Université Claude Bernard Lyon 1

NOTES DE COURS DE STATISTIQUES

INTRODUCTION À LA STATISTIQUE INFÉRENTIELLE

Formation : M2IGAPAS

UE : STATISTIQUE

2015-2016, Automne

Jérôme BASTIEN

Document compilé le 23 mai 2016

Identification Apogée

Matière	Statistiques
Formation	Master 2 I.F.A.P.SouA (P)
Formation (code)	SPM208
UE	3MIGAPAS1 Statistique
UE (code)	SPT3023M

Table des matières

Avant-propos

Ces notes de cours constituent un support de cours, TD et TP de Statistiques pour l'UE Statistique du M2IGAPAS (2015-2016, Automne). Chacun des chapitres de ce poly s'inspire de polycopiés déjà existant réalisés par des statisticiens de l'université Lyon I, qui seront cités en début de chapitre (avec les URL où leurs polycopiés 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 polycopié 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 '`M2IGAPAS`'.

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

Vous trouverez

- au chapitre 6, l'essentiel (et l'exigible aux examens !) des notions, définitions, propriétés, exercices et manipulations avec R qu'il faut savoir (ou retrouver dans le polycopié de cours) ;
- en annexe A, un petit guide d'installation du logiciel R 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 ;
- en annexe J, l'essentiel (et l'exigible aux examens !) des notions et commandes avec R.

Une référence à consulter : [**7**] (voir bibliographie en page 205).

Pour ceux qui utilisent le package Rcmdr sur les ordinateurs [1] 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 R 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 ! ♢

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

CHAPITRE 1

Révisions de statistiques descriptives

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.

Ce chapitre, traité en un ou deux séances, vous permettra de réviser les notions de statistiques univariées et bivariées.

Consulter les annexes D, E, G, H, I, et l'annexe récapitulative J.

À titre de révisions, vous pourrez aussi traiter l'ancien projet, qui ne sera plus étudié cette année (cf annexe K).

CHAPITRE 2

Créer ses propres fichiers de données

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

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 2.1. Quelques données

Considérons les données très simples du tableau 2.1. Les données manquantes correspondent à des cases vides ou contenant la chaîne de caractère "NA".

Pour créer un fichier de données lisible par R 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 2.1 ;

(4) enregister 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 2.1. Ces manipulations peuvent aussi être faites avec excel, la sauvegarde se faisant alors au format `xls`.

REMARQUE 2.2. OpenOffice est libre, gratuit et téléchargeable sur internet.
Voir `http://fr.openoffice.org/`

Une fois que le fichier est prêt, il faut alors l'enrergister sous un format que R 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 R :

- *Avec Rcmdr :*

  Pour que cette manipulation fonctionne, il est nécessaire que les données manquantes soient codées par la chaîne "NA" et non par des cases vides.

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis un fichier texte ou le presse-papier...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier texte, ici, par exemple `dudu`. pour le Séparateur de champ, choisir "Autre" et Spécifier ";". Pour le Séparateur Décimal, choisir ",". Laisser les autres champs avec les valeurs choisies par défaut.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier à importer ;

(3) Cliquer alors éventuellement sur le bouton "Visualiser".

- *Sans Rcmdr :*

  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 2.3. On peut aussi faire au export au format `xls`, mais R ne gère pas toujours ce format correctement. Voir remarque D.1.

REMARQUE 2.4. Il existe des tas de solutions différents : exports au format texte, csv, excel... On peut aussi le faire grâce à Rcmdr.Seule une méthode est ici présentée.

CHAPITRE 3

Introduction aux probabilités

On pourra consulter [**6**], le chapitre 10 de [**4**] ainsi que le chapitre 3 de [**7**], dont s'inspire ce chapitre.

3.1. Introduction

Il est choisi de faire le moins de mathématiques possible!!

Une expérience aléatoire (jeté de dès, choix d'une carte dans un jeu, mesure de la taille d'un étudiant) a par définition un résultat que l'on ne peut prévoir. En revanche, sur un grand nombre d'expériences, il sera possible, de calculer la proportion d'apparition des résultats, grâce aux notions de probabilité et fréquences.

3.2. Notions de probabilités

3.2.1. Définitions

Si on jette en l'air, une pièce non truquée, nous avons une chance sur deux de faire "pile". On dit que la probabilité de faire pile est de $p = 1/2 = 0.5$ Si non lançons, un dés non truquée, la probabilité de faire quatre est de $1/6$. La probabilité de tirer l'as de pique dans un jeu de 32 cartes est de $1/32$.

DÉFINITION 3.1. On parle d'expérience aléatoire, s'il est impossible d'en prévoir l'issue mais qu'en revanche, sur un grand nombre de répétitions, on peut connaître la fréquence (ou la proportion) avec laquelle les différents résultats apparaitront.

La probabilité d'un événement se définit donc comme le rapport entre le nombre de cas favorable à l'apparition de cet événement et le nombre total de cas possibles. C'est un nombre compris entre 0 et 1.

La somme des probabilités de toutes les éventualités possibles est de 1. La probabilité d'un événement impossible est nulle, celle d'un événement certain est 1.

Par exemple, nous jetons un dé à six faces ; la probabilité d'obtenir :

- un 2 est $1/6$;
- un nombre pair est est $1/6 + 1/6 + 1/6 = 1/2$;
- un nombre égal à 1 et 3 et 4 et 5 et 6 est nulle ;
- un nombre égal à 1 ou 3 ou 4 ou 5 ou 6 est égale à 1.

3.2.2. Probabilités et fréquences

3.2.2.1. "Simulation" avec R.

Ceux qui souhaitent utiliser Rcmdr doivent s'habituer à taper les commandes données directement dans la fenêtre de "Rgui", puisqu'elle n'existent pas dans Rcmdr.

EXERCICE 3.2. Taper les lignes suivantes dans la fenêtre de script et soumettez les successivement.

```
 1:6
```

```
[1] 1 2 3 4 5 6
```

```
 sample(1:6, size = 1, replace = T)
```

```
[1] 2
```

```
sample(1:6, size = 10, replace = T)
```

```
[1] 4 6 6 5 4 2 6 1 4 3
```

```
dede <- sample(1:6, size = 10, replace = T)
```

Ici `dede` est le nom de la variable dans laquelle on stocke un résultat ; on aurait pu l'appeler autrement. Cette dernière ligne permet de figer le résultat donné par `sample`.

```
dede
```

```
[1] 3 3 5 5 6 2 4 4 2 3
```

```
table(dede)
```

```
dede
2 3 4 5 6
2 3 2 2 1
```

```
table(sample(1:6, size = 10000, replace = T))/10000
```

```
     1      2      3      4      5      6
0.1656 0.1644 0.1678 0.1668 0.1659 0.1695
```

```
table(sample(1:6, size = 1e+05, replace = T))/1e+05
```

```
      1       2       3       4       5       6
0.16962 0.16548 0.16699 0.16543 0.16821 0.16427
```

Attention, puisque des fonctions aléatoires interviennent, elles fournissent donc des valeurs différentes à chaque appel (et donc *a priori* différent de ce qui est écrit en bleu ici !).

Voir éléments de correction page 25.

Exercice 3.3 (facultatif).

(1) Quelle ligne de commande vous permet de simuler un tirage d'une grille de loto ?

(2) Comment simuler un nombre entier quelconque $n \geq 1$ grilles de loto ?

(3) Pourriez-vous retirer un enrichissement personnel (au sens propre du terme) de cet exercice ?

Voir éléments de correction page 25.

3.2.2.2. Définitions.

Définition 3.4. Si une expérience est aléatoire, la probabilité d'un événement est la "limite" quand la série est "longue" de la proportion de fois (les fréquences) où il se réalise.

Supposons par exemple que l'on considère le jetté de dé et que l'on compte les fréquences d'apparition du 1 ; par exemple,

- pour 10 lancers, on peut obtenir 2 fois le 1 ;
- pour 100 lancers, on peut obtenir 16 fois le 1 ;
- pour 1000 lancers, on peut obtenir 157 fois le 1.

Les nombres 2, 16, 157 sont les réalisations de l'événement "tirer le 1". Les rapports 2/10, 16/100, 157/1000 sont les proportions (ou les fréquences) et se rapprochent de la probabilité théorique 1/6. Attention, cette limite est théorique. On peut très bien obtenir, un jour, 500 fois la face 1 sur 1000 lancers (mais c'est très rare !).

Voir sur la figure 3.1 page ci-contre une simulation faite avec R. Sur cette figure, on peut voir six courbes, correspondant aux proportions expérimentales d'apparition de chacun des numéros, qui se rapprochent de 1/6.

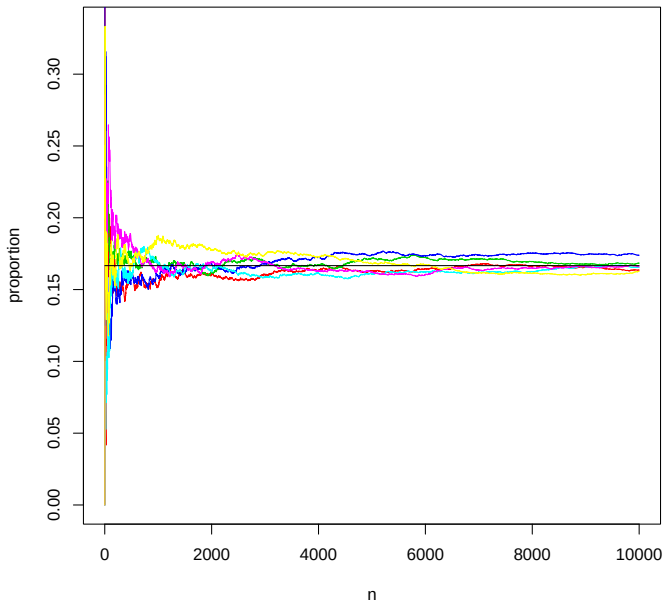


FIGURE 3.1. Une simulation d'un jeté de dès pour $n = 10000$.

3.2.2.3. Variables aléatoires.

DÉFINITION 3.5. On parle de *variable aléatoire* lorsqu'on associe à des résultats d'expérience aléatoire des valeurs x (numériques ou non).

Dans le cas où l'ensemble des valeurs possibles est fini, on parle de *variable aléatoire discrète*. La liste des probabilités associées aux résultats possibles d'une variable aléatoire est appelée *loi de probabilités*. On note $P(X = x)$ la probabilité qu'une variable aléatoire X prenne la valeur x.

Si au contraire, une variable aléatoire peut prendre toutes les valeurs dans un intervalles, on parle de *variable aléatoire continue*.

