (extrait de [17]) Pour chaque étudiant on a observé 3 variables qualitatives : la couleur des cheveux, la couleur des yeux et le sexe. Les données se trouvent dans le fichier 'qualitatif.txt' que vous pouvez télécharger à l'URL habituelle.

Calculer la taille d'effet, le coefficient de Cramer et la probabilité critique entre les deux variables la couleur des cheveux et la couleur des yeux. Conclure

Voir éléments de correction page 65

## 5.6. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 5.5

La table de contingence donne

x	y			
	Bleu	Marron	Noisette	Vert
Blond	94	7	10	16
Marron	84	119	54	29
Noir	20	68	15	5
Roux	17	26	14	14

	Bleu	Marron	Noisette	Vert
Blond	94	7	10	16
Marron	84	119	54	29
Noir	20	68	15	5
Roux	17	26	14	14

TABLE 5.5. Table de contingence des couleurs de cheveux en fonction de celle des yeux

Ces résultats sont présentés dans le tableau 5.5. Avec les notations précédentes, on a

$$\begin{aligned} p &= 4, \\ q &= 4, \\ n &= 592 \end{aligned}$$

On obtient grâce à  $\mathbb{R}$ ,

$$\begin{aligned} \chi^2 &= 138.289842, \\ p_c &= 2.32529e - 25. \end{aligned}$$

On peut donc successivement calculer le coefficient de Cramer  $V$

$$\begin{aligned} V &= \sqrt{\frac{\chi^2}{n \min(p-1, q-1)}}, \\ &= \sqrt{\frac{138.289842}{592 \min(4-1, 4-1)}}, \\ &= 0.279045 \end{aligned}$$

puis

$$\begin{aligned} w &= \sqrt{\frac{\chi^2}{n}}, \\ &= 0.483319 \end{aligned}$$

On peut aussi la fonction `determin.qualiquali` qui fournit directement les valeurs de  $\chi^2$ ,  $V$ ,  $w$  et  $p_c$  :

```
determin.qualiquali(qualitatif$cheveux, qualitatif$yeux)
```

```
$chi2
```

```
[1] 138.2898
```

```
$V
```

```
[1] 0.2790446
```

```
$w
```

```
[1] 0.4833195
```

```
$pc
```

```
[1] 2.325287e-25
```

renvoie bien les valeurs précédentes.

Ainsi, au vu des seuil de Cohen [16] (0.1, 0.3, 0.5), la liaison est considérée comme forte. Au vu de la probabilité critique de seuil (5 %), la liaison est significativement statistique, c'est-à-dire non due au hasard.

## Croisement d'une variable qualitative et d'une variable quantitative

Ce chapitre s'inspire fortement du document [3] et du chapitre 5 de [5].

### 6.1. Introduction

La situation statistique visée est extrêmement courante : il s'agit du cas où deux mesures sont prises sur un même échantillon d'unités statistiques, *l'une étant numérique et l'autre catégorielle*. On peut ainsi comparer la taille dans des groupes d'hommes et de femmes. On peut comparer des performances de vitesse selon différentes méthodes d'entraînement. On peut comparer des chiffres de vente selon les vendeurs, ou les journées de la semaine.

Nous allons pour commencer prendre une situation simple : le fichier 'notes3TDbis.txt' contient les notes d'étudiants répartis dans trois groupes de TD.

On souhaite donc sur cet exemple comparer les 15 groupes d'étudiants sur la base de leur note. En particulier, on va se demander si les notes dépendent du groupe. Le même problème peut se poser en termes de liaison et non pas de comparaison : Est-ce que la note de statistique est liée au groupe ? C'est cette deuxième formulation qui va expliquer l'organisation informatique des données.

### 6.2. Avec R

Nous avons deux mesures pour chaque unité statistique, ce qui signifie que le tableau de données va avoir 15 lignes (les 15 étudiants) et 2 colonnes (les 2 variables mesurées). On peut voir cette organisation en allant chercher le fichier 'notes3TDbis.txt' et en l'important sous le nom 'notes3TDbis'.

On vérifie que la variable 'notes' est bien une variable numérique et la variable 'groupes' est bien une variable catégorielle<sup>1</sup>.

#### 6.2.1. La description des groupes par les graphiques

Le principe de la description graphique est de réaliser un graphique pour chacun des groupes afin de pouvoir détailler, à un premier niveau, les caractéristiques de chaque groupe, mais aussi de les réaliser tous à la même échelle afin de permettre, à un deuxième niveau, une comparaison des différents groupes. On parle de *collection de graphiques*.

Il est assez difficile de comparer des histogrammes. En revanche, il faut sans doute se tourner vers la *collection de boîtes de dispersion*. Pour réaliser ce graphique, il faut

Taper la ligne de commande

```
boxplot(notes~groupes,data=notes3TDbis)
```

ou

```
boxplot(notes~groupes,xlab="groupes",ylab="notes",data=notes3TDbis)
```

On peut aussi utiliser une autre syntaxe équivalente (plus universelle) :

```
boxplot(notes3TDbis$notes~notes3TDbis$groupes)
```

---

1. un **factor** dans la terminologie du logiciel R

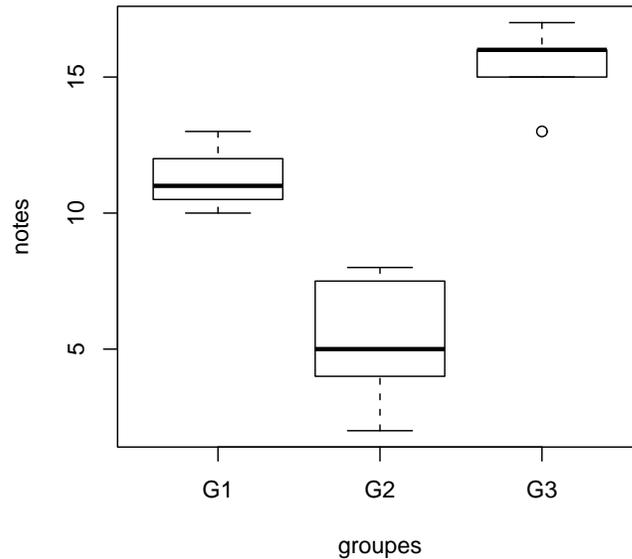


FIGURE 6.1. La collection de boîtes de dispersion des notes par groupe (données 'notes3TDbis.txt')

Que voit-on sur le graphique 6.1? Commençons par décrire la centralité : les trois médianes semblent être assez différentes (11, 5 et 16)

En ce qui concerne la dispersion, les trois groupes sont différentes. Enfin, il ne semble pas y avoir de valeurs extrêmes (il est vrai qu'avec des notes de 0 à 20, c'est peu probable).

Comme il y a très peu de données par groupe, on peut également envisager la création d'une collection de ligne de points (cf. figure 6.2 page ci-contre) Pour obtenir ce dessin, il faut taper dans la fenêtre "Rgui"

```
stripchart(notes3TDbis$notes~notes3TDbis$groupes,xlab="notes",ylab="groupes",method="stack")
```

### 6.2.2. La description des groupes par les statistiques

Afin de décrire ces aspects des deux distributions, on va utiliser les statistiques classiques : moyenne, médiane, écart-type...

Pour obtenir ces calculs,

Plusieurs solutions différentes :

- (1) Il faut calculer les statistiques, groupes par groupes, en tapant par exemple

```
summary(notes3TDbis$notes[notes3TDbis$groupes=="G1"])
```

ce qui donne

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.00  10.50   11.00   11.33  12.00   13.00
```

puis

```
mean(notes3TDbis$notes[notes3TDbis$groupes=="G1"])
```

ce qui donne

```
[1] 11.33333
```

et enfin

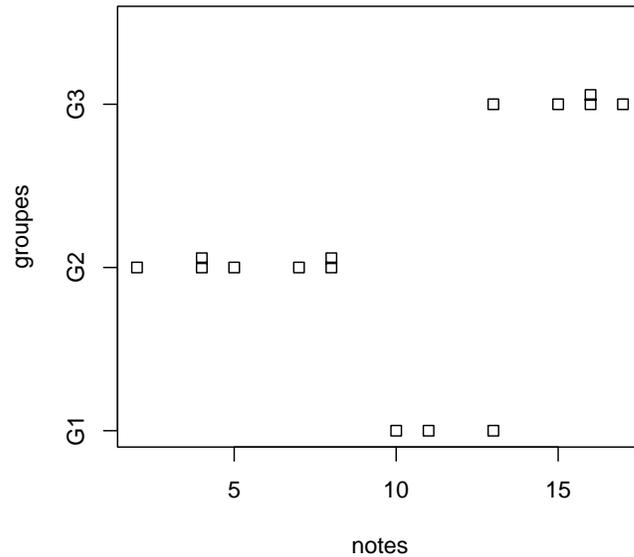


FIGURE 6.2. La collection de lignes de points des notes par groupes (données 'notes3TDbis.txt')

```
sd(notes3TDbis$notes[notes3TDbis$groupes=="G1"])
```

ce qui donne

```
[1] 1.527525
```

Et ainsi de suite pour les autres groupes ce qui donnerait

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	4.000	5.000	5.429	7.500	8.000

```
[1] 5.428571
```

```
[1] 2.299068
```

puis

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13.0	15.0	16.0	15.4	16.0	17.0

```
[1] 15.4
```

```
[1] 1.516575
```

- (2) On peut automatiser cela en utilisant une boucle sur les niveaux de la variable `Cheveux`

```
ni <- levels(notes3TDbis$groupes)
for (i in 1:length(ni)) {
  print(ni[i])
  print(summary(notes3TDbis$notes[notes3TDbis$groupes == ni[i]]))
  print(sd(notes3TDbis$notes[notes3TDbis$groupes == ni[i]]))
}
```

- (3) On pourra aussi utiliser la fonction `determin.qualiquanti`, disponible à l'URL habituelle et taper (pour comprendre la syntaxe utilisée, voir l'annexe D page 115)

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)
```

ou encore

```
res<-determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)
res$stat.groupe
```

ou ce qui est équivalent

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)$stat.groupe
```

Vous devriez obtenir le résultat de sortie

	mean	sd	0%	25%	50%	75%	100%	n
G1	11.333333	1.527525	10	10.5	11	12.0	13	3
G2	5.428571	2.299068	2	4.0	5	7.5	8	7
G3	15.400000	1.516575	13	15.0	16	16.0	17	5

Ici, on indique d'abord la donnée quantitative (ou numérique) puis la qualitative (ou catégorielle).

On constate que les trois groupes sont assez hétérogènes.

### 6.2.3. La significativité pratique de la liaison

EXEMPLE 6.1. Avant de commencer à quantifier, il faut d'abord comprendre dans quelles situations on considère qu'une liaison est intense. Le graphique 6.3 montre quatre situations possibles avec deux groupes représentées par des collections de lignes de points empilés.

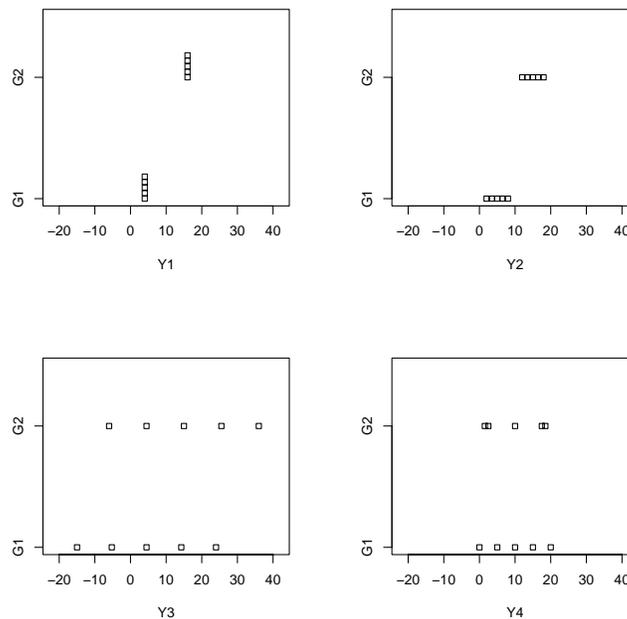


FIGURE 6.3. Quatre situations concernant deux groupes avec 5 mesures numériques pour chacun

- Dans la première situation, tous les éléments de la même catégorie ont exactement la même valeur numérique et ces valeurs diffèrent d'un groupe à l'autre. C'est la situation de relation parfaite. Lorsque l'on connaît le groupe, on peut dire exactement la valeur que va prendre un individu de ce groupe. On peut le formuler autrement en disant que la variabilité que l'on observe entre les valeurs (prises dans leur ensemble) est entièrement due aux différences entre les groupes.

- Dans la deuxième situation, les éléments d'un groupe n'ont pas tous la même valeur, mais on voit que ces groupes sont relativement homogènes (leur écart-type est petit) et qu'on peut assez bien différencier un groupe de l'autre (leurs moyennes sont différentes). La relation est forte. Pour reprendre la formulation précédente, la variabilité que l'on observe entre les valeurs provient largement de la différence entre les moyennes des groupes et dans une moindre mesure de la variabilité interne aux groupes.
- Dans la troisième situation, les éléments d'un groupe ont des valeurs assez différentes, en tout cas, on a plus de mal à distinguer les différences entre les groupes. La relation est faible. La variabilité entre les valeurs provient largement de la variabilité interne aux groupes.
- Dans la dernière situation, les éléments d'un groupe ont des valeurs différentes, mais surtout les moyennes des groupes ne permettent plus de les distinguer. La relation est nulle. La variabilité entre les valeurs est entièrement causée par la variabilité interne aux groupes et plus du tout par les différences entre les moyennes des groupes.

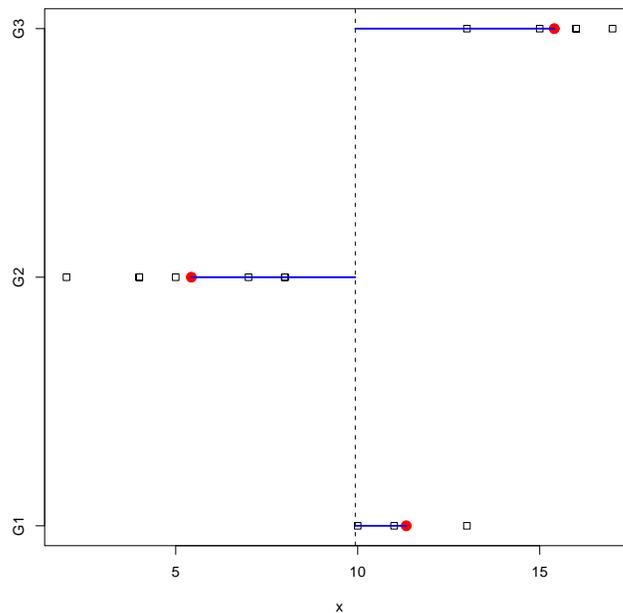


FIGURE 6.4. La collection de lignes de points des notes par groupes (données 'notes3TDbis.txt') et les moyennes des groupes.

Afin de bien visualiser la relation entre une variable quantitative et une variable qualitative, nous avons construit la représentation suivante des notes issues du fichier 'notes3TDbis.txt' en figure 6.4 avec

- Les groupes sont représentés en vertical, la variable quantitative en horizontal
- Un carré blanc représente un individu
- Les points rouges représentent les moyennes dans chaque groupe
- La ligne en pointillé représente la moyenne de l'ensemble des individus
- Les traits bleus représentent les écarts entre les moyennes des groupes et la moyenne de l'ensemble.

Ainsi, pour chaque note, on s'intéresse à déterminer si sa part dans la variabilité générale provenait plus des différences entre les groupes ou bien de la différence à l'intérieur du groupe auquel elle appartient.

Sur le graphique 6.4, on voit pour une donnée que sa part dans la variabilité générale est mesurable par sa distance à la moyenne générale (en bleu) dite distance totale. Cette distance est elle-même décomposable en

deux parties, une première partie qui mesure la différence de son groupe par rapport à l'ensemble et qui est la distance entre la moyenne du groupe et la moyenne générale, et une seconde partie qui est la distance entre la valeur et la moyenne de son groupe. En d'autres termes, pour chaque valeur  $y_{ij}$  qui appartient au  $j^{\text{ème}}$  groupe, on se demande si sa distance à la moyenne générale  $M$  soit  $y_{ij} - M$  provient plus de la distance interne à son groupe, c'est-à-dire à la moyenne  $M_j$  de son groupe, soit  $y_{ij} - M_j$ , ou bien de la distance de la moyenne de son groupe à la moyenne générale soit  $M_j - M$ . On peut écrire

$$y_{ij} - M = (y_{ij} - M_j) + (M_j - M).$$

Or, cette relation reste vraie lorsque ses éléments sont mis au carré et ajoutés (c'est une forme du théorème de Pythagore). C'est ce qu'on appelle la décomposition de la somme des carrés

$$SC_{\text{Totale}} = SC_{\text{Inter-Groupes}} + SC_{\text{Intra-Groupes}},$$

où  $SC_{\text{Totale}}$  est la somme totale des carrés,  $SC_{\text{Inter-Groupes}}$  est la variation inter-groupes et  $SC_{\text{Intra-Groupes}}$  est la variation intra-groupes. Si on fait le lien avec ce qui a été dit précédemment, on peut alors quantifier la force de la relation suivant l'importance que prend la somme des carrés inter-groupes par rapport à la somme des carrés totale. Plus sa part est grande, plus les deux variables sont reliées, car la plus grande partie de la variabilité vient des différences entre les groupes et non pas des différences internes aux groupes.

On a en fait avec  $n$  unités statistiques au total et  $n_j$  unités dans le groupe  $j$  ( $j = 1, \dots, G$ ) :

$$\underbrace{\sum_{j=1}^G \sum_{i=1}^{n_j} (y_{ij} - M)^2}_{SC_{\text{Totale}}} = \underbrace{\sum_{j=1}^G n_j (M_j - M)^2}_{SC_{\text{Inter-Groupes}}} + \underbrace{\sum_{j=1}^G \sum_{i=1}^{n_j} (y_{ij} - M_j)^2}_{SC_{\text{Intra-Groupes}}}. \quad (6.1)$$

◇

DÉFINITION 6.2. Le *rapport de corrélation* donne la somme des carrés (de la variable numérique) qui est expliquée par la prise en compte de la variable catégorielle, c'est-à-dire le rapport de la somme des carrés inter-groupes sur la somme des carrés totale :

$$RC = \frac{SC_{\text{Inter-Groupes}}}{SC_{\text{Totale}}} \quad (6.2)$$

Il est toujours compris entre 0 et 1. Proche de zéro, les deux variables sont peu reliées, proche de 1 elles le sont fortement.

EXEMPLE 6.3. Donnons les différentes valeurs des RC pour l'exemple 6.1 page 70 qui sont dans l'ordre des graphiques de la figure 6.3 page 70

$$\begin{aligned} RC_1 &= 1, \\ RC_2 &= 0.847458, \\ RC_3 &= 0.118357, \\ RC_4 &= 0. \end{aligned}$$

Ainsi, les quatre graphiques montrent l'exemple de variables allant successivement d'une situation très fortement liée à une situation non liée.

En fait, grâce à (6.1), on a l'expression exacte du RC :

$$RC = \frac{\sum_{j=1}^G n_j (M_j - M)^2}{\sum_{j=1}^G \sum_{i=1}^{n_j} (y_{ij} - M)^2}. \quad (6.3)$$

◇

Pour calculer cette quantité,  
Il suffit de taper