EXEMPLE 3.6. Dans le cas de l'exercice 3.2 qui simule un tirage au dès, l'ensemble des valeurs possible pour un jeté de dès est $\{1, 2, 3, 4, 5, 6\}$. X est ici la variable aléatoire discrète égale au numéro de la face obtenue et on a

$$P\big(X = 1\big) = P\big(X = 2\big) = \ldots = P\big(X = 6\big) = \frac{1}{6}. \tag{3.1}$$

De façon plus générale, on donne la définition suivante :

DÉFINITION 3.7. Pour tout entier n non nul, nous dirons que la variable aléatoire X suit une loi uniforme si

$$\forall k \in \{1, \ldots, n\}, \quad P(X = k) = \frac{1}{n}. \tag{3.2}$$

0n notera

$$X \rightsquigarrow \frac{1}{n} \sum_{k=1}^{n} \delta_k. \tag{3.3}$$

♢

EXEMPLE 3.8. Supposons maintenant que l'on s'intéresse aux tailles d'un ensemble d'étudiants. *A priori*, les tailles peuvent prendre toutes les valeurs possibles, même si en pratique, elles sont arrondie à l'unité. On

considère l'expérience aléatoire qui consiste à prendre la taille d'un étudiant (issu d'un groupe défini à l'avance). X est ici la variable aléatoire continue égale à la taille.

3.2.2.4. Notions de fonction de densité (ou densité de probabilité).

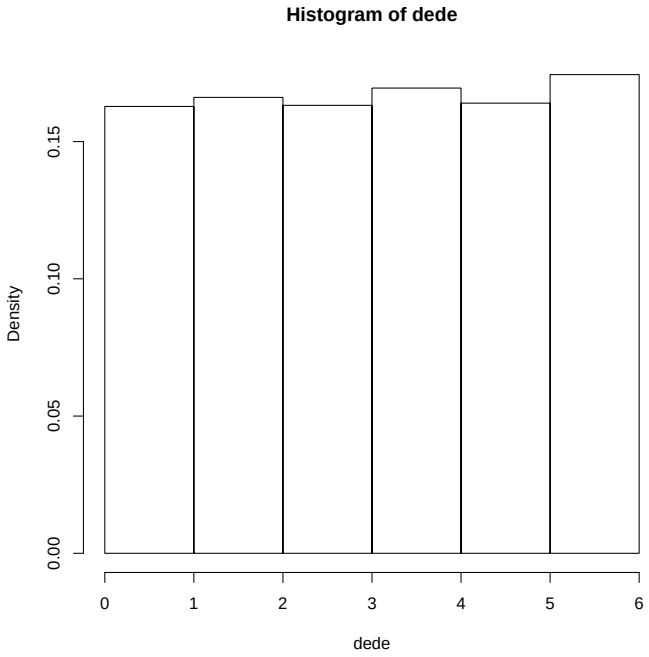


FIGURE 3.2. Histogramme de $n = 10000$ tirages de dé.

Dans le cas de l'exercice 3.2, on peut tracer l'histogramme des résultats obtenus pour $n = 10000$, en tapant par exemple

```
dede<-sample(1:6,size=10000,replace=T)
hist(dede,freq=F,breaks=0:6)
```

On obtient par exemple la figure 3.2.

REMARQUE 3.9. Pour tracer les histogrammes, on a le choix entre
- fréquence (nombre d'unités statistiques par classe) ;
- pourcentage (nombre d'unités statistiques par classe divisée par 100) ;
- densité (fréquence divisée par le produit de la largeur de la classe par le nombre total d'individu).

Ici, il y a peu de différences entre ces trois possibilités. Retenez que les histogrammes en densité sont plus "stables" par rapport aux nombre de classes choisie. De plus, l'histogramme en densité est "normalisé", c'est-à-dire que son aire totale est égale à 1 et il pourra être ainsi comparé à des lois théoriques de probabilité.

Puisque la variable est discrète et prend un nombre fini de valeurs, on peut remplacer l'histogramme par la courbe de fréquence (ou de probablité expérimentale) obtenue en tapant

```
plot(as.numeric(table(dede))/10000,ylim=0:1)
```

On obtient la courbe de la figure 3.3 page ci-contre.

Traçons l'histogramme d'une variable aléatoire continue, égale par exemple à la taille d'un individu. Pour ce faire,

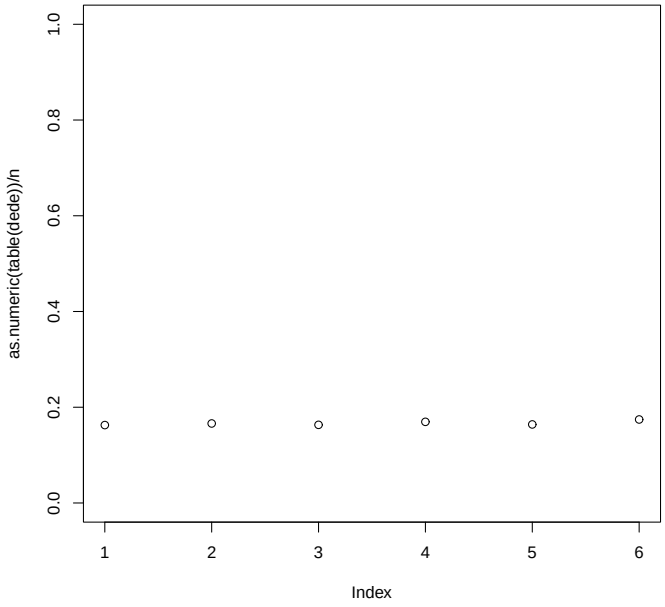


FIGURE 3.3. Courbe de fréquence de la variable aléatoire "numéro de la face obtenue" pour $n = 10000$.

- *Avec Rcmdr :*

  Il faut choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Echantillon d'une distribution normale". Dans la fenêtre de dialogue qui s'ouvre il faut indiquer pour
  - "mu" : 171,
  - "sigma" : 8.7,
  - "Nombre d'échantillons" : 10000,
  - "Nombre d'observations" : 1

  Il se crée alors un tableau à 10000 lignes et 1 colonne.

  On trace ensuite un histogramme en densité avec 100 classes.
- *Sans Rcmdr :*

  On tape dans "Rgui"

```
dede<-rnorm(10000, mean = 171, sd = 8.7)
hist(dede,breaks=100,freq=F)
```

  Ici la variable `breaks` correspond au nombre de classes choisi.

On obtient la figure 3.4 page suivante.

Sur cette figure, à cause du grand nombre de polygone, l'histogramme semble se rappprocher d'une courbe continue (tracée en rouge sur la figure). Cette courbe est la représentation d'une loi idéale de disribution, appelée la fonction de densité.

Cette courbe peut être alors considérée comme la représentation géométrique d'une loi idéale de distribution. On parle de fonction de densité. Par définition, l'aire totale située entre la cette courbe et l'axe des x est égale à 1 (somme des probabilité). Sur la figure 3.5 page 11, la portion d'aire comprise entre les abscisses 177 et 189, représentée en rouge, est la probabilité qu'une valeur issue de cette loi soit située dans l'intervalle

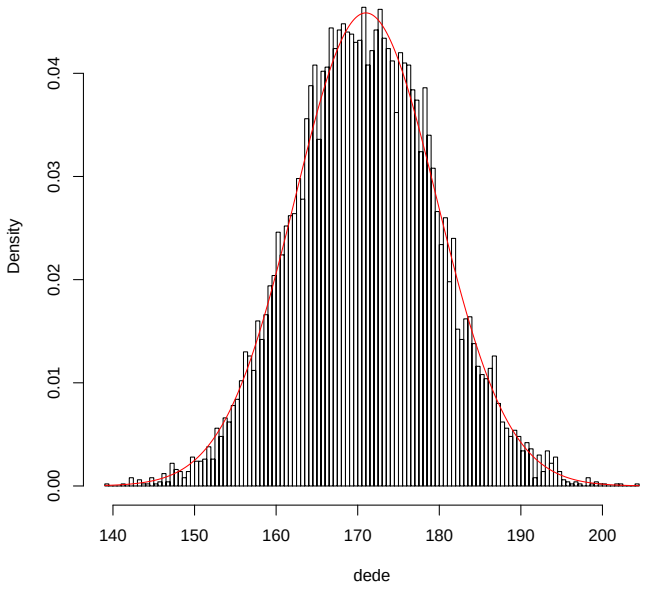


FIGURE 3.4. Histogramme des $n = 10000$ tailles.

$[177, 189]$. La portion d'aire en deça de l'abscisse 164, représentée en bleue, est la probabilité qu'une valeur issue de cette loi soit inférieure à 164.

3.2.3. Espérance et écart-type d'une variable aléatoire

On considère une variable aléatoire X, discrète ou continue.
La notion d'espérance et écart-type peut être approchée de manière expériementale de la façon suivante :

DÉFINITION 3.10. On considère un échantillon formé d'un grand nombre de réalisations de la variable aléatoire X. On calcule la moyenne et l'écart-type de cet échantillon. Ces deux nombres doivent se rapprocher respectivement d'un nombre appelé moyenne de la variable aléatoire X et d'un nombre appelé écart-type de la variable aléatoire X. Ils sont respectivement noté $\mathbb{E}(X)$ et σ.

EXEMPLE 3.11. Reprenons l'exemple de l'exercice 3.2 page 5. Quels valeurs semble-t-on obtenir en tapant

```
n <- 10000
dede <- sample(1:6, size = n, replace = T)
mean(dede)
```

```
[1] 3.4823
```

```
sd(dede)
```

```
[1] 1.710257
```

Deux autres définitions théoriques peuvent être données :

DÉFINITION 3.12. L'*espérance* d'une variable aléatoire discrète X est la somme des résultats possibles pondérés par leurs probabilités. On la note généralement $\mathbb{E}(X)$.

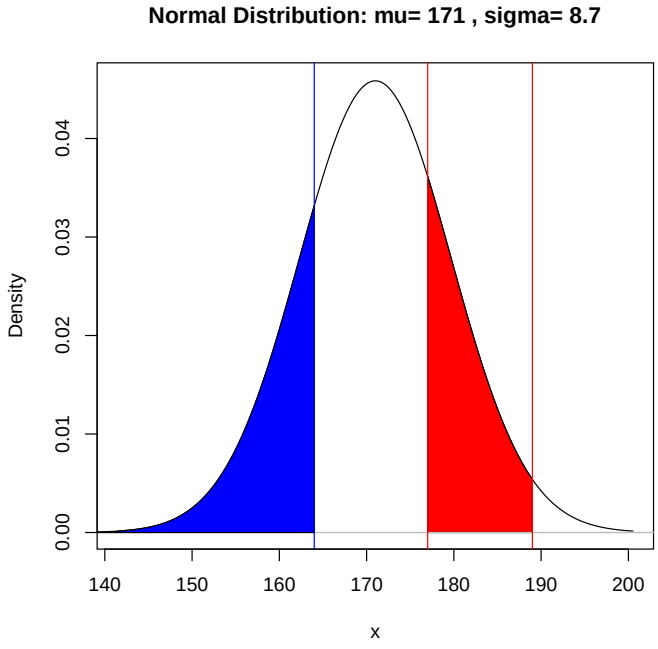


FIGURE 3.5. Une fonction de densité p

Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$, alors

$$\mathbb{E}(X) = P(X = n_1)n_1 + P(X = n_2)n_2 + \ldots + P(X = n_q)n_q = \sum_{i=1}^{q} P(X = n_i)n_i \tag{3.4}$$

L'*écart-type* d'une variable aléatoire discrète X est la racine carrée de la somme des carrés des écarts entre les résultats possibles et l'espérance, pondérée par les probabilités respectives.

Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$, alors

$$\sigma = \sqrt{P(X = n_1)(n_1 - \mathbb{E}(X))^2 + \ldots + P(X = n_q)(n_q - \mathbb{E}(X))^2} = \sqrt{\sum_{i=1}^{q} P(X = n_i)(n_i - \mathbb{E}(X))^2} \tag{3.5}$$

La moyenne et l'écart-type d'une variable aléatoire continue X de loi de densité p sont données par

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx,$$
$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx}.$$

Une preuve de l'équivalence de cette définition et de la précédente peut être trouvée en annexe O. ♢

EXERCICE 3.13 (facultatif). Retrouver grâce à la définition 3.12 page ci-contre les valeurs obtenues dans l'exemple 3.11 page précédente, en montrant que

$$\mathbb{E}(X) = \frac{7}{2}, \tag{3.6a}$$
$$\sigma \approx 1.707825. \tag{3.6b}$$

Voir éléments de correction page 26.

3.3. La distribution binomiale

3.3.1. À partir d'un exemple

Considérons un établissement scolaire, pour lequel la proportion de garçons est p et celle de fille q, avec $p+q=1$.

Si nous tirons au sort, la probabilité d'obtenir un lycéen est de p, une lycéenne de q.

Supposons maintenant que l'on fasse deux tirages indépendant : on choisit une première fois un élève, on le remet dans l'urne (pour que la proportion soit conservée) et on choisit de nouveau un éléve. Les différents événements possibles sont (G,G), (F,G), (G,F) et (F,F). On applique la règle selon laquelle les probabilités d'événements indépendants sont le produits des probabilités. La probabilité d'obtenir :

- (G,G) est $p \times p = p^2$;
- (F,G) est $p \times q = pq$;
- (G,F) est $q \times p = pq$;
- (F,F) est $q \times q = q^2$.

Si on ne s'intéresse plus à l'ordre mais seulement à la répartition par rapport au sexe des couples formés, les différents événements et leur probabilités sont

- deux garçons ((G,G)), de probabilité p^2 ;
- un garçon et une fille ((F,G) ou (G,F)), de probabilité $pq+pq=2pq$;
- deux filles ((F,F)), de probabilité q^2.

On remarque que

$$1^2 = (p+q)^2 = p^2 + 2pq + q^2,$$

autrement dit que la somme des différentes probabilités est de 1 et que la probabilité des différents événements est donné par le développement de $(p+q)^2$. Autrement dit la probabilité d'obtenir

- deux garçons est p^2 ;
- un garçon est $2pq = 2p(1-p)$;
- zéro garçon est $q^2 = (1-p)^2$.

Si nous prenons un tirage de trois élèves (avec remise), la probabilité d'obtenir 0, 1 , 2 ou 3 garçons serait de même obtenue en développant $(p+q)^3$

$$1^2 = (p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3,$$

soit la probabilité d'obtenir

- 3 garçons est p^3 ;
- 2 garçons est $3p^2q = 3p^2(1-p)$;
- 1 garçons est $3pq^2 = 3p(1-p)^2$;
- zéro garçon est $q^3 = (1-p)^3$.

De façon plus générale, on suppose que l'on réalise une expérience à deux issues, succès (ici obtenir un garçon) avec une probabilité p ou échec (une fille) avec une probabilité q. On réalise ensuite, de façon indépendante, n fois cette expérience élémentaire et on s'intéresse aux probabilités d'obtenir 0, 1, 2, ..., n succès qui sont égales aux termes successifs du développement de $(p+q)^n$.

3.3.2. La théorie

Définition 3.14. Le modèle probabiliste *binomial* correspond à des circonstances où

- il n'y a que deux résultats possibles dans une unique expérience aléatoire : garçon/fille, pile/face, correct/défectueux, mort/vivant ; l'un est considéré comme un succès et l'autre comme un échec [1],
- les répétitions de cette expérience se réalisent indépendamment les unes des autres et
- la probabilité de succès reste la même à chaque répétition.

1. Sans qu'il y ait pour autant jugement de valeur...

Le modèle probabiliste binomial décrit le comportement de la variable aléatoire "nombre de succès apparus" sur l'ensemble des répétitions.

La famille binomiale est engendrée par deux *paramètres* : le nombre de répétitions et la probabilité de succès, notés respectivement n et p.

Définition 3.15. Nous dirons que la variable aléatoire X suit une loi binomiale de paramètres n et p et on notera

$$X \rightsquigarrow \mathcal{B}(n,p) \tag{3.7}$$

Exercice 3.16. Quels sont les paramètres correspondant au nombre de "pile" dans un lancer de pièce cinq fois de suite ? Quels sont les paramètres correspondant à la roulette russe ?

Voir éléments de correction page 26.

Proposition 3.17. *Pour n répétitions avec une probabilité de succès p, la probabilité qu'une variable aléatoire binomiale*[2] *X soit égale à $k \in \{0,\dots,n\}$ est*

$$P(X=k) = C_n^k p^k (1-p)^{n-k} \tag{3.8}$$

avec C_n^k qui est appelé coefficient binomial.

Remarque 3.18 (facultative). Dans la proposition 3.17, on a

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

Preuve facultative. La proposition 3.17 peut se "montrer" (en simplifiant) de la façon suivante : pour obtenir x succès et $n-x$ échecs dans un ordre donné, la probabilité est de $p^k(1-p)^{n-k}$ (on multiplie les diférentes probabilités de chacun des événements, indépendants). On multiplie ce résultas par C_n^k qui correspond au nombre total de combinaisons de k éléments parmi n, puisqu'ici l'ordre des succès ne compte pas. □

Manipulation avec R 3.19. On peut calculer cette loi de probabilité :

- *Avec Rcmdr :*

  En utilisant le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis "Probabilités binomiales". Dans la fenêtre de dialogue, il faut alors préciser le nombre d'essais (n) et la probabilité de succès (p).

  Ainsi pour $n=5$ essais et une probabilité de succès de $p=0.3$, on obtient

```
      Pr
0 0.16807
1 0.36015
2 0.30870
3 0.13230
4 0.02835
5 0.00243
```

  et donc en particulier $P(X=2) = 0.3087$.

2. ce nom vient de la formule du binôme de Newton :

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}.$$

On a un moyen "mnémotechnique" de se rappeler (3.8) :

$$g(X) = (p+1-p)^n = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} =$$

Cela nous montre de plus que la somme des probabilité est bien 1.

- *Sans Rcmdr :*
  On tape dans "Rgui" pour $n = 5$ et $p = 0.3$
  ```
  dbinom(0:5, size=5, prob=0.3)
  ```
  et on obtient alors les différentes probablités correspondant à 0, 1, ..., n succès. On peut aussi taper directement pour k dans $\{0, ..., n = 5\}$
  ```
  dbinom(k, size=5, prob=0.3)
  ```
  Ainsi pour $n = 5$ essais et une probabilité de succès de $p = 0.3$, on obtient et donc en particulier $P(X = 2) = 0.3087$.

On peut aussi tracer le graphe de la loi de probabilité : Sur ce graphe, pour chaque k dans $0, \ldots, n$, on trace un segment d'abscisse k et de hauteur $P(X = k)$.

Manipulation avec R 3.20.
- *Avec Rcmdr :*
  On utilise l'option "Graphe de la distribution binomiale".
- *Sans Rcmdr :*
  On tape dans "Rgui"
  ```
  x <- 0:n
  plot(x, dbinom(x, size = n, prob = p), type = "h")
  points(x, dbinom(x, size = n, prob = p), pch = 16)
  abline(h = 0, col = "gray")
  ```
  où $n = 5$ et $p = 0.3$.

Exercice 3.21. Observer comment évolue le graphe de la distribution binomiale évolue lorsque vous modifiez la probabilité de succès en $p = 0$, $p = 0.4$, $p = 0.5$, $p = 0.7$, $p = 0.9$, $p = 0.95$ et $p = 1$ en conservant le paramètre $n = 5$.

Voir éléments de correction page 26.

Exercice 3.22. Représenter graphiquement les lois de probabilités correspondant aux jeux de paramètres suivants : $(n = 10, p = 0.1)$, $(n = 10, p = 0.25)$, $(n = 10, p = 0.5)$, $(n = 10, p = 0.85)$.

Voir éléments de correction page 26.

3.3.3. Espérance et écart-type binomiaux

Proposition 3.23. *L'espérance d'une variable aléatoire binomiale de paramètres n et p est égale à np et son écart-type à $\sqrt{np(1-p)}$.*

Preuve facultative. Voir la preuve de ce résultats en annexe P. □

Preuve partielle "avec les mains". On peut remarquer formellement et facilement que l'espérance d'une variable aléatoire binomiale de paramètres n et p est égale à np.

En effet, pour un essai, la proportion de succès est égale à p ; pour n essais, la proportion de succès est égale au nombre moyen de succès (soit $\mathbb{E}(X)$) divisée par le nombre d'essais (n) ; bref, $p = \mathbb{E}(X)/n$, d'où le résultat ! □

Afin de calculer l'espérance, la variance et l'écart-type pour une distribution binomiale comportant $n = 5$ essais et une probabilité de succès de $p = 0.3$, il faut écrire dans la fenêtre de script les trois commandes suivantes et les soumettre successivement :

```
5*0.3
5*0.3*(1-0.3)
sqrt(5*0.3*(1-0.3))
```

On obtient donc

```
[1] 1.5
```

```
[1] 1.05
```

```
[1] 1.024695
```

c'est-à-dire

$$\mathbb{E}(X) = 1.5, \quad \sigma = 1.024695 \tag{3.9}$$

EXERCICE 3.24. Trouver l'espérance et l'écart-type pour les distributions binomiales suivantes :

- $n = 20$ et $p = 0{,}50$;
- $n = 40$ et $p = 0{,}20$;
- $n = 200$ et $p = 0{,}80$.

Voir éléments de correction page 27.

3.3.4. Les probabilités cumulés

Il est souvent utile de calculer des probabilités binomiales cumulées, par exemple la probabilité d'observer 2 succès ou moins ou bien la probabilité d'observer plus de 4 succès.

DÉFINITION 3.25. La *probablité cumulée* jusqu'au *quantile* $i \in \{0, ..., n\}$ correspond à la somme

$$P(X = 0) + P(X = 1) + \ldots + P(X = i)$$

On notera cette somme

$$P(X \leq i)$$

MANIPULATION AVEC R 3.26.

Pour ce faire,

- *Avec Rcmdr :*

  Il est nécessaire d'employer le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis l'option "Probabilités binomiales cumulées" (en prenant par défaut l'aire à gauche)
- *Sans Rcmdr :*

  On tapera dans "Rgui"

```
pbinom(k,size=5,prob=0.3)
```

De façon plus générale,

DÉFINITION 3.27. Si l'ensemble des valeurs prise par la variable aléatoire X est noté $\{n_1, n_2, \ldots, n_q\}$ (*ici, dans un ordre croissant*), alors la *probablité cumulée* jusqu'au *quantile* n_i pour $i \in \{1, ..., q\}$ correspond à la somme

$$P(X = n_i) + P(X = n_{i+1}) + \ldots + P(X = n_q) \tag{3.10}$$

On notera cette somme

$$P(X \leq n_i) \tag{3.11}$$

On note de même

$$P(X > n_i) = P(X \geq n_{i+1}) = P(X = n_{i+1}) + P(X = n_{i+2}) + \ldots + P(X = n_q) \tag{3.12}$$

REMARQUE 3.28. Notons aussi que, de façon plus générale, si on se donne la probabilité p, le quantile est le nombre q tel que

$$P(X \leq q) = p.$$

On choisira parfois la définition (équivalente ici) : si on se donne la probabilité p, le quantile est la plus petite valeur x telle que

$$P(X \leq x) \geq p.$$

♢

REMARQUE 3.29. On a naturellement

$$P(X \leq n_i) + P(X > n_i) = 1 \tag{3.13}$$

MANIPULATION AVEC R 3.30.