```
summary(lm(notes3TDbis$notes~notes3TDbis$groupes))
```

et on obtient

Call:

```
lm(formula = notes[, ind.travail[1]] ~ notes[, ind.travail[2]])
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-3.4286 -1.3810 -0.3333  1.5857  2.5714
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.333      1.125  10.072 3.31e-07 ***
notes[, ind.travail[2]]G2  -5.905      1.345  -4.390 0.00088 ***
notes[, ind.travail[2]]G3   4.067      1.423   2.857 0.01443 *
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.949 on 12 degrees of freedom

Multiple R-squared: 0.8671, Adjusted R-squared: 0.8449

F-statistic: 39.14 on 2 and 12 DF, p-value: 5.514e-06

Il ne faut pas paniquer, seules les deux dernières lignes nous intéressent dans la fenêtre de sortie. Le rapport de corrélation se lit en face de la mention Multiple R-Squared, il est donc égal à  $RC=0.8671$  Il est très petit. On en déduit que la relation est faible entre la couleur de cheveux et la note de statistique.

On pourra aussi utiliser la fonction `determin.qualiquanti`, disponible à l'URL habituelle et taper

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)$RC
```

Ici, on indique d'abord la donnée quantitative (ou numérique) puis la qualitative (ou catégorielle).

La somme des carrés est proportionnelle à un coefficient près à la variance (qui est, elle-même, le carré de l'écart-type). On peut donc interpréter le rapport de corrélation comme un pourcentage de variabilité expliquée.

On trouve sur l'exemple des notes des étudiantes

$$RC = 0.867085 = 86.71\%$$

ce qui signifie que la prise en compte du groupe permet d'expliquer 86.71 % de la variabilité de la note de statistique. Pour le dire autrement, les variables notes et groupes sont fortement reliées.

Comment savoir à partir de quel seuil pourra-t-on déclarer qu'il y a relation ?

Nous allons voir deux façons de répondre à cette question : la *significativité pratique* et la *significativité statistique* (dans la section 6.2.4).

En ce qui concerne la significativité pratique, une grille d'interprétation qualitative a été proposée par Cohen [16] en considérant le rapport de corrélation comme une *taille d'effet* (*effect size*) : il introduit trois seuils  $RC_1 = 0.01$ ,  $RC_2 = 0.05$  et  $RC_3 = 0.15$  tels que

$$\text{si } RC \begin{cases} < RC_1, & \text{la significativité pratique de la liaison est faible,} \\ \in [RC_1, RC_2[, & \text{la significativité pratique de la liaison est moyenne,} \\ \in [RC_2, RC_3[, & \text{la significativité pratique de la liaison est forte,} \\ > RC_3, & \text{la significativité pratique de la liaison est très forte} \end{cases} \quad (6.4)$$

### 6.2.4. La significativité statistique de la liaison

Comme dans la section 4.4, on introduit une probabilité critique  $p_c$ , comprise entre 0 et 1. Proche de zéro (inférieure ou égale à  $0.05 = 5\%$ , valeur traditionnellement choisie) elle indique une relation statistiquement significative, c'est-à-dire qui a peu de chance d'être due au hasard. En revanche, strictement supérieure à 0.05, elle indique que la relation n'est pas statistiquement significative donc qu'elle peut-être due au hasard.

Pour la calculer,  
Il suffit de taper

```
summary(lm(notes3TDbis$notes~notes3TDbis$groupes))
```

et on obtient comme ci-dessous.

Dans le cas du fichier 'notes3TDbis.txt', on obtient donc

$$p_c = 5.51369e - 06 = 0.00055\%, \quad (6.5)$$

donc ici statistiquement significative.

Dans les deux cas, on pourra aussi utiliser la fonction `determin.qualiquanti` et taper

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)$pc
```

### 6.3. Calculer tous les indicateurs

On peut aussi utiliser la fonction `determin.qualiquanti.R` disponible sur le site et qui fournit directement les valeurs des statistiques par groupes, du  $RC$  et  $p_c$  :

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes)
```

```
$RC
```

```
[1] 0.8670851
```

```
$pc
```

```
[1] 5.513688e-06
```

```
$stat.groupe
```

	mean	sd	0%	25%	50%	75%	100%	n
G1	11.333333	1.527525	10	10.5	11	12.0	13	3
G2	5.428571	2.299068	2	4.0	5	7.5	8	7
G3	15.400000	1.516575	13	15.0	16	16.0	17	5

renvoie bien les valeurs indiquées précédemment. On pourra aussi obtenir directement les collections de boîtes à moustaches et de lignes de points en tapant

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes,fig=T)
```

ou encore

```
determin.qualiquanti(notes3TDbis$notes,notes3TDbis$groupes,fig=T,labelX="notes",labelgpe="groupes")
```

ce qui donne la figure 6.5 page suivante.

### 6.4. Quelques exercices

EXERCICE 6.4.

À l'occasion d'une surveillance d'examen, un enseignant a décidé de relever la couleur des cheveux des étudiantes en L3 Management des Organisations Sportives et, un peu plus tard, en corrigeant leurs copies de statistique à l'aveugle, a conservé leurs notes, ce qui donne le tableau suivant 6.1.

Voir le fichier 'blondes.txt'.

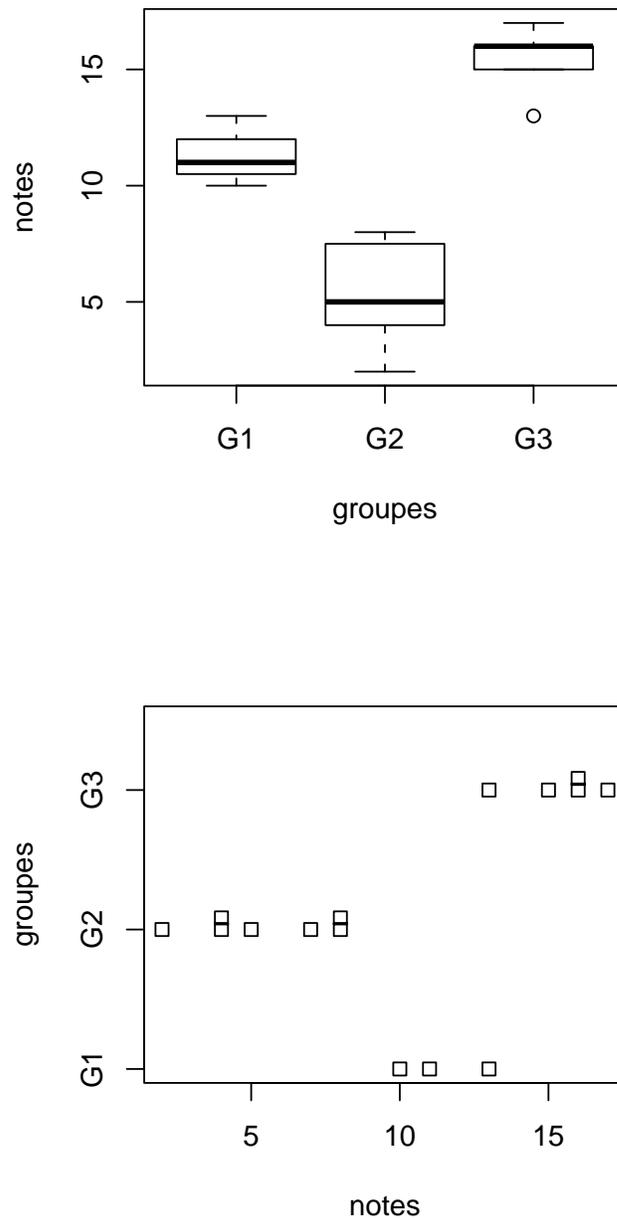


FIGURE 6.5. Les collections de boites de dispersion et des lignes de points des notes par couleurs de cheveux (données 'notes3TDbis.txt')

Est-ce que la note de statistique est reliée à la couleur des cheveux ?  
 Voir éléments de correction page 76

	blonde	brune
1	14	4
2	6.5	4
3	11	17.5
4	10	12.5
5	7.5	7.5
6	13.5	8
7	14.5	13
8		12

TABLE 6.1. Notes de statistique selon la couleur de cheveux des étudiantes

EXERCICE 6.5 (facultatif). Le fichier `'notes3TDter.txt'` contient *exactement les mêmes données* que son homologue `'notes3TDbis.txt'`, *excepté les fait que les groupes sont appelés '1', '2' et '3', au lieu de 'G1', 'G2' et 'G3'!*

Décrire graphiquement et numériquement les groupes de ce fichier.

Voir éléments de correction page 78

EXERCICE 6.6 (facultatif). L'entraînement intensif conduit à des perturbations physiologiques chez les sportifs de haut niveau. Cela est bien connu chez les femmes avec des dysfonctionnements de leur cycle mensuel. Peut-on également constater un impact de l'entraînement sur les fonctions hormonales et reproductives des hommes ?

Après prélèvement de sang veineux, Ayers *et al.* évaluent la testostérone totale pour vingt coureurs à pieds, parcourant au moins 45 kilomètres par semaine, et un groupe de contrôle de dix individus, d'âges similaires, en bonne santé

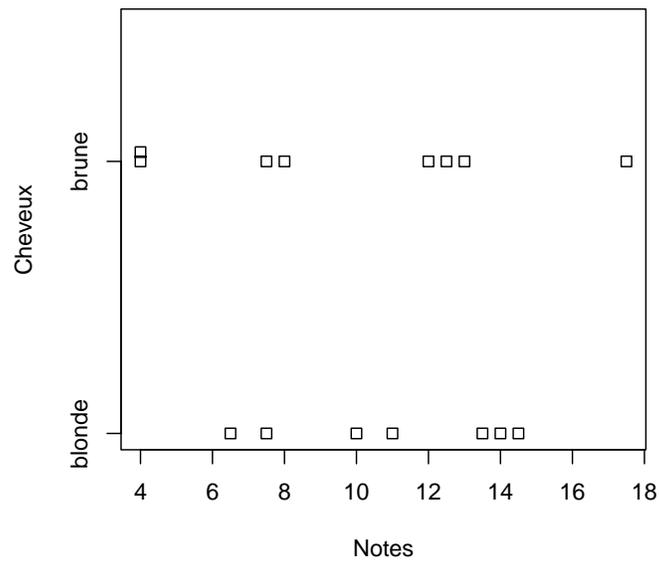
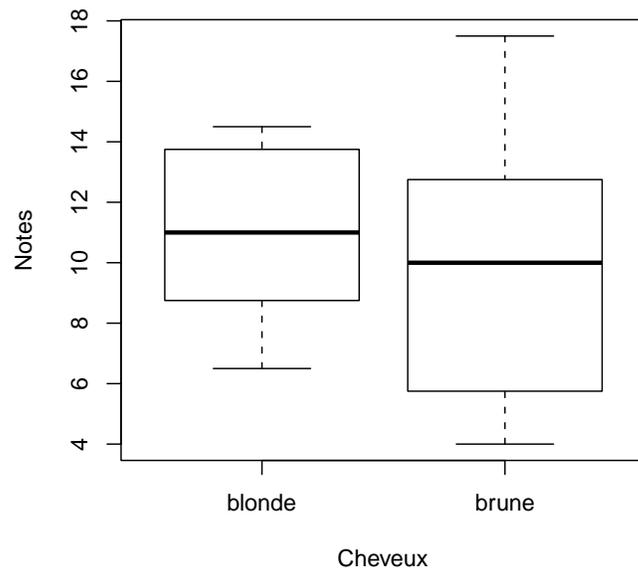
Les données sont réunies dans le fichier `TESTOSTERONE.txt`, la variable `'Taux'` correspond à la mesure de testostérone et la variable `'Sujets|'` permet de distinguer les deux groupes. Décrire graphiquement et numériquement les deux groupes. Qu'en pensez-vous ?

Voir éléments de correction page 79

## 6.5. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.4

- On étudie le croisement de la variable quantitative (ou numérique) `'Notes'` et de la variable qualitative (ou catégorielle) `'Cheveux'`.
-



Voir la figure ci-dessous.

- Avec , on obtient les statistiques par groupes données dans le tableau suivant ;

	moyenne	écart-type	0%	25%	50%	75%	100%	n
blonde	11.00	3.19	6.50	8.75	11.00	13.75	14.50	7
brune	9.81	4.74	4.00	6.62	10.00	12.62	17.50	8

On rappelle que, dans ce tableau :

- le nombre noté 0% est le quartile à 0 % (c'est le minimum) ;
- le nombre noté 25% est le quartile à 25 % (c'est  $Q_1$ ) ;
- le nombre noté 50% est le quartile à 50 % (c'est la médiane) ;
- le nombre noté 75% est le quartile à 75 % (c'est  $Q_3$ ) ;
- le nombre noté 100% est le quartile à 100 % (c'est le maximum).

Commençons par décrire la centralité : dans chaque groupe, il semble que le centre soit proche de 10, il y a peu de différences. En ce qui concerne la dispersion, le groupe des jeunes femmes brunes paraît avoir plus d'hétérogénéité. Enfin, il ne semble pas y avoir de valeurs extrêmes (il est vrai qu'avec des notes de 0 à 20, c'est peu probable). On constate, grâce aux statistiques par groupes que les moyennes (et les médianes) sont proches et qu'elles sont plutôt plus élevées chez les étudiantes blondes. Les statistiques de dispersion montrent que les étudiantes brunes sont un peu plus hétérogènes.

Confirmons cela grâce à  $\mathcal{R}$ .

Les autres résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.023531
probabilité critique $p_c$	0.585196

On compare le rapport de corrélation  $RC=0.023531$  aux seuils de Cohen (0.01,0.05,0.15) (voir [16]) et la probabilité critique  $p_c=0.585196$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	<b>moyenne</b>
significativité statistique	<b>non</b>

- On peut donc affirmer qu'il existe peu de relation entre les variables 'Notes' et 'Cheveux'.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.5

*Attention*, si on tape sans prendre de précaution

```
determin.qualiquanti(notes3TDter$notes,notes3TDter$groupes)
```

la probabilité critique vaudrait

$$p_c = 0.09300072, \quad (6.6)$$

à comparer avec (6.5).

En fait, ici  $\mathcal{R}$  considère la deuxième variable (dont les valeurs dont '1',...) comme numérique et la probabilité critique renvoyée est celle du croisement de deux variables numériques! En effet, si on utilise la fonction `determin.quantiquanti` du chapitre 4 et que l'on tape

```
determin.quantiquanti(notes3TDter$notes,notes3TDter$groupes)$pc
```

on obtient  $p_c = 0.09300072$ , ce qui est bien la valeur donnée par (6.6)!

Pour palier ce problème, regardez ce qui se passe si on tape

```
notes3TD$groupes
is.factor(notes3TDter$groupes)
as.factor(notes3TDter$groupes)
is.factor(as.factor(notes3TDter$groupes))
```

Il faudra donc taper

```
determin.qualiquanti(notes3TDter$notes,as.factor(notes3TDter$groupes))
```

La probabilité critique vaut alors bien  $5.513688e-06$ , comme donné par (6.5).

En revanche, il n'y a pas de problème pour la figure, le RC ou les statistiques par groupes.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 6.6

Comme décrit précédemment, on peut tracer pour les données 'TESTOSTERONE.txt' une collection de ligne de points et une collection de boîte de dispersion (voir figure 6.6).

Les statistiques par groupes fournissent :

```

      mean      sd 0%   25% 50%   75% 100%  n
athlète  4.20  2.110375 0.9  2.775 3.9  5.700  9.0 20
contrôle 6.94  1.075174 5.4  6.175 6.9  7.375  8.9 10

```

*A priori*, au vu des graphiques 6.6 et des résultats précédent, il y a une forte dépendance de la note par rapport au groupe : deux moyennes très différentes selon les groupes et des écart-types différents.

Calculons maintenant le rapport de corrélation et la probabilité critique en tapant

```
determin.qualiquanti(TESTOSTERONE$Taux,TESTOSTERONE$Statut)
```

```
$RC
```

```
[1] 0.3449994
```

```
$pc
```

```
[1] 0.0006437071
```

```
$stat.groupe
```

```

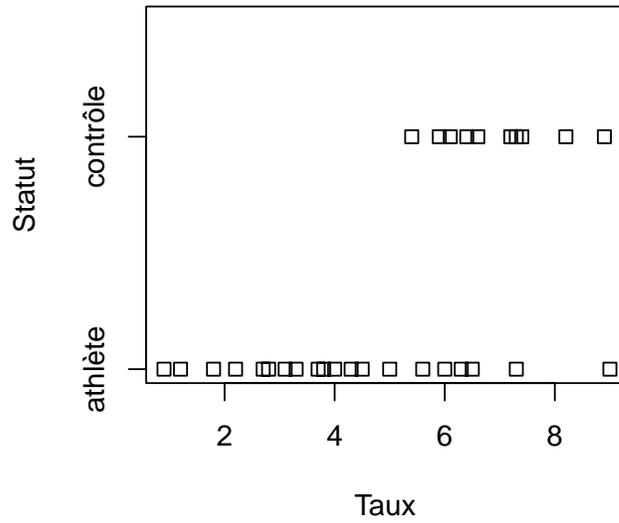
      mean      sd 0%   25% 50%   75% 100%  n
athlète  4.20  2.110375 0.9  2.775 3.9  5.700  9.0 20
contrôle 6.94  1.075174 5.4  6.175 6.9  7.375  8.9 10

```

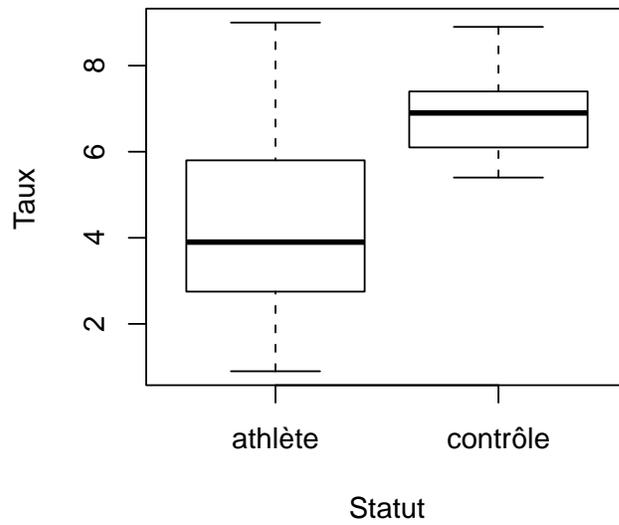
On obtient donc

$$RC = 0.344999, \quad p_c = 0.000643707.$$

Ainsi, les variables très sont fortement liées, ce qui confirme notre observation faite *a priori*. La probabilité critique nous indique que la liaison est statistiquement significative.



(a) : lignes de points



(b) : boîtes de dispersion

FIGURE 6.6. Deux collections pour les données 'TESTOSTERONE.txt'.

## Créer ses propres fichiers de données

Ce chapitre vous montre comment créer vos propres fichiers de données dans un format adapté et lisible par .

prenom	age	taille	masse	sport
jean		1,85	85	foot
pierre	25	2,01	80	Gymnastique aquatique
paul	32	NA	75	foot
jean-jacques	12	1,78	69	pétanque

TABLE 7.1. Quelques données

Considérons les données très simples du tableau 7.1. Les données manquantes correspondent à des cases vides ou contenant la chaîne de caractère "NA".

Pour créer un fichier de données lisible par  sous la forme d'un data frame : il faut :

- (1) lancer le tableur libre et gratuit OpenOffice;
- (2) choisir OpenOffice.org Calc;
- (3) dans la feuille de calcul, saisir les données du tableau 7.1;
- (4) enregistrer ensuite au format OpenOffice : le nom du fichier aura pour extension `ods`. Choisir, par exemple `dudu.ods`. Vous pouvez alors modifier le tableau, l'ouvrir de nouveau, le modifier ... et toujours avec OpenOffice.

REMARQUE 7.1. Ces manipulations peuvent aussi être faites avec excel, la sauvegarde se faisant alors au format `xls`.

REMARQUE 7.2. OpenOffice est téléchargeable sur internet! (<http://fr.openoffice.org/>).

Une fois que le fichier est prêt, il faut alors l'enregistrer sous un format que  peut lire; car malheureusement, il ne peut lire le format OpenOffice! Il faut procéder ainsi : quand le fichier est ouvert dans OpenOffice, il faut alors

- (1) "l'enregistrer sous" au format CSV (d'extension `csv`);
- (2) Il faut répondre "choisir le format actuel" (et non "au format ODS");
- (3) Il faut ensuite choisir en cliquant sur le séparateur de champ ";".

On peut ensuite ouvrir le fichier avec  :

On pourra taper, par exemple,

```
dudu <- read.csv("dudu.csv", header = TRUE, sep = ";", dec = ",", na.strings = c("NA",
""))
```

et travailler ensuite sur le data frame `dudu`, comme d'habitude. Taper par exemple

```
dudu  
dudu$sport
```

REMARQUE 7.3. On peut aussi faire au export au format `xls`, mais `R` ne gère pas toujours ce format correctement.

REMARQUE 7.4. Il existe des tas de solutions différents : exports au format texte, csv, excel... Seule une méthode est ici présentée.

## Une introduction à la statistique univariée. Les représentations graphiques avec $\mathbb{R}$

Ce chapitre s’inspire fortement des documents [18, 19]. Son objectif est de réviser et de poursuivre l’acquisition des notions de base de la statistique descriptive univariée tant sur le plan des paramètres descriptifs que sur celui des représentations graphiques.

*Pour des raisons de lisibilité, toutes les figures de ce chapitre sont données à partir de la page 89.*

### 8.1. Introduction

Afin de gérer proprement votre travail, nous vous rapelons les conseils ci-dessous.

- (1) Songez toujours à la documentation (l’aide) de  $\mathbb{R}$  (voir section 2.3.2 page 15).
- (2) Lorsque nous ouvrons un data frame fichier `fichier` sous  $\mathbb{R}$ , ce dernier est constitué de variables qui lui sont attachées. Nous avons vu deux possibilités pour tenir compte de cette information par exemple :
  - `names(fichier)`  
Chacune des variables de ce data frame peut être lue avec `names(fichier$var)`
  - ou `attach(fichier)`  
`var`  
N’oubliez pas en fin de travail `detach(fichier)`
- (3) Dans cette découverte du logiciel, nous avons vu qu’il y avait acquisition d’un vocabulaire statistique et d’un vocabulaire spécifique à R. Nous vous conseillons d’ouvrir en parallèle un document word (ou tout autre format) et de vous créer votre propre cours. Vous pouvez ainsi noter les commandes, coller les graphiques, insérer des commentaires personnels.
- (4) Nous avons déjà vu un certain nombre de fonctions graphiques :
  - `pie` et `barplot`, voir exercice 3.1 page 26;
  - `dotchart`, voir exercice 3.2 page 27; voir
  - `hist` et `stripchart`, voir exercice 3.5 page 28;
  - `boxplot`, voir exercice 3.7 page 29.

### 8.2. Étude de variables qualitative

Pour représenter une variable qualitative, on peut utiliser soit un diagramme en secteur (ou camembert) (commande `pie`), soit une représentation en batons (commande `barplot`).

#### EXERCICE 8.1.

Considérons le jeu de données portant sur 592 étudiants (extrait de [17]) Pour chaque étudiant on a observé 3 variables qualitatives : la couleur des cheveux, la couleur des yeux et le sexe. Les données se trouvent dans le fichier "qualitatif.txt", que vous pouvez télécharger à l’URL habituelle.

- (1) (a) Importez le fichier dans R à l'aide de la commande `read.table`. Appelez le fichier `qualitatif`.  
 (b) Représentez les données sur la couleur des cheveux sous la forme d'un diagramme en secteurs en tapant la commande suivante :

```
pie(table(qualitatif$cheveux), col = c("yellow", "chocolate4", "black",
  "orange"), main = "Couleur des cheveux de 592 etudiants")
```

Quelle est la couleur de cheveux dominante? Dans quel ordre peut-on classer les couleurs? Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir?

- (c) La représentation en secteurs n'est pas la représentation optimale; il faut s'en méfier! Consultez en effet la documentation de la fonction `pie` à l'aide de la commande

```
help(pie)
```

- (d) La commande `dotchart` permet d'obtenir un graphique plus lisible :

```
dotchart(sort(table(qualitatif$cheveux)), xlim = c(0, max(table(qualitatif$cheveux))),
  pch = 20, cex = 1.5, color = c("orangered", "black", "yellow",
  "chocolate4"), main = "Couleur des cheveux de 592 etudiants")
```

Peut-on maintenant répondre à la question : "Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir?"

- (e) Utilisez la commande `barplot` pour représenter la variable `qualitatif$cheveux`.  
 (2) Le graphe de Cleveland (commande `dotchart`) donne une représentation agréable pour les variables qualitatives avec de nombreuses modalités de même que pour les variables qualitatives ordonnées.  
 (a) Utilisez les commandes `summary` et `dotchart` pour représenter la variable `sport` du fichier `L3APA06.txt`.  
 (b) Représentez la variable `mention` du fichier `L3APA06.txt` grâce à un graphe de Cleveland et un diagramme en baton.

Voir les éléments de correction page 85.

### 8.3. Étude de variables quantitatives

Pour représenter une variable quantitative, on peut utiliser soit un histogramme (commande `hist`), soit une boîte à moustaches pour les variables continues (commande `boxplot`), soit un diagramme en bâtons pour les variables discrètes (commande `plot(table)`).

EXERCICE 8.2.