- *Avec Rcmdr :*

  Avec Rcmdr, dans l'option "Probabilités binomiales cumulées", la somme (3.11) est appelée "aire à gauche" tandis que la somme (3.12) est appelée "aire à droite".

  Plus précisémment, on utilise le menu déroulant "Distributions" , l'option "Distributions discrètes" puis "Distribution binomiale" puis l'option "Probabilités binomiales cumulées", puis en choisissant "aire à gauche" ou "aire à droite".
- *Sans Rcmdr :*

  Pour calculer la somme (3.11) (l'aire à gauche), on tapera

```
pbinom(ni, size = n, prob = p)
```

  tandis que pour calculer la somme (3.12) (l'aire à droite), on taper

```
pbinom(ni, size = n, prob = p, lower.tail = FALSE)
```

On peut aussi tracer le graphique des probabilités cumulées :

- *Avec Rcmdr :*

  On utilise l'option "Graphe des probabilités cumulées".
- *Sans Rcmdr :*

  On tape dans "Rgui"

```
x <- 0:n
x <- rep(x, rep(2, length(x)))
plot(x[-1], pbinom(x, size = n, prob = p)[-length(x)], type = "l")
```

On a aussi le résultat suivant

LEMME 3.31. *Si pour $i < j$, on note*

$$P(n_i \leq X \leq n_j) = P(X = n_i) + P(X = n_{i+1}) + \ldots + P(X = n_j), \tag{3.14}$$

alors

$$P(n_i \leq X \leq n_j) = P(X \leq n_j) - P(X \leq n_{i-1}). \tag{3.15}$$

EXERCICE 3.32. Pour une variable aléatoire binomiale X de paramètres $n = 7$ et $p = 0.2$

(1) Calculer à l'aide du logiciel R : $P(X = 2)$, $P(X = 0)$, $P(X = 9)$, $P(X \leq 5)$, $P(X \geq 5)$, $P(X > 5)$, et $P(2 \leq X \leq 5)$

(2) Vérifier que la probabilité cumulée $P(X \leq 5)$ est bien égale à

$$P(X \leq 5) = P(x = 0) + \ldots + P(x = 5)$$

(3) Vérifier que la probabilité cumulée $P(X > 5)$ est bien égale à

$$P(X > 5) = P(X = 6) + P(X = 7)$$

(4) Représenter par un graphique sa loi de probabilités cumulées.

Voir éléments de correction page 27.

EXERCICE 3.33. La probabilité de contact des sondés au téléphone est généralement estimée à 60 %. Nous décidons de lancer une vague d'appels de $n = 200$ personnes. On considérera que le nombre de personnes que l'on va parvenir à contacter suit une loi binomiale de paramètres $n = 200$ et $p = 0.6$. En effet, dans ce cas, on répète $n = 200$ fois l'expérience "appeler quelqu'un au téléphone" qui a deux issues : le succès est la prise de contact de probalité $p = 0.6$ et l'échec est la non prise de contact.

(1) Quel nombre moyen de personnes peut-on espérer toucher dans cette première vague d'appels ?

(2) Quelle est la probabilité de contacter au moins 120 personnes ?

(3) Quelle est la probabilité de contacter au moins $n_1 = 150$ personnes ?

(4) Combien de personnes faudrait-il appeler pour espérer, en moyenne, contacter 150 personnes ?

Voir éléments de correction page 29.

3.4. La distribution normale

3.4.1. Introduction

La loi normale intervient très fréquemment en probabilité et statistique. Contrairement à la distribution binomiale, associée à une variable aléatoire discrète, elle est associée à une variable aléatoire continue.

Nous l'introduisons de deux façon différentes, à partir de la loi binomiale (section 3.4.2.1) ou à partir d'une population (section 3.4.2.2)

3.4.2. Deux approches possibles

3.4.2.1. Par la "limite" d'une loi binomiale.

Considérons la loi binomiale de paramètres $p = 0.3$ et pour des valeurs de n "de plus en plus grandes". Si on tape par exemple dans "Rgui" la séquence (ou que l'on utilise, pour ceux qui utilisent Rcdmr, la manipulation donnée page 14)

```
p<-0.3
n<-100
x<- 0:n
plot(x, dbinom(x, size=n, prob=p))
points(x, dbinom(x, size=n, prob=p), pch=16)
```

et que l'on donne des valeurs de n de plus en plus grandes, la courbe de probabilité (voir figure 3.6 page suivante) semblera se rapprocher d'une loi continue.

On peut montrer sur le plan théorique que cette distribution est la loi normale que l'on va définir théoriquement.

Pour une approche plus précise, on pourra consulter l'annexe Q.

3.4.2.2. Par une population "normale".

Vous pouvez charger le fichier '`studenth.txt`' et tracer l'histogramme, en densité (voir remaque 3.9 page 8) et le graphe quantile-quantile des données '`Taille`'. On obtient les deux graphes de la figure 3.7 page 19. Le graphe quantile-quantile nous montre un bon accord avec la loi normale.

Sur la figure 3.8 page 20, nous avons rajouté en rouge, la loi normale idéale, loi théorique en forme de cloche. Cette loi théorique est centrée sur la moyenne de l'échantillon (tracée en bleu en pointillé sur la figure) et présente, dans un grand nombre de cas, un phénomène idéal de répartition de données.

En fait, les données sont dites normales, mais on verra que ce sont leurs moyennes sur un "grand nombre" de valeurs qui le sont réellement. Nous reviendrons longuement sur ces moyennes au cours du chapitre. ♢

On pourra aussi relire la section 3.2.2.4 qui montre que la loi continue peut être approchée par une histogramme d'un échantillon, extrait de cette population, de plus en plus grand. Nous constaterons aussi cela de façon expérimentale au cours de l'annexe N.

3.4.3. Définitions théoriques

Définition 3.34. En probabilité, une variable aléatoire suit une loi normale (ou loi normale gaussienne ou loi de Laplace-Gauss) d'espérance $\mu \in \mathbb{R}$ et d'écart-type $\sigma > 0$ si elle admet une densité de probabilité p

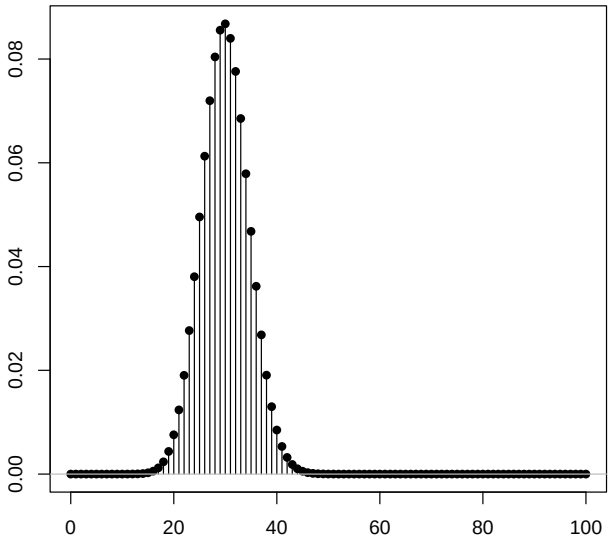


FIGURE 3.6. La loi de probabilité (discrète) binomiale $n = 100$ et $p = 0.3$.

telle que :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{3.16}$$

Nous dirons que la variable aléatoire X suit une loi normale de moyenne μ et d'écart-type σ et on notera

$$X \rightsquigarrow \mathcal{N}(\mu, \sigma) \tag{3.17}$$

Attention, on voit aussi la notation, non utilisée dans ce cours,

$$X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$$

♢

Pour tracer le graphe de cette fonction :

- *Avec Rcmdr :*

  Choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Graphe de la distribution normale". Il préciser dans la fenêtre de dialogue les valeurs de *mu* et de *sigma* (par défaut égaux respectivement à 0 et 1).
- *Sans Rcmdr :*

  Il faut télécharger la fonction `trace.loi.normale.R` et taper

  ```
  trace.loi.normale(mu, sigma)
  ```

  ou encore

  ```
  trace.loi.normale(mu = mu, sigma = sigma)
  ```

  Voir par exemple, la figure 3.9 page 21.

REMARQUE 3.35. Pour la fonction `trace.loi.normale`, un argument optionnel égal à un intervalle (sous la forme `c(a,b)` où $-\infty \leq a \leq b \leq +\infty$) permet de rajouter sur la courbe, une aire en rouge et comprise entre

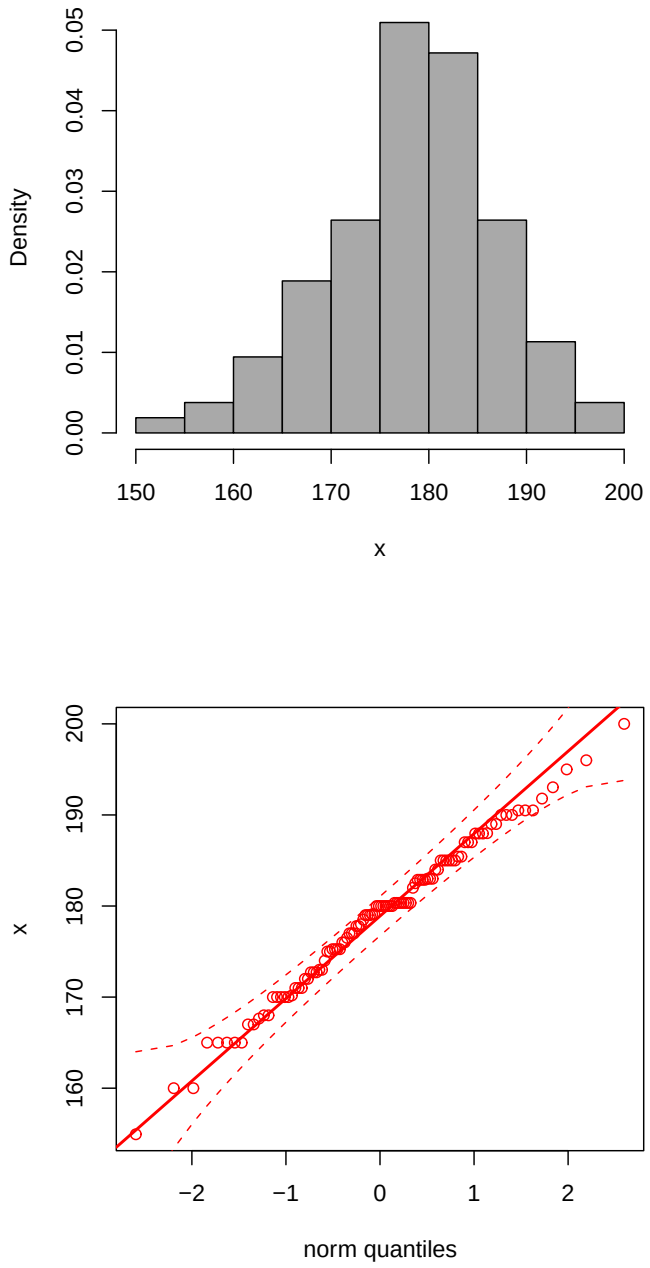


FIGURE 3.7. Histogramme (en densité) et graphe quantile-quantile sur les données de 'Taille' de 'studenth.txt'.

les abscisses a et b, ce qui représente donc la probabilité qu'une variable aléatoire X normale, appartienne à l'intervalle $[a, b]$.