- (1) Voici le poids (en ordre croissant) de 10 marathoniens : 61 62 67 67 68 69 76 77 78 79.  
 Donnez une représentation sous forme d'histogramme de ces données.  
 (2) Notez, en utilisant le `help(hist)` que différents arguments de la fonction permettent de réaliser le découpage en classes :  
 – `breaks` donne les valeurs de scission.  
 – `include.lowest = TRUE` signifie que la valeur la plus petite est incluse dans la première classe.  
 – `right = TRUE` signifie que les intervalles sont ouverts à gauche et fermés à droite.  
 – `freq` est un logique. S'il est égal à `TRUE` l'histogramme indique les fréquences, c'est-à-dire le nombre de chaque éléments dans une classe; s'il est égal à `FALSE`, les densités sont représentées, c'est-à-dire le rapport de la fréquence par le produit de la largeur de la classe et du nombre total d'individus (de telle sorte que l'histogramme soit d'aire totale égale à 1).

Réalisez les histogrammes en choisissant respectivement les séries d'intervalles suivants :

- (a) les intervalles  $[50, 60]$ ,  $]60, 70]$  et  $]70, 80]$  ;
  - (b) les intervalles  $[55, 65]$ ,  $]65, 75]$ , et  $]75, 85]$ .
- (3) Que peut-on dire de l'allure de ces 2 figures ?
- (4) Étudions maintenant le poids des étudiants du fichier de données `L3APA06.txt`. Le sexe est une variable qu'on ne saurait exclure de l'étude. D'une manière générale, toute étude portant sur la morphologie doit tenir compte du dimorphisme sexuel.
- (a) Construire les data frames `sexeM` et `sexeF` qui contiennent respectivement les réponses des élèves de chacun des sexes grâce à
 

```
sexeM <- L3APA06[L3APA06$sexe == "masculin", ]
sexeF <- L3APA06[L3APA06$sexe == "féminin", ]
```
  - (b) Construire l'histogramme du poids des élèves de chaque sexe.
  - (c) Afin de pouvoir comparer le poids selon les 2 sexes, imposez les mêmes classes grâce à la commande `breaks`, représentez les densités et imposez les mêmes ordonnées extrémales.
  - (d) Utilisez la commande `boxplot(poids~sexe)` pour avoir une représentation sous la forme d'une collection de boîtes à moustaches

Voir les éléments de correction page 87.

## 8.4. Pour aller plus loin

REMARQUE 8.3 (Variables quantitatives). Lorsque le data frame contient des données manquantes symbolisées par `NA`, les calculs et les graphiques sont réalisés bien sûr sans tenir compte de ces dernières. Le nombre de données manquantes apparaît dans le `summary`.

REMARQUE 8.4 (Variables qualitatives). Lorsque le data frame contient des données manquantes symbolisées par `NA`, les calculs et les graphiques considèrent celui-ci comme une modalité. C'est intéressant pour repérer le nombre de `NA`. Si vous voulez enlever les données manquantes, vous pouvez adopter deux solutions : `summary(na.omit())` ou `table()`.

REMARQUE 8.5 (Variables qualitatives). Il serait très long et fastidieux de séparer les variables les unes après les autres en fonction de critères qualitatifs. Il existe une commande qui permet de séparer directement un data frame en deux sous data frames : dans le cas de la question 4 de l'exercice 8.2, regardez ce que donne :

```
neofich <- split(L3APA06, L3APA06$sexe)
class(neofich)
names(neofich)
neofich$masculin
neofich$féminin
```

## 8.5. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 8.1

- (1) (a)
- (b) On obtient la figure 8.1 page 89.  
On peut répondre à quelques-unes des questions posées :
  - "Quelle est la couleur de cheveux dominante ? " : c'est le "maron".

- "Dans quel ordre peut-on classer les couleurs?" : à part la couleur dominante pas de classement possible.
- "Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir?" : ce n'est pas très clair sur le graphique.

(c)

(d) La commande `dotchart` permet d'obtenir un graphique plus lisible. Voir figure 8.2 page 90. On peut maintenant répondre à la question "Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir?"

(e) Voir figure 8.3 page 90, obtenue à partir de la ligne de commande :

```
barplot(sort(table(qualnom$cheveux)))
```

(2) (a) On a obtenu la figure 8.4 page 91 grâce à la ligne de commande :

```
dotchart(sort(table(L3APA06$sport)))
```

(b) On a obtenu la figure 8.5 page 91 grâce aux lignes de commande :

```
par(mfrow = c(1, 2))
barplot(summary(L3APA06$mention))
dotchart(summary((L3APA06$mention)), pch = 15)
```

Nous pouvons émettre deux constats. La modalité très bien n'est pas présente. Les modalités sont rangées par ordre alphabétique et non ordonnées.

```
summary(L3APA06$mention)
```

```
AB   B   P NA's
11   1  40   6
```

Nous constatons qu'aucun étudiant n'a eu la mention très bien, ce qui explique son absence dans les modalités de la variable `mention`. Remarquons que là aussi il y a des données manquantes : les étudiants ignorent que le fait d'avoir le baccalauréat avec une moyenne comprise entre 10 et 12 correspond à la mention passable. Pour faire apparaître la mention très bien (avec aucun effectif), il faut "forcer" les modalités :

```
mention2 <- L3APA06$mention
levels(mention2) <- c("P", "AB", "B", "TB")
levels(mention2)
```

```
[1] "P" "AB" "B" "TB"
```

```
summary(mention2)
```

```
 P  AB  B  TB NA's
11  1  40  0   6
```

```
summary(na.omit(mention2))
```

```
 P AB  B TB
11 1 40 0
```

Attention aux calculs des fréquences relatives dans le cas des données manquantes.

```
summary(na.omit(mention2))/length(mention2)
      P      AB      B      TB
0.18965517 0.01724138 0.68965517 0.00000000

sum(summary(na.omit(mention2))/length(mention2))

[1] 0.8965517

summary(na.omit(mention2))/length(na.omit(mention2))
      P      AB      B      TB
0.21153846 0.01923077 0.76923077 0.00000000

sum(summary(na.omit(mention2))/length(na.omit(mention2)))

[1] 1
```

On peut alors tracer les graphiques de la figure 8.6 page 92 grâce à

```
par(mfrow = c(1, 2))
barplot(summary(na.omit(mention2)))
dotchart(summary(na.omit(mention2))), pch = 15)
```

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 8.2

- (1) Pour tracer l'histogramme de la figure 8.7 page 92, on a tapé

```
marathon <- c(61, 62, 67, 67, 68, 69, 76, 77, 78, 79)
hist(marathon, col = "grey")
```
- (2) Pour tracer l'histogramme de la figure 8.8 page 93, on a tapé

```
par(mfrow = c(1, 2))
hist(marathon, col = "grey", breaks = c(50, 60, 70, 80), include.lowest = TRUE,
     right = TRUE)
hist(marathon, col = "grey", breaks = c(55, 65, 75, 85), include.lowest = TRUE,
     right = TRUE)
```
- (3) On ne peut pas bien comparer les aires car l'échelle n'est pas la même. On va donc imposer l'échelle et représenter les densités en tapant :

```
par(mfrow = c(1, 2))
hist(marathon, col = "grey", breaks = seq(from = 50, to = 80, by = 10),
     include.lowest = TRUE, right = TRUE, ylim = c(0, 0.06), freq = F)
hist(marathon, col = "grey", breaks = seq(from = 55, to = 85, by = 10),
     include.lowest = TRUE, right = TRUE, , ylim = c(0, 0.06), freq = F)
```

Voir figure 8.9 page 93. Le choix du découpage en intervalles est un problème délicat qui risque de biaiser fortement notre perception des données.

- (4) (a)
- (b) Pour tracer l'histogramme de la figure 8.10 page 93, on a tapé

```
par(mfrow = c(1, 2))
hist(sexeM$poids, col = "grey", main = "Poids des Hommes")
hist(sexeF$poids, col = "grey", main = "Poids des Femmes")
```

(c) Pour tracer l'histogramme de la figure 8.11 page 94, on a tapé

```
par(mfrow = c(1, 2))
hist(sexeM$poids, freq = F, col = grey(0.8), ylim = c(0, 0.1), breaks = seq(45,
  90, by = 5), main = "Poids des hommes")
hist(sexeF$poids, freq = F, col = grey(0.8), ylim = c(0, 0.1), breaks = seq(45,
  90, by = 5), main = "Poids des femmes")
```

(d) Pour tracer la boîte à moustache de la figure 8.12 page 94, on a tapé

```
boxplot(L3APA06$poids ~ L3APA06$sexe, cex.main = 1, main = "Poids des étudiants en fonction du sexe")
```

## 8.6. Ensemble des figures

### Couleur des cheveux de 592 étudiants

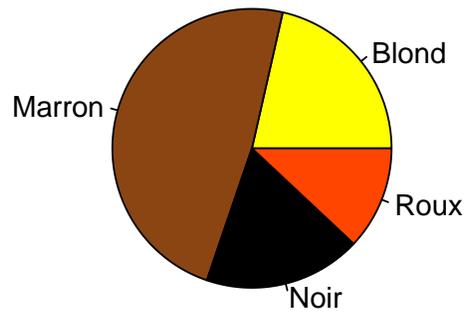


FIGURE 8.1. Le camembert des couleurs de cheveux (données qualitatif).

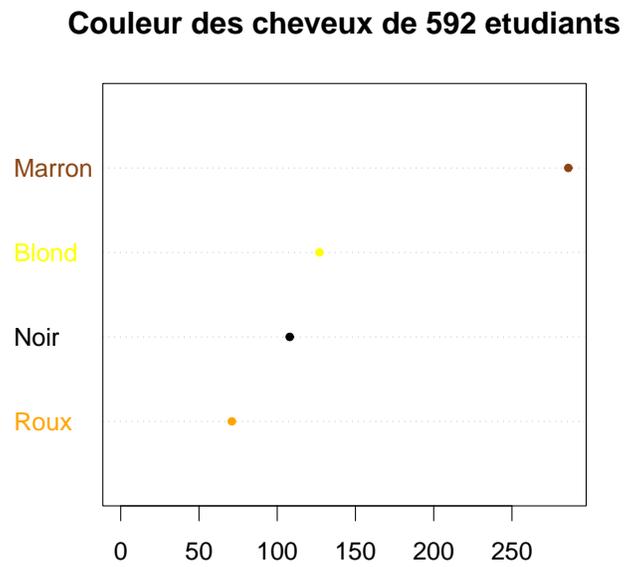


FIGURE 8.2. La ligne de points des couleurs de cheveux (données qualitatif).

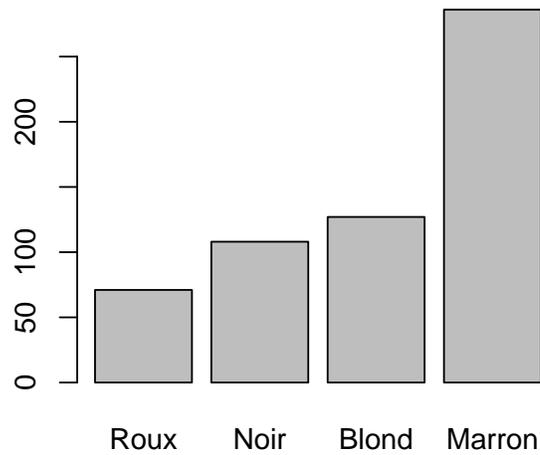


FIGURE 8.3. Le diagramme en barre des couleurs de cheveux (données qualitatif).

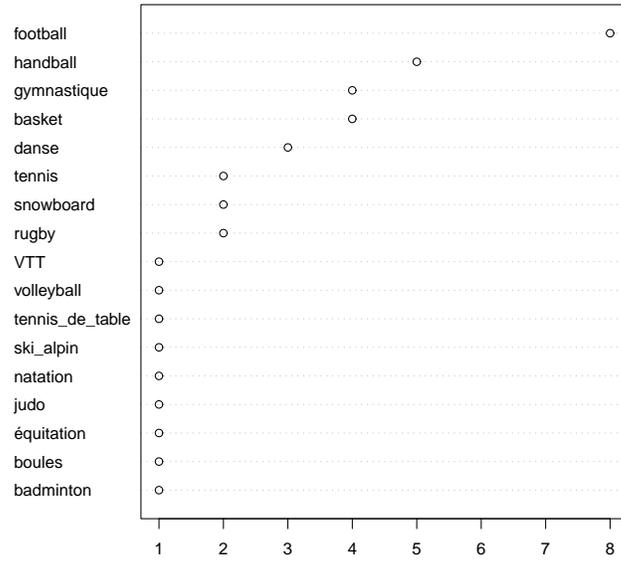


FIGURE 8.4. La ligne de points des sports (données L3APA06.txt).

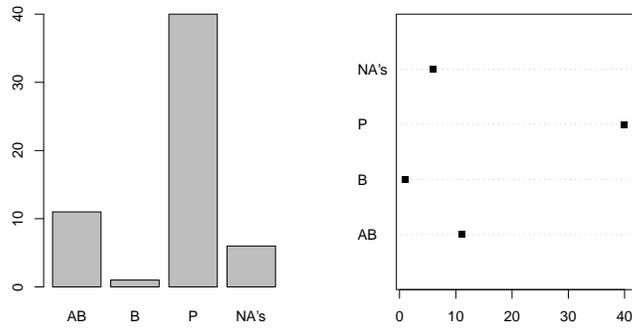


FIGURE 8.5. Le diagramme en baton et le diagramme de Cleveland des mentions (données L3APA06.txt).

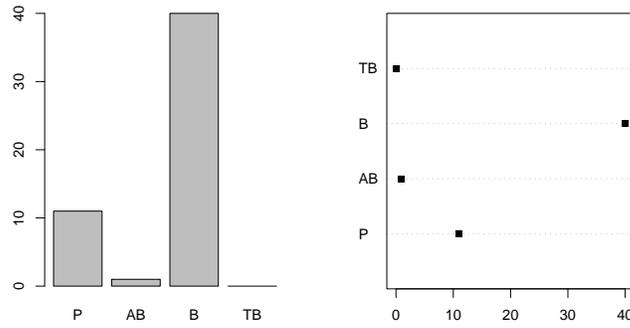


FIGURE 8.6. Le diagramme en baton et le diagramme de Cleveland des mentions avec l'apparition de la mention TB et la suppression des données manquantes (données L3APA06.txt).

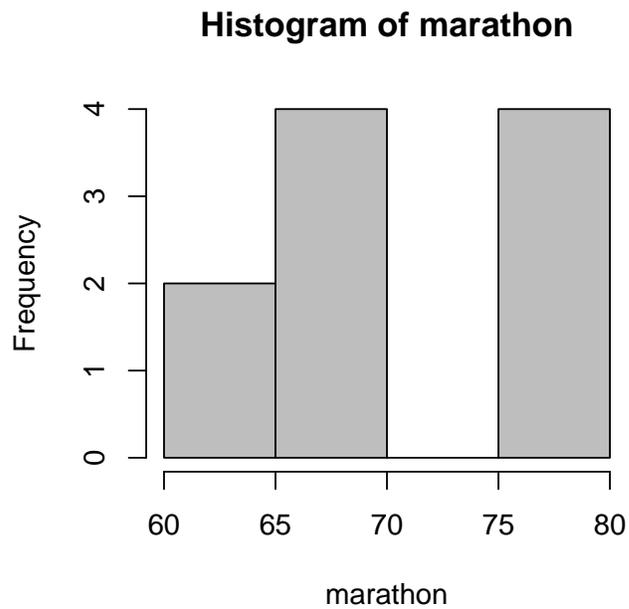


FIGURE 8.7. L'histogramme des 10 marathoniens.

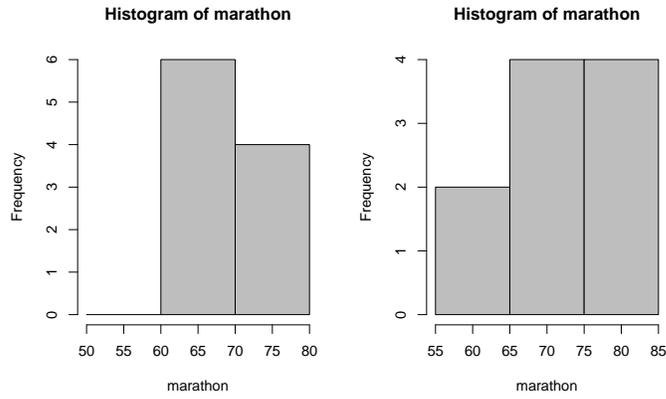


FIGURE 8.8. L'histogramme des 10 marathoniens avec classe imposées.

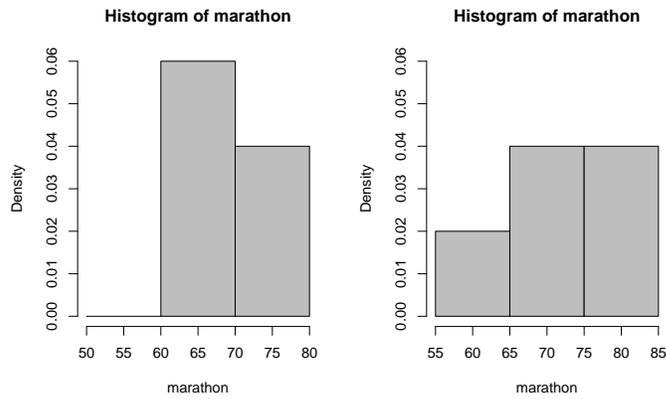
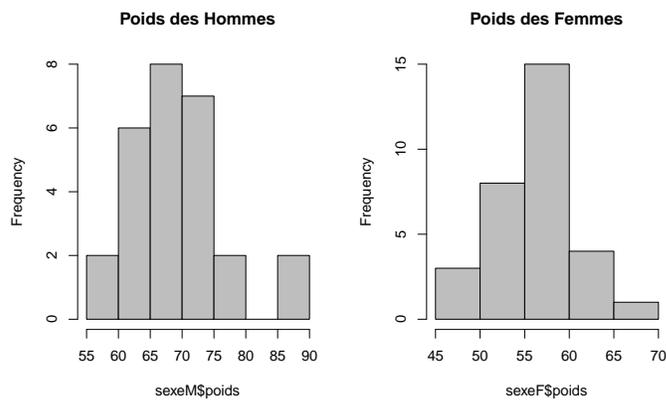
FIGURE 8.9. L'histogramme des 10 marathoniens avec classe imposées en densité avec  $y_{\max} = 0.06$ .

FIGURE 8.10. Histogrammes des poids des hommes et des femmes pour les données L3APA06

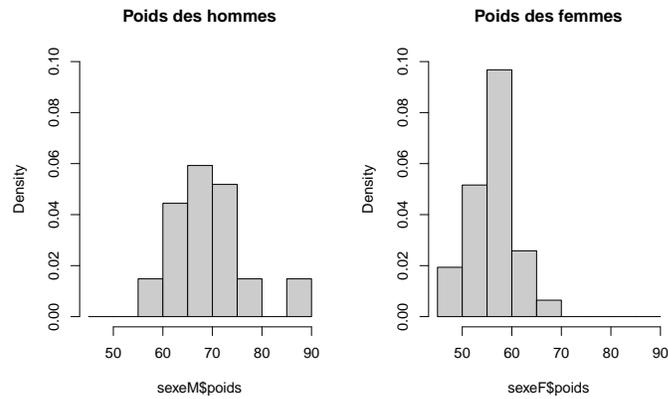


FIGURE 8.11. Histogrammes des poids des hommes et des femmes pour les données L3APA06 avec classes imposées, en densité avec ordonnées extrémales imposées.

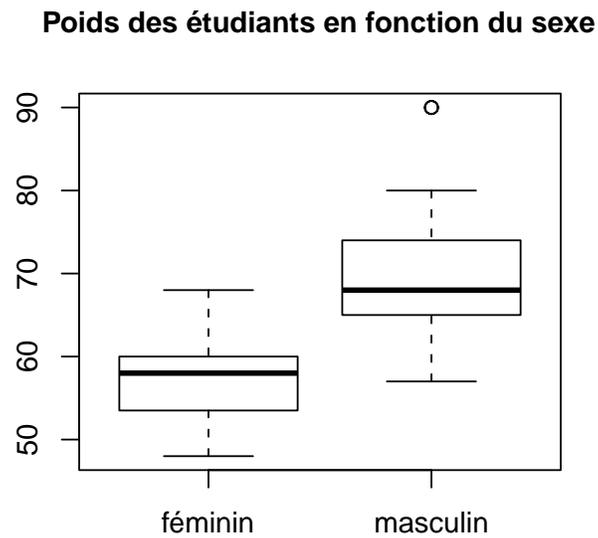


FIGURE 8.12. Boîte à moustache des poids des étudiants par sexe (données L3APA06).

## Récapitulatif des notions et commandes essentielles (statistiques descriptives)

Vous trouverez dans ce chapitre les notions et commandes essentielles à retenir.

*Compte tenu des modifications mineures faites en cours de semestre, les numéros de pages indiquées peuvent avoir changé par rapport à la version papier distribuée : il faut donc se référer au dernier document électronique de ce cours en pdf, disponible sur le web et sur le réseau de l'université.*

### 9.1. Analyse univariée

#### 9.1.1. Importer des données

Voir section 3.2 page 25.

- Pour stocker dans la variable (data frame) 'nom', le contenu du fichier 'nom.txt', on tape :  
`nom <- read.table("nom.txt", h = T)`
- Pour avoir la totalité des statistiques de chacune des variables du data frame 'nom', on tape :  
`summary(nom)`
- Faire un éventuel attachement :  
`attach(nom)`
- Pour voir le nom des différentes variables de ce data frame :  
`names(nom)`
- Pour voir seulement le haut de ce data frame :  
`head(nom)`
- Pour obtenir uniquement la variable 'variableY' du data frame 'nom' :  
`nom$variableY`  
ou directement (en cas d'attachement) :  
`variableY`
- *Attention*, si vous avez un tableau de nom 'Y' qui ne provient pas d'un data frame, il faudra juste taper  
`Y`
- Faire un éventuel détachement (si l'attachement a été précédemment fait) :  
`detach(nom)`

#### 9.1.2. Variable qualitative (ou catégorielle)

Voir section 3.3 page 26.

On suppose que la commande 'read.table' a déjà été tapée!

##### 9.1.2.1. Indicateurs.

- Afficher les différentes fréquences de la variable 'variableY' :  
`table(nom$variableY)`
- Afficher les différentes fréquences de la variable 'variableY' en les triant :  
`sort(table(nom$variableY))`
- Afficher les différents pourcentages de la variable 'variableY' :

```
u <- table(nom$variableY)
100 * u/sum(u)
```

- Afficher les différents pourcentages de la variable 'variableY' en les triant :

```
u <- sort(table(nom$variableY))
100 * u/sum(u)
```

#### 9.1.2.2. Graphiques.

(1) Peu de données :

- Camembert (tourte)
 

```
pie(table(nom$variableY))
```
- Diagramme en barre
 

```
barplot(table(nom$variableY))
```
- Diagramme en barre (classé)
 

```
barplot(sort(table(nom$variableY)))
```

(2) Beaucoup de données (attention, voir la remarque 3.3 page 27) :

- Diagramme de Cléveland
 

```
dotchart(table(nom$variableY))
```
- Diagramme de Cléveland (classé)
 

```
dotchart(sort(table(nom$variableY)))
```

### 9.1.3. Variable quantitative (ou numérique)

Voir section 3.4 page 28.

On suppose que la commande 'read.table' a déjà été tapée!

#### 9.1.3.1. Indicateurs.

- Afficher les différentes statistiques de la variable 'variableY' (sauf écart-type) :

```
summary(nom$variableY)
```

- Afficher l'écart-type de la variable 'variableY' :

```
sd(nom$variableY)
```

#### 9.1.3.2. Graphiques.

(1) Peu de données :

- Ligne de point (avec empilement)
 

```
stripchart(nom$variableY, method = "stack")
```

(2) Beaucoup de données :

- Histogramme
 

```
hist(nom$variableY)
```
- Boîte de dispersion
 

```
boxplot(nom$variableY)
```

## 9.2. Analyse bivariée

On veut croiser les variables 'variableX' et 'variableY' du data frame 'nom'.

### 9.2.1. Quantitatif $\times$ quantitatif

Voir chapitre 4 page 41.

- Après avoir téléchargé la fonction 'determin.quantiquanti', il faut la sourcer.
- Pour obtenir la pente  $a$ , l'ordonnée à l'origine  $b$  de la droite de régression linéaire, ainsi que la corrélation linéaire  $r$  et la probabilité critique  $p_c$  :

```
determin.quantiquanti(nom$variableX, nom$variableY)
```

- On compare

- (1)  $|r|$  à 0 et 1 ; si elle est proche de 0, les points du nuage ne sont pas alignés et il n'y a pas de relation (de type affine) ; si elle est proche de 1, les points du nuage ne sont alignés et il a une relation (de type affine) ;
  - (2)  $|r|$  aux seuils de Cohen (0.1, 0.3, 0.5) (voir équation (4.7) page 43) ;
  - (3)  $p_c$  à la valeur seuil de 0.05 (voir définition 4.4 page 45).
- Pour obtenir en plus la figure (nuage de point et droite de corrélation) :  
`determin.quantiquanti(nom$variableX, nom$variableY, fig = T)`

### 9.2.2. Qualitatif × qualitatif

Voir chapitre 5 page 57.

- Après avoir téléchargé la fonction 'determin.qualiquali', il faut la sourcer.
- Pour obtenir le coefficient  $\chi^2$ , le coefficient de Cramer  $V$ , la taille d'effet  $w$  et la probabilité critique  $p_c$  :  
`determin.qualiquali(nom$variableX, nom$variableY)`
- Pour obtenir en plus la table de contingence :  
`determin.qualiquali(nom$variableX, nom$variableY, tabcontingence = T)`
- On compare
  - (1)  $V$  à 0 et 1 ; s'il est proche de 0, les variables ne sont pas liées ; s'il est proche de 1, les variables sont liées.
  - (2)  $w$  aux seuils de Cohen (0.1, 0.3, 0.5) (voir équation (5.5) page 60) ;
  - (3)  $p_c$  à la valeur seuil de 0.05 (voir définition 4.4 page 45).

### 9.2.3. Qualitatif × quantitatif

Voir chapitre 6 page 67.

- Après avoir téléchargé la fonction 'determin.qualiquanti', il faut la sourcer.
- Pour obtenir les statistiques par groupes, le rapport de corrélation  $RC$  et la probabilité critique  $p_c$  :  
`determin.qualiquanti(nom$variableX, nom$variableY)`  
 où 'nom\$variableX' est la variable quantitative et 'nom\$variableY' est la variable qualitative.
- On compare
  - (1)  $RC$  à 0 et 1 ; s'il est proche de 0, les variables ne sont pas liées ; s'il est proche de 1, les variables sont liées.
  - (2)  $RC$  aux seuils de Cohen (0.01, 0.05, 0.15) (voir équation (6.4) page 73) ;
  - (3)  $p_c$  à la valeur seuil de 0.05 (voir section 6.2.4 page 74).
- Pour obtenir en plus la figure (boîtes de dispersion et lignes de point par groupes) :  
`determin.qualiquanti(nom$variableX, nom$variableY, fig = T)`



## Exercices de révision (statistiques descriptives)

### 10.1. Énoncés

EXERCICE 10.1.

On étudie le fichier 'L3APA06.txt'.

Analyser la variable 'sport'.

EXERCICE 10.2.

On étudie le fichier 'L3APA06.txt'.

Analyser la variable 'sport'. On prendra en compte les valeurs non déterminées.

EXERCICE 10.3.

On étudie le fichier 'L3APA06.txt'.

Analyser la variable 'rythmcard'.

EXERCICE 10.4.

On étudie le fichier 'L3APA06.txt'.

Étudier le croisement de la variable 'taille' et de la variable 'rythmcard'.

EXERCICE 10.5.

On étudie le fichier 'L3APA06.txt'.

Étudier le croisement de la variable 'sexe' et de la variable 'baccalaureat'.

EXERCICE 10.6.

On étudie le fichier 'L3APA06.txt'.

Étudier le croisement de la variable 'sexe' et de la variable 'rythmcard'.

EXERCICE 10.7.

On étudie le fichier 'L3APA06.txt'.

Étudier le croisement de la variable 'taille' et de la variable 'oeil'.