Par exemple,

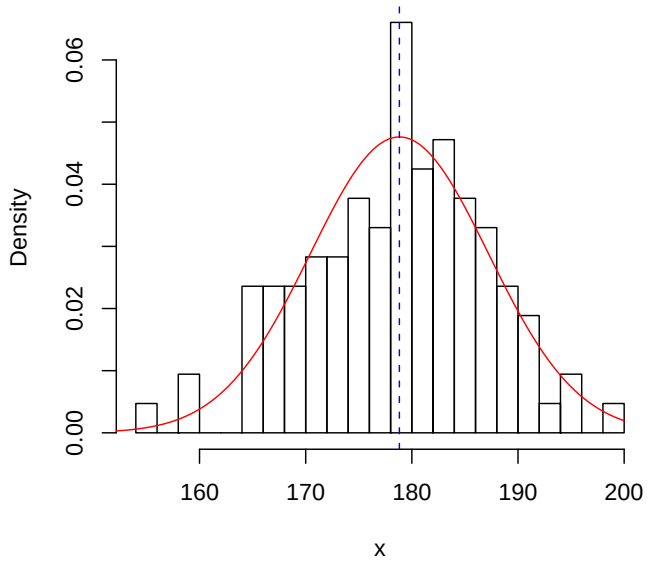


FIGURE 3.8. Histogramme (en densité et avec 18 classes) et la loi normale associée sur les données de 'Taille' de 'studenth.txt'.

```
trace.loi.normale(mu = 1, sigma = 1, aire = c(1.2, 3))
```

produirait la figure 3.10 page 22.

La moyenne μ correspond à l'abscisse du maximum de cette courbe, c'est un paramètre de position. L'écart-type σ correspond à l'étalement de la courbe ; voir la figure 3.11 page 23, où plusieurs courbes correspondant à plusieurs valeurs de σ ont été tracées.

DÉFINITION 3.36. La loi normale centrée réduite correspond au cas où $\sigma = 1$ et $\mu = 0$.

Cette fonction est donnée sur la courbe 3.12 page 24.

Cette fonction de densité est symétrique, elle atteint son maximum en zéro et est pratiquement nulle au delà de trois (en valeur absolue).

REMARQUE 3.37. On peut montrer que

$$X \rightsquigarrow \mathcal{N}(\mu, \sigma) \Longleftrightarrow \frac{X - \mu}{\sigma} \rightsquigarrow \mathcal{N}(0, 1).$$

Autrement dit on passe la densité de la loi normale centrée réduite à celle de la loi normale générale par une "dilatation-translation". Ainsi, les deux graphes des figures 3.9 et 3.12 sont identiques et surperposables. Seules les échelles sont différentes.

REMARQUE 3.38. De façon analogue à la remarque 3.28 page 15 (dans le cas discret), notons aussi que, si on se donne la probabilité p, le quantile est le nombre q tel que

$$P(X \leq q) = p.$$

c'est-à-dire

$$\int_{-\infty}^{q} p(x)dx = p.$$

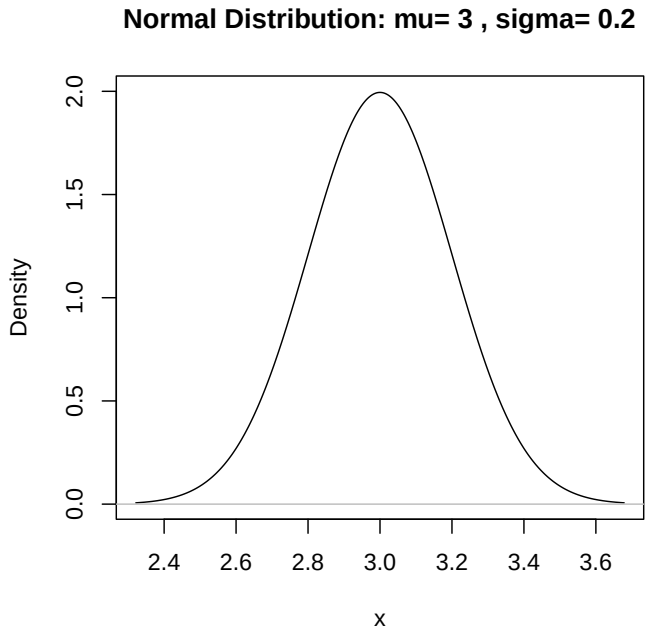


FIGURE 3.9. La loi normale avec $\mu = 3$ et $\sigma = 0.2$.

On choisira parfois la définition (équivalente ici, quand la densité est "intégrable" et de primitive continue) : si on se donne la probabilité p, le quantile est la plus petite valeur x telle que

$$P(X \leq x) \geq p.$$

♢

3.4.4. Espérance et écart-type normaux

PROPOSITION 3.39. *L'espérance d'une variable aléatoire normale de moyenne μ et d'écart-type σ valent* justement *μ et σ.*

3.4.5. Calculs de probabilités normales et applications

Si X est une variable aléatoire de fonction de densité p :

$$P(a \leq X \leq b) = \int_a^b p(x)dx \tag{3.18}$$

On rappelle que de façon analogue au cas discret, l'aire à gauche désigne :

$$P(X \leq a) = \int_{-\infty}^a p(x)dx \tag{3.19}$$

et que l'aire à droite désigne

$$P(X \geq a) = \int_a^{\infty} p(x)dx \tag{3.20}$$

et que la somme de ces deux aires vaut 1. Pour calculer d'autres types de probabilité, on pourra utiliser la formule (3.18) qui s'écrit aussi

$$P(a \leq X \leq b) = \int_a^b p(x)dx = \int_{-\infty}^b p(x)dx - \int_{-\infty}^a p(x)dx. \tag{3.21}$$

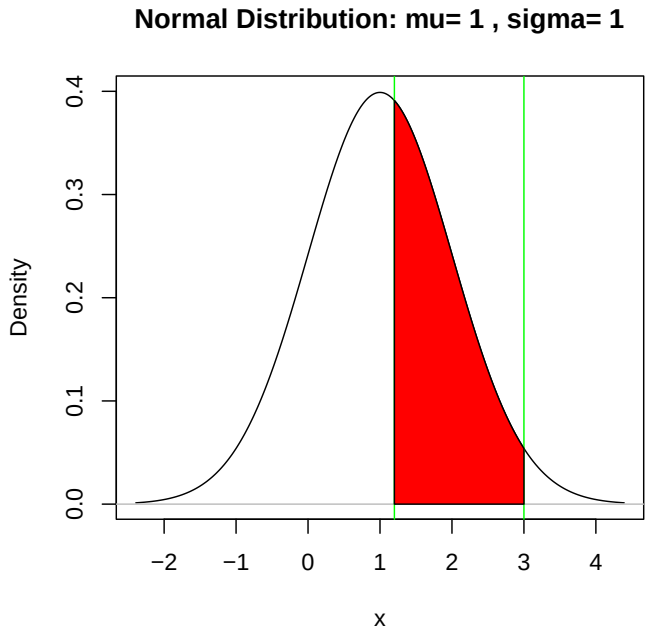


FIGURE 3.10. La loi normale avec $\mu = 1$ et $\sigma = 1$ et l'aire en rouge comprise entre les abscisses 1.2 et 3.

soit encore

$$P(a \leq X \leq b) = P(x \leq b) - P(x \leq a) \tag{3.22}$$

Si on passe par les "aires à droites", on a

$$P(a \leq X \leq b) = P(x \geq a) - P(x \geq b) \tag{3.23}$$

♢ On rappelle aussi que

$$P(X = a) = 0 \tag{3.24}$$

et donc

$$P(X \leq a) = P(X < a). \tag{3.25}$$

MANIPULATION AVEC R 3.40.
Avec R, pour calculer les probabilités normales :

- *Avec Rcmdr :*
  On va dans le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Probabilités normales". Il faut choisir dans la fenêtre de dialogue les valeurs de *mu* et *sigma*. On peut alors calculer l'aire, à gauche ou à droite, sous la courbe de densité pour n'importe quelle valeur (appelée *quantile*, voir remarque 3.38 page 20).
- *Sans Rcmdr :*
  On peut utilise la fonction `pnorm` qui fournit directement dans "Rgui", la fonction de distribution (et donc le calcul des probabilités) pour la loi normale normale : si `x` est un nombre (dit quantile)
  - la commande
    ```
    pnorm(x, mean = mu, sd = sigma)
    ```
    fournit la probabilité $\big(P(X \leq x)$, associée à la lois normales de moyenne `mu` et d'écart-type `sigma`,
  - et la commande

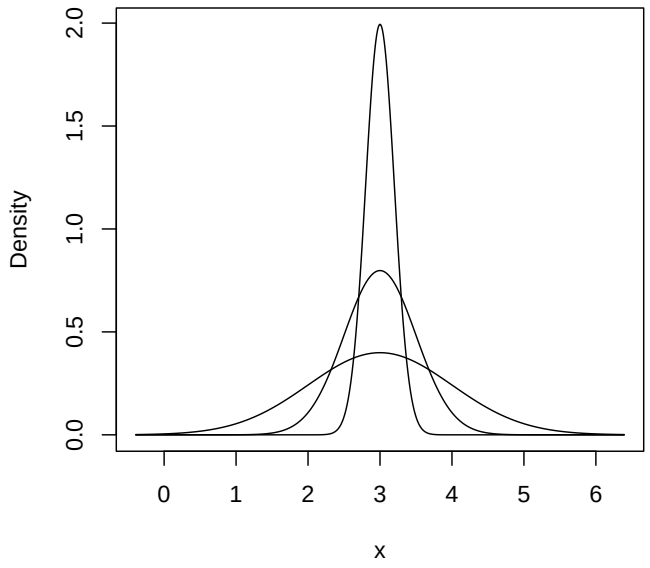


FIGURE 3.11. La loi normale avec $\mu = 3$ et $\sigma \in \{0.2, 0.5, 1\}$.

`pnorm(x, mean = mu, sd = sigma, lower.tail = FALSE)`
fournit la probabilité $\big(P(X \geq x)\big)$, associées à la même loi normale. Dans ce dernier cas, notons que $P(X \geq x) = P(X > x)$!

Plus généralement, si $X = (x_1, \ldots, x_n)$ est un tableau de valeurs (dites quantiles) , alors

- la commande
  `pnorm(X, mean = mu, sd = sigma)`
  fournit le vecteur des probabilités $\big(P(X \leq x_1), \ldots, P(X \leq x_n)\big)$, associées à la lois normales de moyenne `mu` et d'écart-type `sigma`,
- et la commande
  `pnorm(X, mean = mu, sd = sigma, lower.tail = FALSE)`
  fournit le tableau des probabilités $\big(P(X \geq x_1), \ldots, P(X \geq x_n)\big)$, associées à la même loi normale.