### 10.2. Corrigés

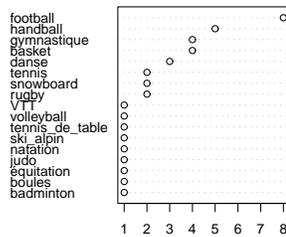
ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.1

- On étudie la variable qualitative (ou catégorielle) 'sport'. Pour les manipulations avec  $\mathbb{R}$ , on renvoie donc à la section 3.3 et aux sections récapitulatives 9.1.1 et 9.1.2 du document de cours.
- Les effectifs et les pourcentages déterminés par  $\mathbb{R}$  sont donnés dans le tableau suivant

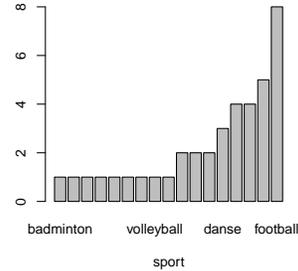
-

	effectifs	pourcentages
badminton	1	2.564
boules	1	2.564
équitation	1	2.564
judo	1	2.564
natation	1	2.564
ski_alpin	1	2.564
tennis_de_table	1	2.564
volleyball	1	2.564
VTT	1	2.564
rugby	2	5.128
snowboard	2	5.128
tennis	2	5.128
danse	3	7.692
basket	4	10.256
gymnastique	4	10.256
handball	5	12.821
football	8	20.513

Cléviand pour sport



Barres pour sport



Camembert pour sport



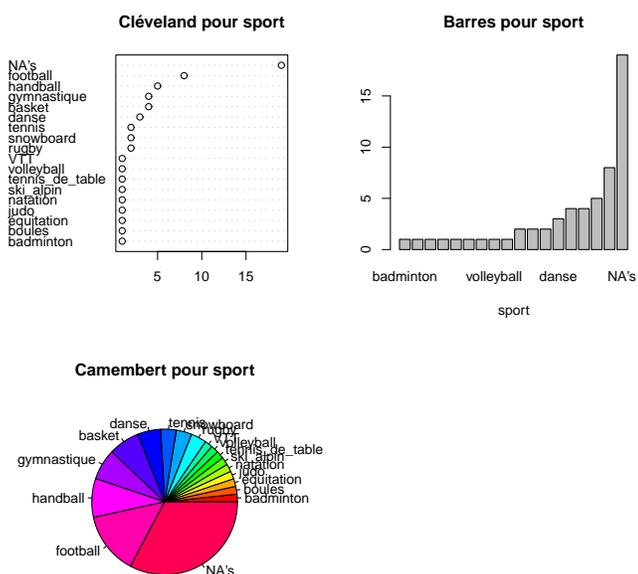
Voir les trois graphiques ci-dessus pour la variable 'sport'.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.2

- On étudie la variable qualitative (ou catégorielle) 'sport'. Pour les manipulations avec  $\mathcal{R}$ , on renvoie donc à la section 3.3 et aux sections récapitulatives 9.1.1 et 9.1.2 du document de cours.
- Les effectifs et les pourcentages déterminés par  $\mathcal{R}$  (en prenant en compte les éventuelles valeurs non déterminées (NA)) sont donnés dans le tableau suivant

•

	effectifs	pourcentages
badminton	1	1.724
boules	1	1.724
équitation	1	1.724
judo	1	1.724
natation	1	1.724
skiAlpin	1	1.724
tennis_de_table	1	1.724
volleyball	1	1.724
VTT	1	1.724
rugby	2	3.448
snowboard	2	3.448
tennis	2	3.448
danse	3	5.172
basket	4	6.897
gymnastique	4	6.897
handball	5	8.621
football	8	13.793
NA's	19	32.759



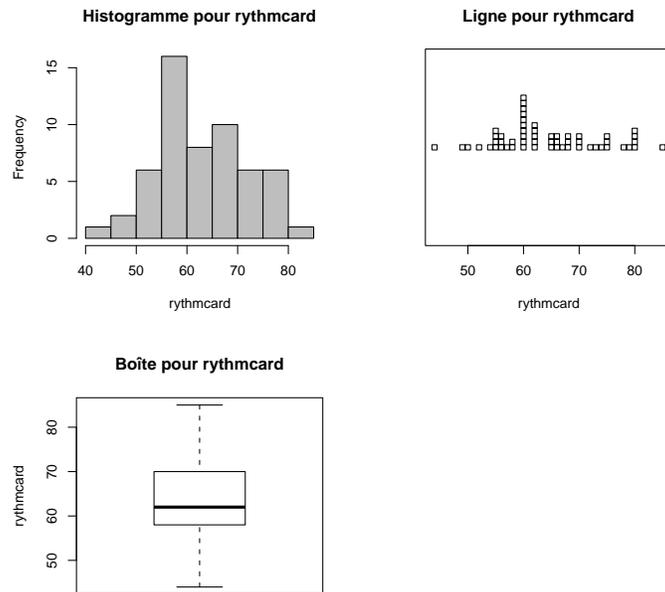
Voir les trois graphiques ci-dessus pour la variable 'sport'.

### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.3

- On étudie la variable quantitative (ou numérique) 'rythmcard'. Pour les manipulations avec  $\mathbb{R}$ , on renvoie donc à la section 3.4 et aux sections récapitulatives 9.1.1 et 9.1.3 du document de cours.
- Les différents résultats déterminés par  $\mathbb{R}$  sont donnés dans le tableau suivant

noms	valeurs
moyenne	64.285714
écart-type	9.096881
$Q_1$ (quartile à 25 %)	58
médiane	62
$Q_3$ (quartile à 75 %)	70
minimum	44
maximum	85
nombre	58

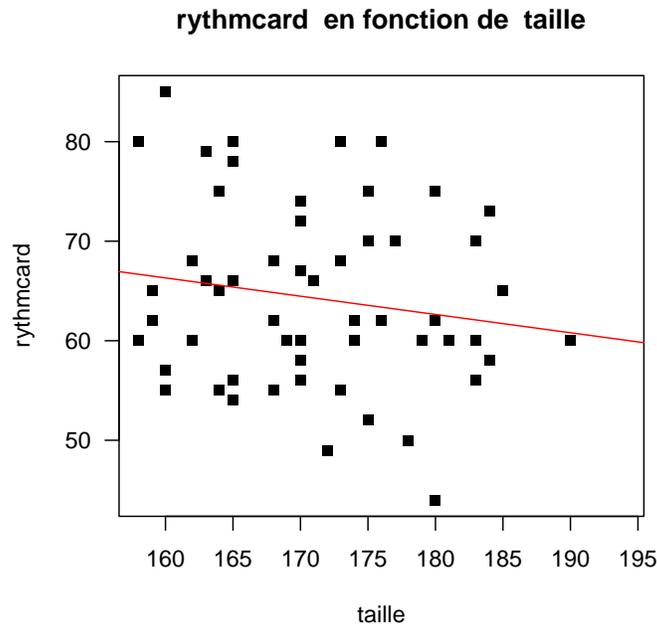
•



Voir les trois graphiques ci-dessus pour la variable 'rythmcard'.

#### ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.4

- On étudie le croisement de la variable quantitative (ou numérique) 'taille' et de la variable quantitative (ou numérique) 'rythmcard'. Pour les manipulations avec  $\mathcal{R}$ , on renvoie donc à la section 4.5 et la section récapitulative 9.2.1 du document de cours.
- Voir la figure ci-dessous.



Sur cette figure, les points semblent peu alignés.

- Confirmons cela grâce à  $\mathcal{R}$ .

Les résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
pente $a$	-0.183626
ordonnée à l'origine $b$	95.692316
corrélation linéaire $r$	-0.163126
probabilité critique $p_c$	0.229651

On compare la valeur absolue de la corrélation linéaire  $r = -0.163126$  aux seuils de Cohen (0.1,0.3,0.5) (voir [16]) et la probabilité critique  $p_c = 0.229651$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	<b>moyenne</b>
significativité statistique	<b>non</b>

- On peut donc affirmer qu'il existe une relation faible entre les variables 'taille' et 'rythmcard'.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.5

- On étudie le croisement de la variable qualitative (ou catégorielle) 'sexe' et de la variable qualitative (ou catégorielle) 'baccalaureat'. Pour les manipulations avec  $\mathcal{R}$ , on renvoie donc à la section 5.5 et la section récapitulative 9.2.2 du document de cours.
- La table de contingence déterminée par  $\mathcal{R}$  est donnée dans le tableau suivant

Les autres résultats donnés par  $\mathcal{R}$  sont les suivants :

	ES	L	S	SMS	STAE
féminin	10	6	14	1	0
masculin	6	0	20	0	1

Noms des indicateurs	Valeurs
$\chi^2$	9.829714
coefficient de Cramer $V$	0.411677
taille d'effet $w$	0.411677
probabilité critique $p_c$	0.0433959

On compare la taille d'effet  $w=0.411677$  aux seuils de Cohen (0.1,0.3,0.5) (voir [16]) et la probabilité critique  $p_c=0.0433959$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

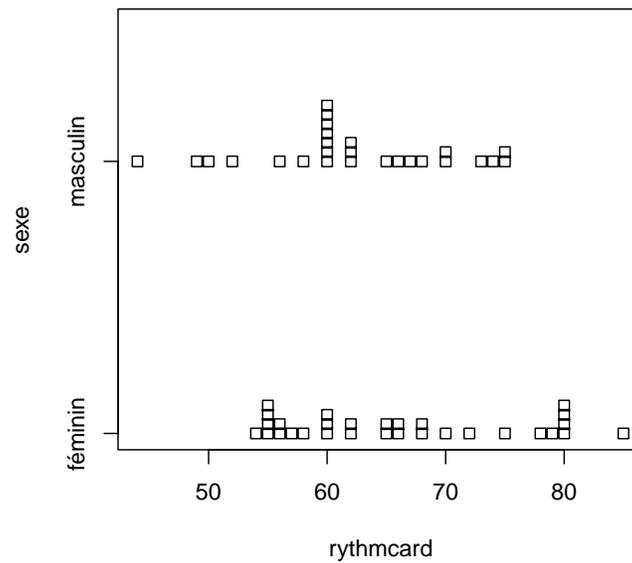
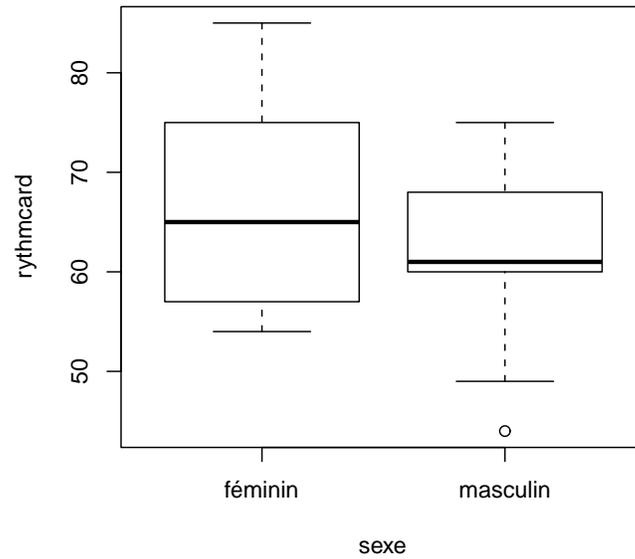
significativité pratique	<b>forte</b>
significativité statistique	<b>oui</b>

- On peut donc affirmer qu'il existe une relation entre les variables 'sexe' et 'baccalaureat'.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.6

- On étudie le croisement de la variable qualitative (ou catégorielle) 'sexe' et de la variable quantitative (ou numérique) 'rythmcard'. Pour les manipulations avec  $\mathcal{R}$ , on renvoie donc aux sections 6.2 et 6.3 et la section récapitulative 9.2.3 du document de cours.

•



Voir la figure ci-dessous.

- Avec `summary()`, on obtient les statistiques par groupes données dans le tableau suivant ;

	mean	sd	0%	25%	50%	75%	100%	n	NAs
féminin	66.07	9.67	54.00	57.25	65.00	74.25	85.00	30	1
masculin	62.23	8.08	44.00	60.00	61.00	67.75	75.00	26	1

On rappelle que, dans ce tableau :

- le nombre noté 0% est le quartile à 0 % (c'est le minimum) ;
- le nombre noté 25% est le quartile à 25 % (c'est  $Q_1$ ) ;
- le nombre noté 50% est le quartile à 50 % (c'est la médiane) ;
- le nombre noté 75% est le quartile à 75 % (c'est  $Q_3$ ) ;
- le nombre noté 100% est le quartile à 100 % (c'est le maximum).

Les graphiques et les statistiques par groupes montrent une certaine hétérogénéité entre les types.

Confirmons cela grâce à  $\mathcal{R}$ .

Les autres résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.045029
probabilité critique $p_c$	0.116395

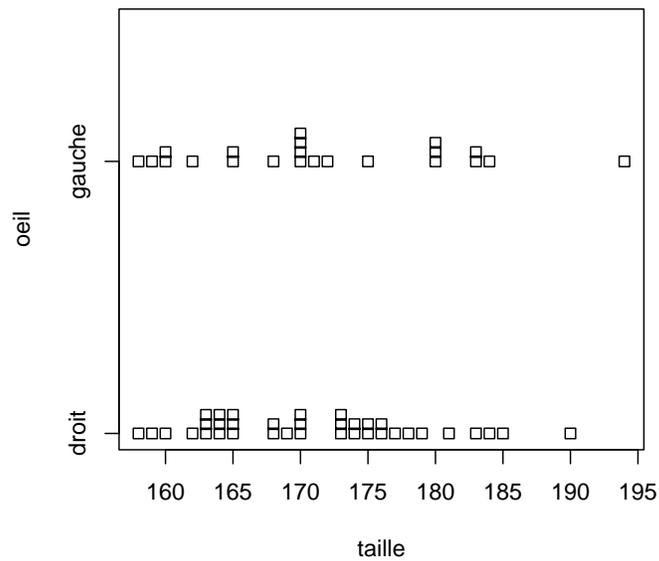
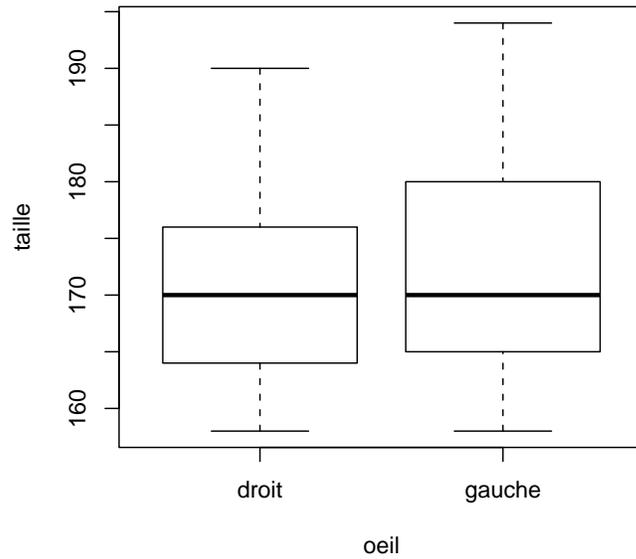
On compare le rapport de corrélation  $RC=0.045029$  aux seuils de Cohen (0.01,0.05,0.15) (voir [16]) et la probabilité critique  $p_c=0.116395$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	<b>moyenne</b>
significativité statistique	<b>non</b>

- On peut donc affirmer qu'il existe une relation entre les variables 'sexe' et 'rythmcard'.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 10.7

- On étudie le croisement de la variable quantitative (ou numérique) 'taille' et de la variable qualitative (ou catégorielle) 'oeil'. Pour les manipulations avec  $\mathcal{R}$ , on renvoie donc aux sections 6.2 et 6.3 et la section récapitulative 9.2.3 du document de cours.
-



Voir la figure ci-dessous.

- Avec , on obtient les statistiques par groupes données dans le tableau suivant ;

	moyenne	écart-type	0%	25%	50%	75%	100%	n
droit	171.00	7.96	158.00	164.00	170.00	176.00	190.00	36
gauche	171.77	9.65	158.00	165.00	170.00	180.00	194.00	22

Les graphiques et les statistiques par groupes montrent peu de différence entre les types.  
 Confirmons cela grâce à  $\mathcal{R}$ .

Les autres résultats donnés par  $\mathcal{R}$  sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.001951
probabilité critique $p_c$	0.742008

On compare le rapport de corrélation  $RC=0.001951$  aux seuils de Cohen (0.01,0.05,0.15) (voir [16]) et la probabilité critique  $p_c=0.742008$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	<b>faible</b>
significativité statistique	<b>non</b>

- On peut donc affirmer qu'il n'existe pas de relation entre les variables 'taille' et 'oeil'.