EXERCICE 3.41. On étudie la loi normale centrée réduite (c'est-à-dire de moyenne nulle et d'écart-type égal à 1)

(1) Calculer la probabilité correspondant aux intervalles suivants : $P(X \leq -0.5)$, $P(X \leq 4.5)$, $P(X \geq 1.25)$ et $P(X \geq -2)$.
(2) Puis, en procédant en deux temps, calculer $P(1.25 \leq X \leq 1.5)$ et $P(-0.65 \leq X \leq 1.4)$.
(3) En utilisant la remarque 3.35 page 18, faites une figure où apparaissent les probabilités $P(X \leq -0.5)$ et $P(1.25 \leq X \leq 1.5)$.

Voir éléments de correction page 30.

EXERCICE 3.42. Reprendre les questions de l'exercice 3.41 en remplaçant la loi normale centrée réduite par la loi $X \rightsquigarrow \mathcal{N}(\mu = 1.5, \sigma = 2)$.

Voir éléments de correction page 30.

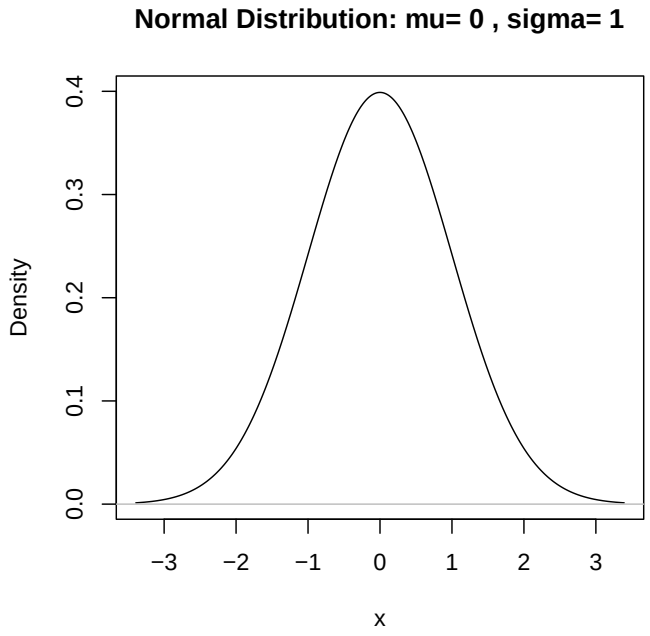


FIGURE 3.12. La loi normale centrée réduite

EXERCICE 3.43. Sundberg étudie la VO_2max d'une population d'enfants (8-15 ans). L'histogramme de ces VO_2max a une allure de courbe en cloche. On considérera donc que la loi normale d'espérance $\mu = 55.4$ et d'écart-type $\sigma = 7.7$ modélise correctement cette mesure.

(1) Tracer la densité de la loi normale associée.

(2) Quelle est la probabilité qu'un enfant de cet âge ait une VO_2max supérieure à 60 ? Supérieure à 70 ? Inférieure à 50 ?

(3) Douze enfants aveugles ont également été évalués. La moyenne observée n'est alors que de 45.3 . Quelle est la probabilité qu'un enfant voyant présente une VO_2max inférieure à cette valeur ?

Voir éléments de correction page 32.

EXERCICE 3.44. Comment établir des limites concernant le taux d'hématocrites (volume total des globules rouges par rapport au sang) ? O'toole *et al.* se sont intéressés à une population de triathlètes et ont établi que les taux observés avant une course chez ces hommes suivent assez fidèlement une loi normale d'espérance $\mu = 43{,}2$ (%) et d'écart-type $\sigma = 2{,}9$.

(1) Quelle est la probabilité d'observer dans ce contexte un triathlète dont le taux dépasse 45 % ? Dont le taux dépasse 50 % ?

(2) Les auteurs suggèrent d'employer une valeur de détection de 52 % car elle est située à trois écarts-types au dessus de la moyenne. Qu'en pensez-vous ?

3.5. La distribution de Student

Nous introduisons, très sommairement cette deuxième loi continue, très utile !

Nous dirons que la variable aléatoire X suit la loi de Student à n (entier non nul) degrés de liberté et on notera

$$X \rightsquigarrow t_n. \tag{3.26}$$

L'entier n sera appelé "degré de liberté" et noté *ddl* ("degree of freedom" ou *df* en anglais).

Retenons que quand n "grand" (supérieur à 30 ou 60, selon les cas !), cette loi se rapproche de la loi normale.

Comme précédemment, on pourra faire des graphiques et calculer des probabilité :

- *Avec Rcmdr :*
  En utilisant le menu déroulant "Distributions" , l'option "Distributions continue" puis "Distribution t".
- *Sans Rcmdr :*
  En utilisant les fonction `dt`, `pt`, `qt` et `rt`.

3.6. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.2

La fonction `sample` effectue un tirage aléatoire dans l'échantillon {1,2,3,4,5,6}. Au vu de la définition 3.4 page 6, on observe bien des proportions qui se rapproche de la probabilité d'apparition de chacune des six valeurs, soit 1/6.

Ces instructions permettent donc de simuler numériquement des tirages de dés à six faces.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.3

(1) Tapez dans "Rgui"

```
sample(1:50, size = 6, replace = F)
```

ou mieux

```
sort(sample(1:50, size = 6, replace = F))
```

(2) Faire plusieurs fois la commande précédente ou alors, plus subtilement,

```
n <- 20
dudu <- matrix(nrow = n, ncol = 6)
for (i in 1:n) {
    dudu[i, ] <- sort(sample(1:50, size = 6, replace = F))
}
dudu
```

Ceux qui souhaitent en savoir plus sur le fonctionnement des commandes `for` et `matrix` pourront taper dans R, les commandes `help("for")` et `help("matrix")` ou `?matrix` ♢

Cela donnerait par exemple

```
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]    8   10   17   18   20   38
 [2,]    3    4   10   39   44   49
 [3,]    1   15   21   28   36   44
 [4,]    4    9   30   45   47   48
 [5,]    1   13   18   20   30   31
 [6,]    8   11   16   20   28   43
 [7,]    2    8   13   25   28   43
 [8,]   17   21   37   43   44   47
 [9,]   17   18   24   43   45   50
[10,]    5   16   22   24   44   50
[11,]    4   11   18   19   23   44
```

```
[12,]    1    9   13   25   27   36
[13,]   11   17   21   27   37   45
[14,]    9   10   20   44   46   50
[15,]    5   19   25   28   44   46
[16,]   16   19   27   30   35   37
[17,]    3   15   29   33   39   42
[18,]    2   15   22   29   42   46
[19,]    2    3    8   19   45   48
[20,]    6    7   13   15   22   33
```

(3) Non, bien sûr ....

Éléments de correction de l'exercice 3.16
On a

$$p = 0.5, \quad n = 5,$$
$$p = 5/6, \quad n = 1.$$

Attention, dans le cas de la roulette russe, cette expérience ne peut être répetée plusieurs fois, car une issue malheureuse ne permet pas de continuer à jouer !

Éléments de correction de l'exercice 3.13 Pour retrouver (3.6a), on peut écrire

$$\mathbb{E}(X) = \frac{1}{6}(1+2+3+4+5+6)$$

On peut calculer cela en se rappelant que

$$1+2+3+4+5+6 = \frac{6\times 7}{2} = 21$$

et donc

$$\mathbb{E}(X) = 3.5$$

On peut aussi le calculer avec Ren tapant

```
sum(1:6)/6
```

Pour retrouver (3.6b), on peut écrire

$$\sigma = \sqrt{\sum_{i=1}^{6} \frac{1}{6}(i-3.5)^2}$$

et grâce à Ren tapant

```
sqrt((sum((1:6 - 3.5)^2))/6)
```

on obtient

$$\sigma \approx 1.70783 \tag{3.27}$$

Éléments de correction de l'exercice 3.21

Voir en figure 3.13 page suivante les cinq courbes obtenues. On constate que quand le paramètre p augmente et se rapproche de 1, les courbes de décalent vers la droite : cela signifie que quand la probabilité de succès se rapproche de 1, on a plus de chance d'obtenir un grand nombre de succès ! Autrement dit, les résultats les plus probables sont à droite quand p se rapproche de 1, à gauche quand p proche de 0, et proche du milieu quand p proche de 0.5.

Pour les cas extrèmes $p = 0$ et $p = 1$, les graphes de probabilités sont très simples !

Éléments de correction de l'exercice 3.22

Voir en figure 3.14 les quatres courbes obtenues.

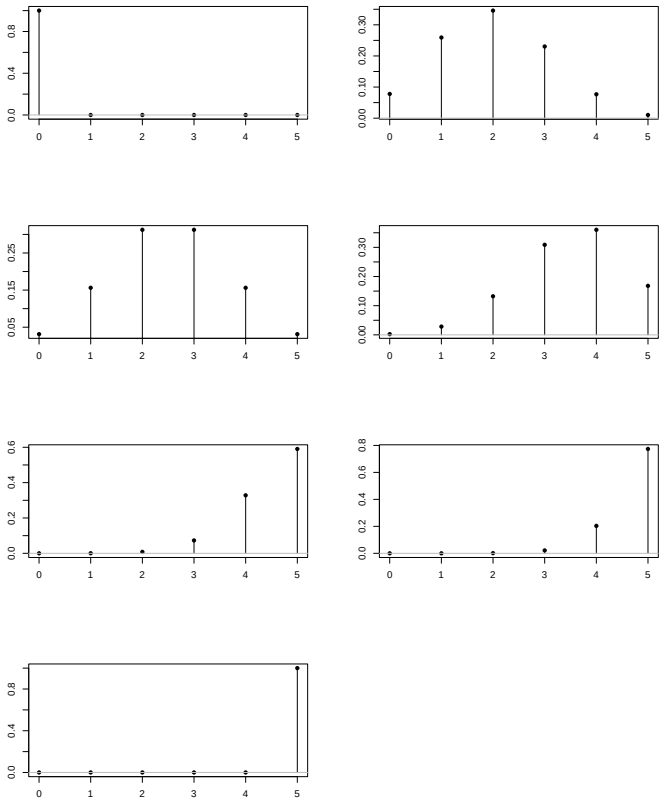


FIGURE 3.13. Les cinq graphes correspondant aux graphes de la loi binomiale avec la probabilité de succès en $p = 0$, $p = 0.4$, $p = 0.5$, $p = 0.7$, $p = 0.9$, $p = 0.95$, $p = 1$ et $n = 5$.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.24
On obtient successivement en écrivant par exemple

```
20*0.5
20*0.5*(1-0.5)
sqrt(20*0.5*(1-0.5))
```

$$\mathbb{E}(X) = 10, \quad \sigma = 2.236068,$$
$$\mathbb{E}(X) = 8, \quad \sigma = 2.529822,$$
$$\mathbb{E}(X) = 160, \quad \sigma = 5.656854.$$

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.32

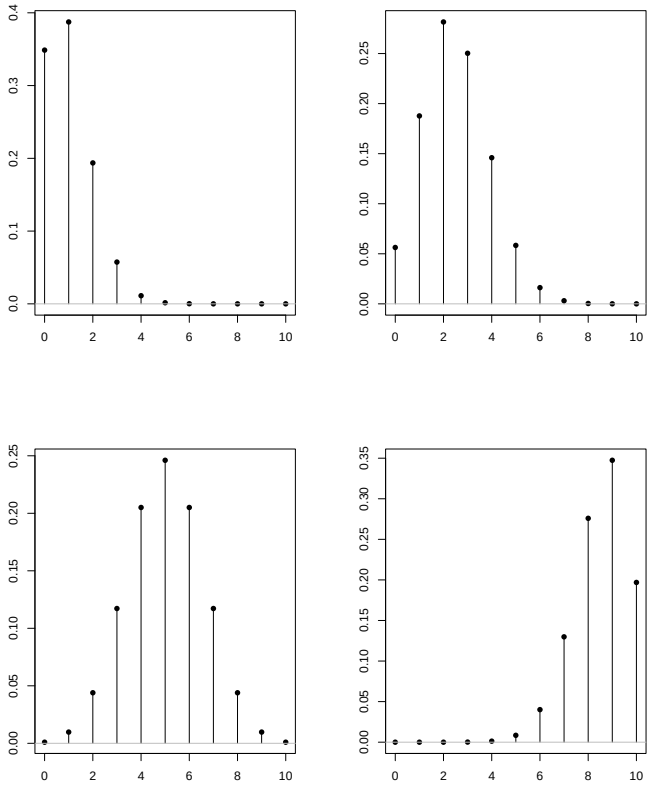


FIGURE 3.14. Les quatre graphes correspondant aux graphes de la loi binomiale avec la probabilité de succès en $p = 0.1$, $p = 0.25$, $p = 0.5$, $p = 0.85$ et $n = 10$.

(1) On obient successivement pour une variable aléatoire binomiale X de paramètres $n = 7$ et $p = 0.2$:

$$P(X = 2) = 0.275251,$$
$$P(X = 0) = 0.209715,$$
$$P(X = 9) = 0,$$
$$P(X \leq 5) = 0.999629,$$
$$P(X \geq 5) = 0.004672,$$
$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.999629 = 0.000371$$

pour cette dernière, on peut aussi passer par l'aire à droite :

$$P(X > 5) = 0.000371,$$
$$P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 1) = 0.999629 - 0.576717 = 0.422912.$$

(2) On a

$$\begin{aligned} P(X = 0) + \ldots + P(X = 5) &= 0.209715 + 0.367002 + 0.275251 + 0.114688 + 0.028672 + 0.004301 \\ &= 0.999629 \end{aligned}$$

et

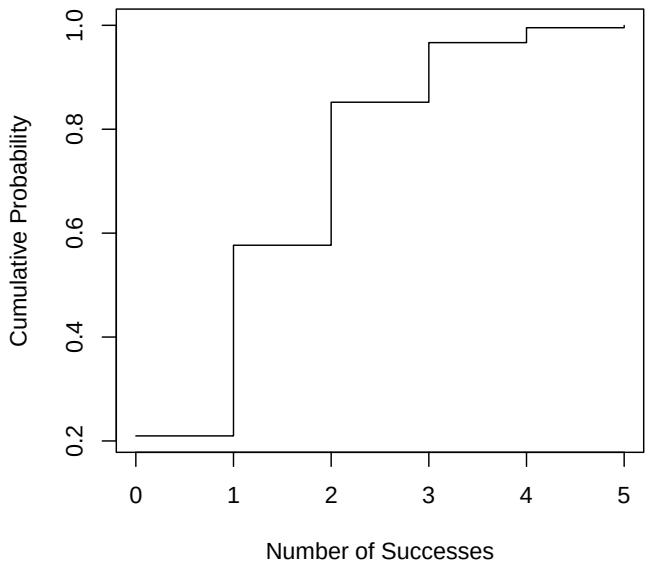
$$P(X \leq 5) = 0.999629$$

(3) On a

$$\begin{aligned} P(X = 6) + P(X = 7) &= 0.000358 + 1.3e - 05 \\ &= 0.000371 \end{aligned}$$

et

$$P(X > 5) = 0.000371$$



(4)

FIGURE 3.15. Le graphes des probabilités cumulées pour $n = 7$ et $p = 0.2$.

Voir le graphique 3.15.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 3.33

(1) Le nombre moyen de personnes que l'on peut espérer toucher dans cette première vague d'appels est par définition la moyenne des valeurs de succès, soit encore l'espérance de la loi binomiale de paramètres n et p, c'est-à-dire $\mathbb{E}(X) = np$, soit 120 personnes.

(2) La probabilité de contacter au moins $n_3 = 120$ personnes est égale à $P(X \geq 120)$. On peut la calculer de deux façon :
   - On écrit et on calcule avec les probabilités cumulées
   $$P(X \geq 120) = 1 - P(X < 120) = 1 - P(X \leq 119) = 0.5306621577.$$
   - On peut aussi passer par "l'aire à droite" :
   $$P(X \geq 120) = P(X > 119) = 0.5306621577.$$

(3) La probabilité de contacter au moins $n_1 = 150$ personnes est égale à $P(X \geq 150)$. On peut la calculer de même de deux façon :

– On écrit et on calcule avec les probabilités cumulées

$$P(X \geq 150) = 1 - P(X < 150) = 1 - P(X \leq 149) = 5.8989e - 06.$$

– On peut aussi passer par "l'aire à droite" :

$$P(X \geq 150) = P(X > 149) = 5.8989e - 06.$$

(4) Si on veut appeler n_2 personnes pour espérer, en moyenne, obtenir 150 succès, il faut utiliser l'argument de la question 1 à "l'envers" :

$$\mathbb{E}(X(n_2, 0.6)) = n_2 p = 150$$

On a donc

$$n_2 = \frac{150}{p} = \frac{150}{0.6} = 250.$$

Il faut donc contacter $n_2 = 250$ personnes.

Éléments de correction de l'exercice 3.41

(1) On obtient

$$P(X \leq -0.5) = 0.308538,$$
$$P(X \leq 4.5) = 0.999997.$$

En passant par "l'aire à droite", on obtient

$$P(X \geq 1.25) = 0.10565,$$
$$P(X \geq -2) = 0.97725.$$

Ces dernières peuvent aussi être obtenue par "l'aire à gauche" :

$$P(X \geq 1.25) = 1 - 0.89435 = 0.10565,$$
$$P(X \geq -2) = 1 - 0.02275 = 0.97725.$$

(2) De même, on obtient en utilisant (3.21) (en passant par "l'aire à gauche") :

$$P(1.25 \leq X \leq 1.5) = P(X \leq 1.5) - P(X \leq 1.25) = 0.933193 - 0.89435 = 0.038843,$$
$$P(-0.65 \leq X \leq 1.4) = P(X \leq 1.4) - P(X \leq -0.65) = 0.919243 - 0.257846 = 0.661397,$$

(3)

En utilisant la remarque 3.35 page 18, on peut tracer les aires correspondant aux probabilités $P(X \leq -0.5)$ et $P(1.25 \leq X \leq 1.5)$: voir figure 3.16 page ci-contre.

Éléments de correction de l'exercice 3.42

(1) On obtient en utilisant "l'aire à gauche" :

$$P(X \leq -0.5) = 0.158655254,$$
$$P(X \leq 4.5) = 0.933192799.$$

En passant par "l'aire à droite", on obtient

$$P(X \geq 1.25) = 0.549738225,$$
$$P(X \geq -2) = 0.959940843.$$

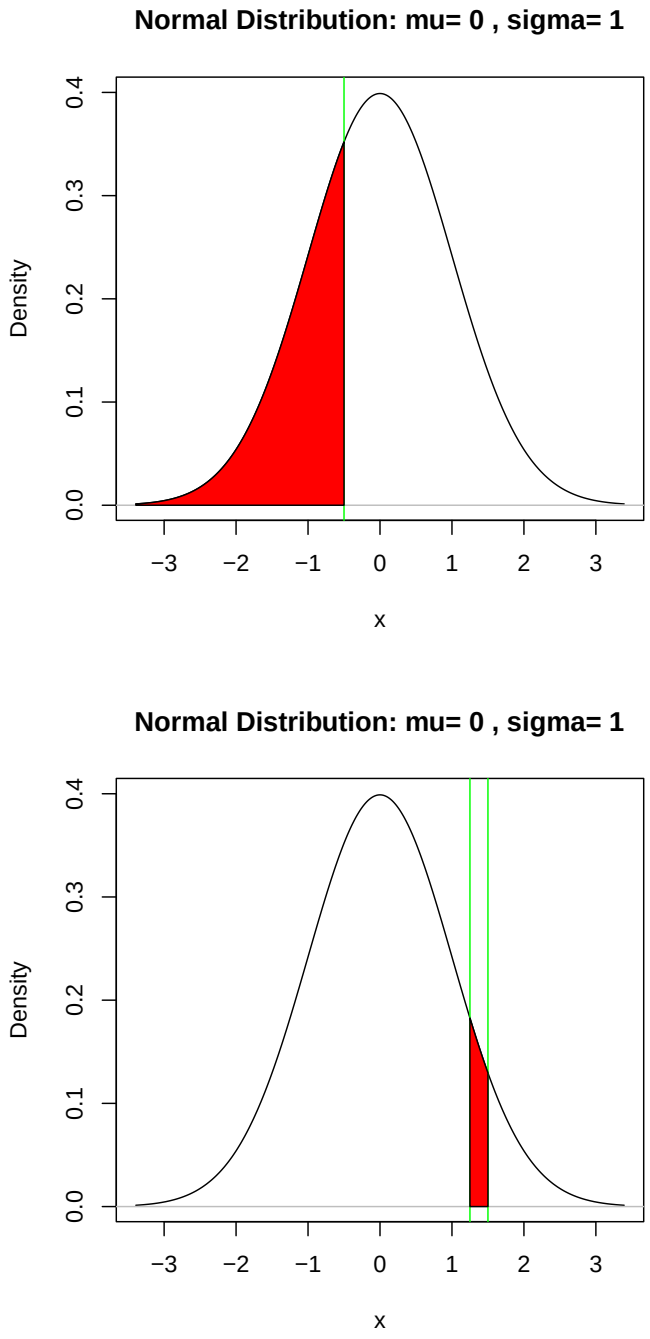


FIGURE 3.16. Les deux aires correspondant aux probabilités $P(X \leq -0.5)$ et $P(1.25 \leq X \leq 1.5)$.