## Installation du logiciel

### A.1. Installation de pour Windows

- (1) Aller sur le site <http://www.r-project.org/>
- (2) Cliquer sur "Download, Packages, CRAN", puis pour limiter le temps de téléchargement, choisir "France", <http://cran.univ-lyon1.fr/>.
- (3) Dans la rubrique "Download and Install R", choisir (pour Windows ; bien entendu, sont aussi distribuées des versions pour Mac et Linux) Windows.
- (4) Puis, cliquer sur "base Binaries for base distribution".
- (5) Cliquer enfin sur "Download R-2.11.1 for Windows (33 megabytes, 32 bits)"
- (6) Télécharger alors le logiciel d'installation `R-2.11.1-win32.exe` *Attention, le numéro de version de  est souvent réactualisé depuis la compilation de ce document !*
- (7) Double-cliquer sur le logiciel `R-2.11.1-win32.exe` (ou la dernière version en date) afin de procéder à l'installation de R. Choisir les options par défaut afin de l'installer sur le disque `C:\`.

REMARQUE A.1. Dans la rubrique "R-2.11.1 for Windows" (ou la dernière en date), vous pouvez aussi télécharger les anciennes versions de , parfois utiles quand les plus récentes peuvent être instables ou présenter un bug non encore résolu ! Voir "Previous releases". Dans cette même rubrique, vous pouvez aussi récupérer des packages des anciennes versions, non distribuées dans la plus récente.

REMARQUE A.2. Dans la rubrique "The Comprehensive R Archive Network", sous-rubrique "Source Code for all Platforms", puis "Contributed extension packages", vous pouvez aussi récupérer des packages des anciennes versions (sous forme de zip) , non distribuées dans la plus récente.

### A.2. Utilisation de

Utiliser le menu démarrer, puis Tous les programmes, puis R, puis R-2.11.1 (ou la dernière version en date). Le logiciel s'ouvre alors et une fenêtre "Rconsole" apparaît.

Il faut indiquer au logiciel R dans quel répertoire windows il doit aller chercher les fichiers (en particulier les jeux de données) dont nous avons besoin et où les sauvegarder également ; ce répertoire est dit *répertoire de travail*. Dans le menu déroulant "Fichier", existe une option "Changer de répertoire courant" qui par l'intermédiaire d'une arborescence permet de choisir le répertoire qui nous convient (par défaut c'est `C:\R\R-2.11.1`).

En quittant R, ne pas sauvegarder la session !



## Prise en main à la première séance

*Cette annexe est destinée à ceux qui se sentent peu habitués aux opérations de téléchargement de fichiers, de démarrage de logiciels et pourra être lue en première séance.*

### B.1. Création d'un dossier de travail (ou répertoire courant)

Il est nécessaire de créer un dossier de travail pour stocker le photocopie de cours et les fichiers de données. Pour cela,

- (1) Ouvrez un "Explorateur Windows" ou dans "poste de travail", allez dans le répertoire W:. Ce répertoire vous est propre et vous y aurez accès à chaque ouverture de session (avec vos propres identifiants).
- (2) Créez-y un dossier (ou répertoire), par exemple appelé "statistiques".

Ce dossier constitue votre répertoire de travail ou répertoire courant.

### B.2. Téléchargement du cours et des fichiers de données

Ce photocopie de cours et les fichiers de données sont normalement disponibles à la fois

- en ligne sur <http://utbmjb.chez-alice.fr/UFRSTAPS/index.html> à la rubrique habituelle ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  - 'Poste de travail',
  - puis sur le répertoire 'P:' (appelé aussi '\\teraetu\Enseignants'),
  - puis 'jerome.bastien',
  - enfin sur 'M1GAPAS'.

Pour l'examen, les données se trouveront aussi, par mesure de précaution à ces deux endroits.

- (1) Rendez-vous sur donc soit sur internet soit (en cas de problème de connexion) sur le réseau et
  - ou bien sur internet, téléchargez dans votre répertoire de travail le photocopie de cours (rubrique "Version provisoire du cours" ou "Version définitive du cours"), grâce au clic droit "enregistrer sous"
  - ou bien sur le réseau, copiez-collez le photocopie de cours vers votre répertoire de travail.
- (2) Faites de même pour les fichiers de données (disponibles soit sous la forme de fichiers txt ou xls, soit la forme d'un fichier "zipé").
- (3) Dans votre répertoire courant, cliquez sur la version pdf du cours.
- (4) Dans votre répertoire courant, dézipiez éventuellement (clic droit, "extraire ici") les fichiers de données.

### B.3. Installation du logiciel

- (1) Bouton "démarrer", puis "tous les programmes", puis "R".
- (2) Déclarer le répertoire courant avec le menu déroulant "Fichier" puis l'option "Changer le répertoire courant", et indiquer le répertoire créé en section B.1.



## ANNEXE C

### **Exemple de d'analyse univariée : quelques statistiques sur les lettres d'un texte**

Cet exercice a déjà été donné en examen de L3APAS.

Voir l'exercice 4 du CCF2 du 13 Mai 2011, à la rubrique L3APAS (ou alors directement sur <http://utbmjb.chez-alice.fr/UFRSTAPS/L3APA/CCF2L3APASP11web.pdf> et <http://utbmjb.chez-alice.fr/UFRSTAPS/L3APA/corCCF2L3APASP11.pdf>).



## Utilisation de fonctions avec

*Ici, la notion de fonction est introduite pour vous simplifier vos démarches avec , mais l'usage de ces fonctions n'est nullement imposé (sauf éventuellement pour faire tourner des démonstrations)!*

### D.1. Une fonction "simple"

Que vous soyez un adepte de Rcmdr ou non, on peut maintenant évoquer la notion de fonction. Quand vous tapez par exemple

```
cos(5)
```

cela calcule le cosinus de 5. On peut aussi sois-même écrire des fonctions. Voir la fonction `somme.R`, disponible à l'URL habituelle (dans la rubrique "fonctionsR"). Vous pouvez visualiser ce fichier et constatez qu'il comporte :

- un entête :
 

```
somme <- fonction(x, y) {
}
```
- éventuellement des commentaires (lignes commençant par # et qui expliquent le fonctionnement de la fonction) :
 

```
# exemple de fonction : somme de deux nombres
# *****
# somme(x,y) :
# * Variables d'entrées :
#   * x,y : les deux nombres dont on veut la somme
# * Variable de sortie :
#   * la somme de x et de y
```
- et un corps de fonction
 

```
s <- x + y
return(s)
```

qui retourne la somme des deux nombre. Ce corps de fonction comporte en fait les différentes étapes (ici très simple) nécessaires au calcul exigé.

Pour utiliser cette fonction, il faut

- (1) d'abord "sourcer" cette fonction. Vous avez quatre possibilités :
  - (a) Soit récupérer le fichier `somme.R` dans le répertoire de travail et faire "fichier", puis "Sourcer du code R" et choisir `somme.R`.
  - (b) Soit récupérer le fichier `somme.R` dans le répertoire de travail et taper la ligne de commande (dans Rgui) :
 

```
source("somme.R")
```
  - (c) Soit s'affranchir de la récupération du fichier `somme.R` en tapant directement
 

```
source("http://utbmjb.chez-alice.fr/UFRSTAPS/M1APA/fonctionsR/somme.R")
```

 ce qui ne marche que si la connexion internet est correcte!

(d) Si vous avez accès au texte de la fonction, ici

```
somme<-function(x,y){
  # exemple de fonction : somme de deux nombres
  # *****
  # somme(x,y) :
  # * Variables d'entrées :
  #   * x,y : les deux nombres dont on veut la somme
  # * Variable de sortie :
  #   * la somme de x et de y

  s<-x+y
  return(s)
}
```

il faut en faire un copier-coller et à partir d'un éditeur simple (type bloc-note) l'enregister dans un fichier de nom `somme.R` dans votre répertoire de travail. Vous pouvez aussi utiliser l'éditeur *ad hoc* de , en allant dans "fichier", "Nouveau script", puis une fois le texte collé, faite "sauver".

REMARQUE D.1. Cette éditeur vous permet aussi de voir des fichiers R déjà écrits en allant dans "fichier", puis "ouvrir un script".

Comme précédemment, il faudra alors le "sourcer".

(2) Cette fonction est donc chargée et, ensuite, vous pourrez la faire tourner en tapant par exemple (ici, les deux arguments sont 'x' et 'y')

```
somme(2, 3)
[1] 5
à comparer à
2 + 3
[1] 5
```

## D.2. Une fonction à deux valeurs de sortie

Considérons maintenant la fonction `somme_rapport.R`, disponible à l'URL habituelle. Comme indiqué dans la section D.1, récupérez et sourcez-la. Cette fonction renvoie deux expressions, la somme et le rapport de deux nombres. Tapez par exemple

```
somme_rapport(2, 3)
$s
[1] 5
```

```
$r
[1] 0.6666667
```

Cette fonction renvoie en fait une liste avec deux éléments (ce qui permet d'avoir plusieurs valeurs de sortie). On pourra pour comprendre comment fonctionne une liste en tapant par exemple

```
res <- somme_rapport(2, 3)
class(res)
[1] "list"
```

```
names(res)
[1] "s" "r"

res$s
[1] 5

res$r
[1] 0.6666667
```

On peut aussi définir les valeurs des arguments "dans le désordre" à condition de spécifier quel argument est  $x$  et quel argument est  $y$ . Comparez ce que donne

```
somme_rapport(2, 3)
```

```
$s
[1] 5
```

```
$r
[1] 0.6666667
```

```
somme_rapport(3, 2)
```

```
$s
[1] 5
```

```
$r
[1] 1.5
```

```
somme_rapport(x = 2, y = 3)
```

```
$s
[1] 5
```

```
$r
[1] 0.6666667
```

```
somme_rapport(y = 3, x = 2)
```

```
$s
[1] 5
```

```
$r
[1] 0.6666667
```

Une fonction peut aussi avoir un argument optionnel. Quand il n'est pas indiqué, il prend la valeur imposée par défaut par la fonction. Par exemple, si  $y$  n'est pas indiqué, il vaut 1. Comparez ce que donne

```
somme_rapport(2, 1)
```

```
$s
[1] 3
```

```
$r
[1] 2
```

```
somme_rapport(y = 1, x = 2)
```

```
$s  
[1] 3
```

```
$r  
[1] 2
```

```
  somme_rapport(2)
```

```
$s  
[1] 3
```

```
$r  
[1] 2
```

### D.3. D'autres fonctions

Nous utiliserons dans ce cours un certain nombre de fonctions, déjà écrites et disponibles à l'URL habituelle. Elles permettent de faire des calculs déjà programmés et fréquemment utilisés.

Bien entendu, vous pourrez vous même écrire vos propres fonctions. Consultez la section 6.3 page 72 de l'excellente introduction à  $\mathbb{R}$ , [20] disponible sur internet.

## Un exemple "pédagogique" sur les danger de la régression linéaire (sous forme d'exercice corrigé)

Cet exercice a déjà été donné en examen de M1IGAPAS (CCF2 Automne 2009).

### Énoncé

EXERCICE E.1.

On étudie le fichier de données 'anscombe.txt'.

- (1) Étudier successivement et de façon graphique les relations linéaires entre les variables :
  - 'X' et 'Y1' ;
  - 'X' et 'Y2' ;
  - 'X' et 'Y3' ;
  - 'Xp' et 'Yp'.
- (2) Pour chacune de ces relations linéaires, déterminer les coefficients de corrélations linéaires et les probabilités critiques.
- (3) Conclure.

### Corrigé

ÉLÉMENTS DE CORRECTION DE L'EXERCICE E.1

Cet exemple pédagogique a été mis au point par Anscombe [21] et provient de l'ouvrage [4].

- (1) On étudie le croisement de la variable quantitative (ou numérique) 'X' et de la variable quantitative (ou numérique) 'Y1'. Pour les manipulations avec  $\mathbb{R}$ , on renvoie donc à la section 4.5 et la section récapitulative 9.2.1 du document de cours.

On a indiqué en figure E.1 page 122 et E.2 page 123, les quatres nuages de points et les droites de régression linéaire.

- Le premier présente un nuages de points qui semblent être à peu près alignés, pour lequel la régression linéaire a l'air pertinente.
  - Le deuxième graphique nous indique un nuage de point en forme de parabole tournée vers le bas ; la régression linéaire n'est donc pas pertinente.
  - Sur le troisième graphique, on peut constater, qu'hormis le dernier point, les points ont l'air d'être alignés. Cependant, ce dernier point, mesure extrême, a tendance à attirer la droite et la modifie par rapport au nuage de point sans cette donnée extrême ; la régression linéaire n'est donc pas pertinente.
  - Enfin, sur le quatrième graphique, on constate que tous les points sauf un, on la même abscisse. Il n'existe donc pas de droite de régression pour les premiers points. Le dernier point modifie sensiblement la droite de régression ; la régression linéaire n'est donc pas pertinente.
- (2) Étudions le croisement des variables 'X' et 'Y1'  
Les résultats donnés par  $\mathbb{R}$  sont les suivants :

Noms des indicateurs	Valeurs
pente $a$	0.812452
ordonnée à l'origine $b$	0.451378
corrélation linéaire $r$	0.786901
probabilité critique $p_c$	0.000298198

On compare la valeur absolue de la corrélation linéaire  $r=0.786901$  aux seuils de Cohen (0.1,0.3,0.5) (voir [16]) et la probabilité critique  $p_c=0.000298198$  à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	<b>très forte</b>
significativité statistique	<b>oui</b>

Croisement	pente $a$	ordonnée à l'origine $b$	corrélation linéaire $r$	probabilités critique $p_c$
('X', 'Y1')	0.81245168	0.45137803	0.78690107	0.0002982
('X', 'Y2')	0.80852812	0.52367369	0.78539584	0.00031191
('X', 'Y3')	0.80619049	0.55752951	0.78427149	0.00032249
('Xp', 'Yp')	0.78255247	1.30087118	0.78384874	0.00032654

Dans le tableau ci-dessous, on a indiqué les quatre pentes et ordonnées à l'origine, ainsi que les quatre coefficients de corrélation linéaire et les quatre probabilités critiques obtenues. Les trois premières pentes et ordonnées l'origine sont à peu près égales! Ainsi, les quatre nuages de points donnent mêmes corrélations linéaires et mêmes probabilités critiques! Cependant, d'après nos observations graphiques précédentes, seule la première régression linéaires est pertinente.

- (3) La morale de l'histoire, c'est qu'il convient donc toujours de commencer par une visualisation des données avant de continuer les calculs de corrélation linéaire et de probabilité critique!

REMARQUE E.2. Les données étudiées ici, créées de façon pédagogiques par Anscombe, sont en fait déjà présentes dans  $\mathbb{R}$ ! Il suffit de taper dans  $\mathbb{R}$  :

```
data(anscombe)
anscombe
```

Les variables du data frame 'anscombe' sont : 'x1', 'x2', 'x3', 'x4', 'y1', 'y2', 'y3' et 'y4'.

On obtient des nuages de points un peu différents que ceux créés par le fichier de données, mais leurs propriétés sont les mêmes!

Croisement	pente $a$	ordonnée à l'origine $b$	corrélation linéaire $r$	probabilités critique $p_c$
('x1', 'y1')	0.50009091	3.00009091	0.81642052	0.00216963
('x2', 'y2')	0.5	3.00090909	0.81623651	0.00217882
('x3', 'y3')	0.49972727	3.00245455	0.81628674	0.00217631
('x4', 'y4')	0.49990909	3.00172727	0.81652144	0.0021646

Dans le tableau ci-dessous, on a indiqué les quatre pentes et ordonnées à l'origine, ainsi que les quatre coefficients de corrélation linéaire et les quatre probabilités critiques obtenues pour les données 'anscombe'.

On a indiqué en figure E.3 page 124 et E.4 page 125, les quatres nuages de points et les droites de régression linéaire.

On pourra aussi consulter la rubrique de Wikipédia sur le quartet d'anscombe : [http://fr.wikipedia.org/wiki/Quartet\\_d'Anscombe](http://fr.wikipedia.org/wiki/Quartet_d'Anscombe)

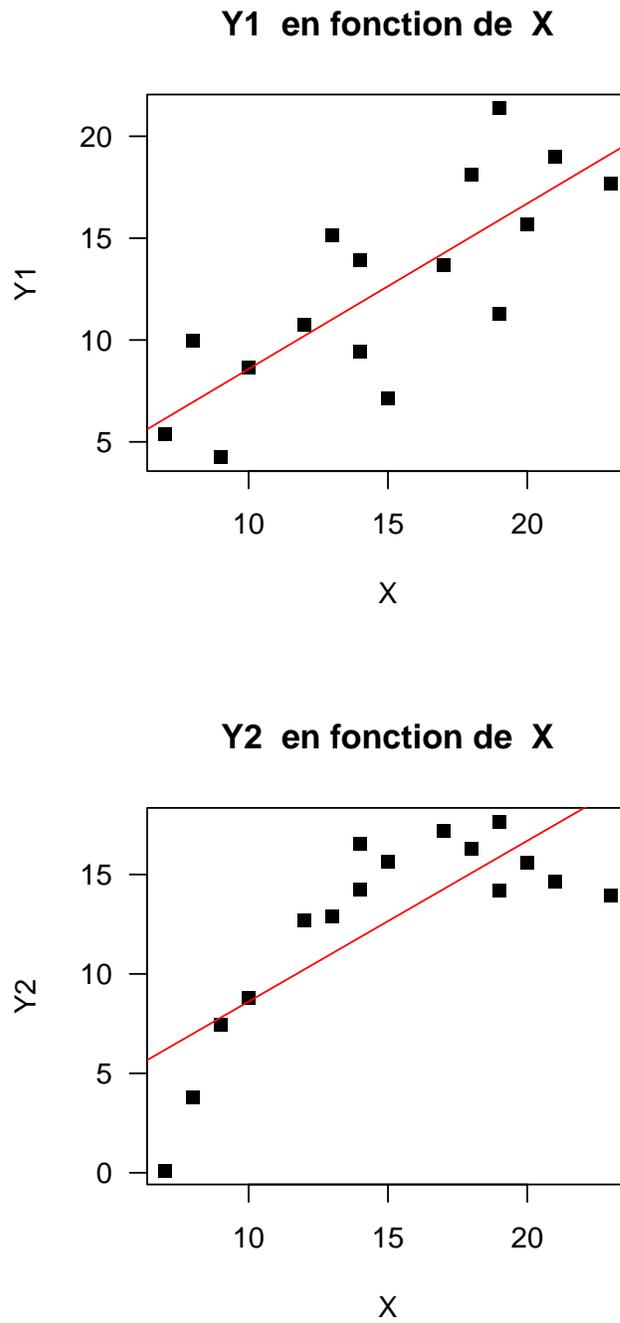


FIGURE E.1. Les deux premiers nuages de points et les droites de régression linéaire.

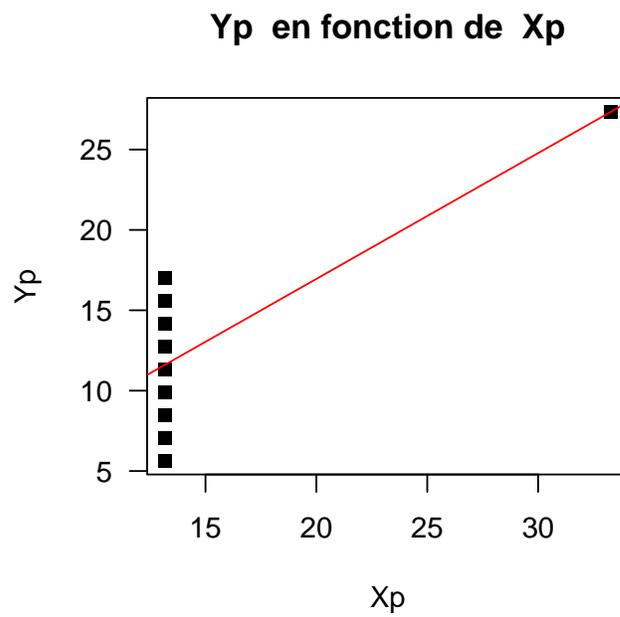
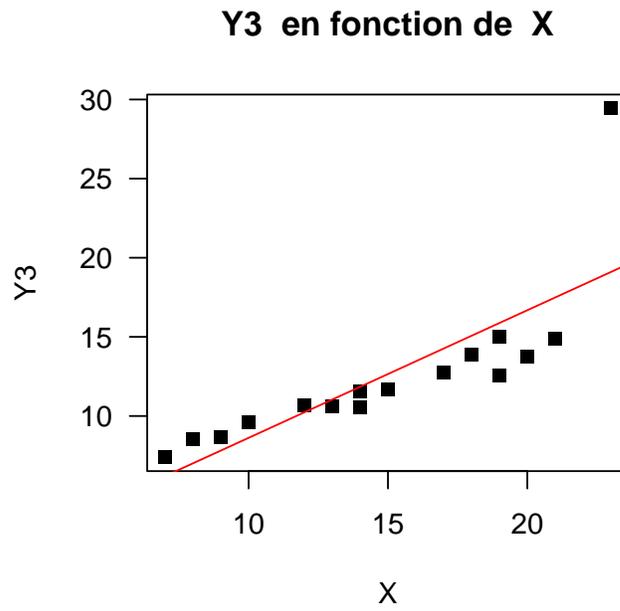


FIGURE E.2. Les deux derniers nuages de points et les droites de régression linéaire.

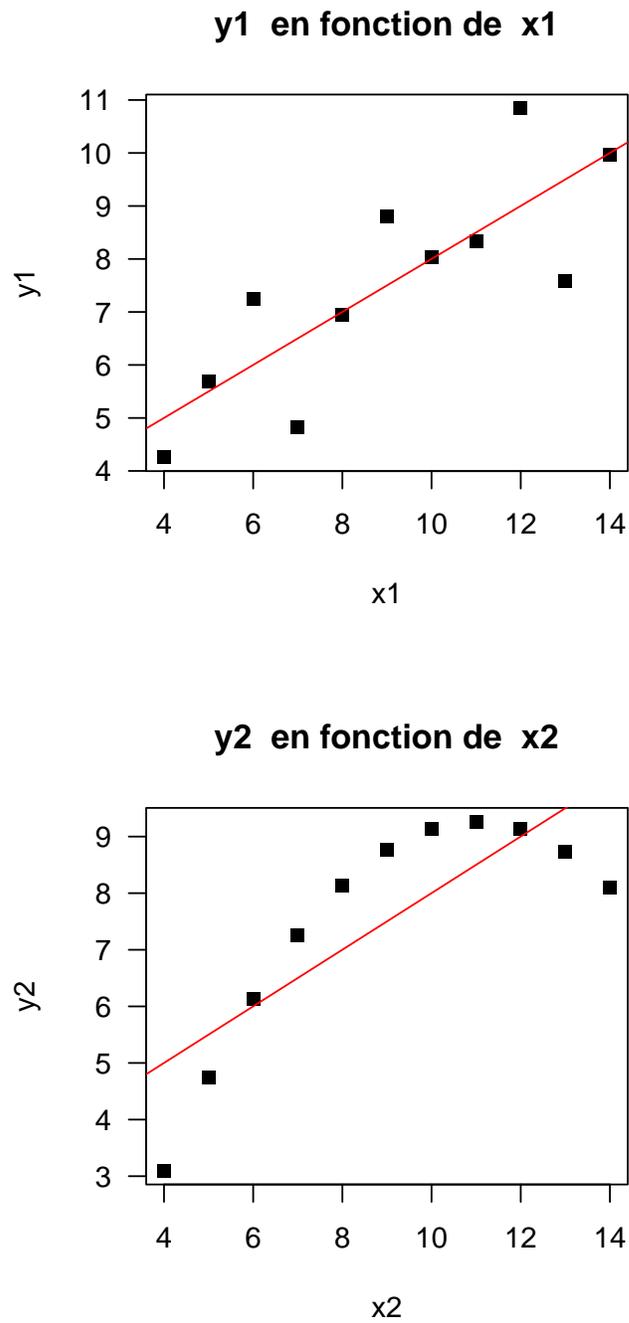


FIGURE E.3. Les deux premiers nuages de points et les droites de régression linéaire pour les données 'anscombe'.

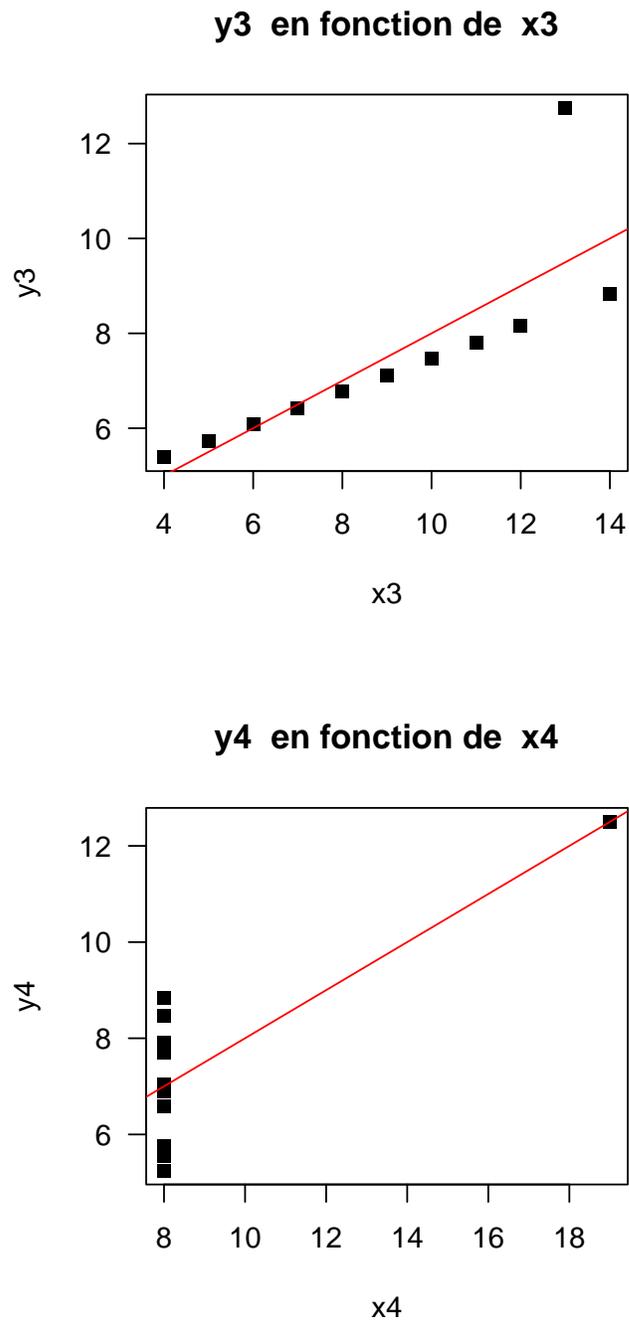


FIGURE E.4. Les deux derniers nuages de points et les droites de régression linéaire pour les données 'anscombe'.



## Bibliographie

- [1] AB Dufour and M Royer. Fiche de td 206 : Croisement de deux variables quantitatives. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [2] AB Dufour and M Royer. Fiche de td 207 : Croisement de deux variables qualitatives. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [3] AB Dufour and M Royer. Fiche de td 208 : Croisement d'une variable qualitative et d'une variable quantitative. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [4] Stéphane Champely. *Statistique vraiment appliquée au sport*. de Boeck, 2004. disponible à la BU de Lyon I sous la cote 519.5 CHA.
- [5] Stéphane CHAMPELY. Introduction à la statistique descriptive (sous R). Note de cours de l'UE de statistique L3MOS, disponible sous spiral, 2007.
- [6] Stéphane CHAMPELY. Inférence statistique, application à l'entraînement sportif. Note de cours de l'UE de statistique M1PPMR, disponible sous spiral, 2007.
- [7] Jean-Claude Porlier and Gabriel Langouet. *Mesure et statistique en milieu éducatif*. Science de l'éducation. ESF, Paris, 1998. réédition de l'édition de 1981.
- [8] Alain Rey, editor. *Le robert, dictionnaire historique de la langue française*. Dictionnaires le Robert, Paris, 1998.
- [9] D Chessel, AB Dufour, J Lobry, and S Penel. Fiche de td 11 : Premier accès au logiciel r. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [10] D Chessel, AB Dufour, and J Lobry. Fiche de td 12 : Première session de travail. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [11] J Lobry, AB Dufour, and D Chessel. Fiche de td 13 : Objets. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [12] AB Dufour and M Royer. Fiche de td 201 : Pour une introduction à la statistique descriptive. quelques manipulations dans R. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [13] AB Dufour and M Royer. Fiche de td 202 : Introduction à la statistique univariée. Variables et Descriptions générales. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [14] AB Dufour, JR Lobry, and M Royer. Fiche de td 205 : Quelques paramètres décrivant la variabilité. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [15] AB Dufour, J.R. Lobry, and D. Chessel. Fiche de cours bs02 : De la stature chez l'Homme ... à la taille des cerveaux chez les mammifères. Réversion, Régression, Correlation. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html> rubrique cours, puis introduction, 2008.
- [16] J Cohen. A power primer. *Psychological bulletin*, 112(1) :155–159, 1992.
- [17] R.D Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28(9–12), 1974.
- [18] AB Dufour and M Royer. Fiche de td 203 : Introduction à la statistique univariée. les représentations graphiques. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [19] AB Dufour and M Royer. Fiche de td 204 : Solutions des exercices de la fiche tdr203. Disponible sur <http://pbil.univ-lyon1.fr/R/enseignement.html>, 2007.
- [20] Emmanuel Paradis. R pour les débutants. disponible sur internet : <http://www.r-project.org/>, puis rubrique Manuals, puis contributed documentation, puis Non-English Documents, puis French, puis "R pour les débutants" by Emmanuel Paradis, the French version of "R for Beginners" (PDF)", ou alors directement sur [http://cran.r-project.org/doc/contrib/Paradis-rdebuts\\_fr.pdf](http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf), 2005.
- [21] F.J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27 :17–21, 1964.