Ces dernières peuvent aussi être obtenue par "l'aire à gauche" :

$$P(X \geq 1.25) = 1 - 0.450261775 = 0.549738225,$$
$$P(X \geq -2) = 1 - 0.040059157 = 0.959940843.$$

(2) De même, on obtient en utilisant (3.21) (en passant par "l'aire à gauche") :

$P(1.25 \leq X \leq 1.5) = P(X \leq 1.5) - P(X \leq 1.25) = 0.5 - 0.450261775 = 0.049738225,$

$P(-0.65 \leq X \leq 1.4) = P(X \leq 1.4) - P(X \leq -0.65) = 0.480061194 - 0.141187364 = 0.33887383.$

Éléments de correction de l'exercice 3.43

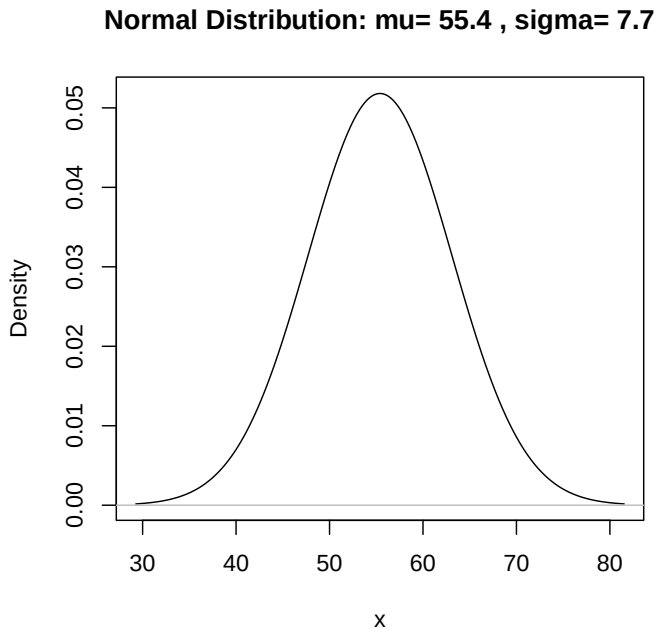


Figure 3.17. La densité de la loi normale de moyenne $\mu = 55.4$ et d'écart-type $\sigma = 1$.

(1) On obtient le graphe de la figure 3.17.

(2) On suppose que $\mathrm{VO_2max} \rightsquigarrow \mathcal{N}(55.4, 7.7)$. Comme dans l'exercice 3.42 page 23, on obtient

$$P(\mathrm{VO_2max} \geq 60) = 0.275119,$$
$$P(\mathrm{VO_2max} \geq 70) = 0.028973,$$
$$P(\mathrm{VO_2max} \leq 50) = 0.241558.$$

(3) De même, on obtient

$$P(\mathrm{VO_2max} \leq 45.3) = 0.094813.$$

CHAPITRE 4

Intervalles de confiance

On pourra consulter [**6**] (dont s'inspire partiellement ce chapitre), [**4**] ou [**8**].

4.1. Rappels sur la densité de probabilité

4.1.1. Rappels théoriques

On rappelle (voir la figure 3.5 page 11 et la remarque 3.38 page 20) que si X est une variable aléatoire continue de densité de probabilité f, alors la probabilité que X appartiennent à l'intervalle $[a, b]$ est l'aire sous la courbe de la fonction f comprise entre les abscisses a et b, ce que l'on écrit avec les intégrales :

$$P(a \leq X \leq b) = \int_a^b f(x)dx. \tag{4.1}$$

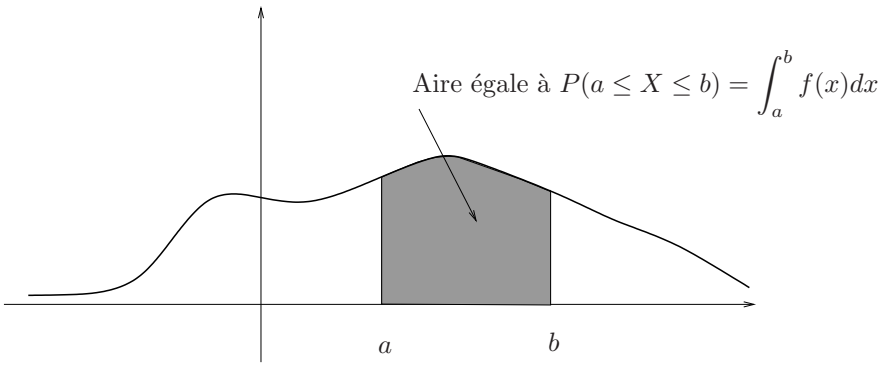


FIGURE 4.1. Une densité de probabilité quelconque avec, en gris, l'aire égale à $P(a \leq X \leq b)$.

Voir figure 4.1.

On a aussi

$$p = \int_{-\infty}^{q} f(x)dx \tag{4.2}$$

où

$$p = P(X \leq q). \tag{4.3}$$

Dans cette égalité, le nombre p est appelée probabilité et le nombre q quantile.

On a déjà vu, dans le cas où la loi de probabilité est normale, comment passer de q à p. On renvoie à la section 3.4.5 page 21. Refaire en particulier les exercices 3.41 page 23 et 3.42, s'ils n'ont pas été compris.

Réciproquement, on peut aussi, avec ®, passer de la probabilité p au quantile q :

MANIPULATION AVEC R 4.1. • *Avec Rcmdr :*

On va dans le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Quantiles normaux". Il faut choisir dans la fenêtre de dialogue les valeurs de *mu* et *sigma*. On peut alors calculer le quantile (en laissant l'aire à gauche).

- *Sans Rcmdr :*

  On peut utiliser la fonction `qnorm` qui fournit directement dans "Rgui", le quantile en fonction de la probabilité pour la loi normale normale ; plus précisement, si $P = (p_1, \ldots, p_n)$ est un vecteur de valeurs (dites probabilités) , alors la commande

  ```
  qnorm(X, mean = mu, sd = sigma)
  ```

  fournit le vecteur des quantiles $Q = (q_1, \ldots, q_n)$ tels que $P(X \leq q_i) = p_i$ associées à la lois normales de moyenne `mu` et d'écart-type `sigma`.

EXERCICE 4.2. On étudie la loi normale centrée réduite (c'est-à-dire de moyenne nulle et d'écart-type égal à 1) calculer les quantiles correspondant aux probabilité suivantes : $p_1 = 0.2$, $p_2 = 0.7$, $p_3 = 0$, $p_4 = 1$, $p_5 = 1.2$.
Commentez !
Voir éléments de correction page 54.

EXERCICE 4.3. Reprendre les questions de l'exercice 4.2 en remplaçant la loi normale centrée réduite par la loi $X \rightsquigarrow \mathcal{N}(\mu = 1.5, \sigma = 2)$.
Voir éléments de correction page 54.

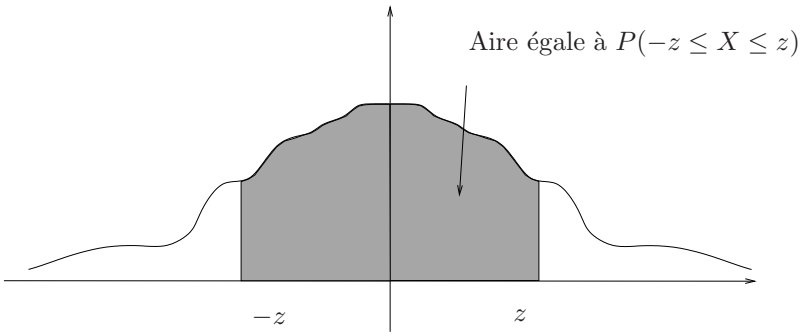


FIGURE 4.2. Une densité de probabilité symétrique

Supposons maintenant que la densité de probabilité soit paire (symétrique), c'est-à-dire : $f(-x) = f(x)$ Voir figure 4.2. On se donne une probabilité $p \in [0, 1]$ et un nombre positif (le quantile) z. On cherche "à résoudre" (dans le sens $p \to z$ ou le sens $z \to p$) :

$$P(-z \leq X \leq z) = p. \tag{4.4}$$

On a la propriété suivante ;

$$P(-z \leq X \leq z) = 2P(X \leq z) - 1 \tag{4.5}$$

DÉMONSTRATION. Par symétrie, on a

$$P(-z \leq X \leq 0) = P(0 \leq X \leq z) \tag{4.6}$$

On a aussi

$$1 = P(-\infty \leq X \leq 0) + P(0 \leq X \leq \infty)$$

Donc, (4.6) fournit

$$P(-\infty \leq X \leq 0) = \frac{1}{2}. \tag{4.7}$$

On déduit donc de (4.6) et (4.7)

$$\begin{aligned} P(-z \leq X \leq z) &= P(-z \leq X \leq 0) + P(0 \leq X \leq z) \\ &= 2P(0 \leq X \leq z), \\ &= 2\left(P(-\infty \leq X \leq z) - P(-\infty \leq X \leq 0)\right), \\ &= 2\left(P(X \leq z) - \frac{1}{2}\right), \\ &= 2P(X \leq z) - 1. \end{aligned}$$

□

♢

Ainsi, l'égalité (4.4) est équivalente à

$$2P(X \leq z) - 1 = p. \tag{4.8}$$

Souvent, on appelle la probabilité p, "niveau de confiance" et on la note NC. Il nous faut donc résoudre

$$2P(X \leq z) - 1 = NC. \tag{4.9}$$

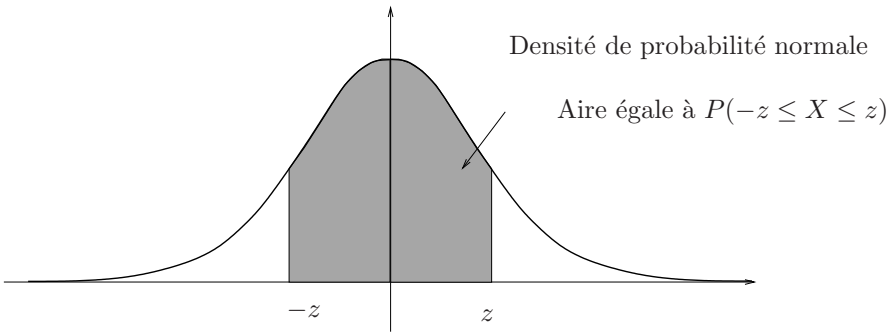


FIGURE 4.3. Une densité de probabilité normale

On suppose maintenant que la densité de probabilité est normale (voir figure 4.3). Sur cette figure, l'aire en gris est donc égale au nombre NC.

(1) Si on connaît z, on calcule la probabilité $P(X \leq z)$, en utilisant ce qu'on a fait en section 3.4.5 page 21 et on déduit donc la valeur de NC grâce à 4.9.

(2) Si on connaît NC, de (4.9), on déduit

$$P(X \leq z) = \frac{NC + 1}{2}. \tag{4.10}$$

On en déduit alors le quantile z en utilisant ce qu'on a fait page 33.

EXEMPLE 4.4. On suppose que la loi normale est centrée réduite. On se donne

$$z = 2. \tag{4.11}$$

On en déduit $P(X \leq z) = 0.97725.$ et donc, d'après (4.9),

$$NC = 2P(X \leq z) - 1 = 2 \times 0.97725 - 1 = 0.9545.$$

soit

$$NC = 0.9545 \tag{4.12}$$

souvent arrondi en

$$NC = 0.95 \tag{4.13}$$

EXEMPLE 4.5. On suppose que la loi normale est centrée réduite. Réciproquement, on se donne

$$NC = 0.95. \tag{4.14}$$

On en déduit d'après (4.10)

$$P(X \leq z) = \frac{NC + 1}{2} = 0.975.$$

On en déduit

$$z = 1.959964 \tag{4.15}$$

souvent arrondi en

$$z = 2. \tag{4.16}$$

4.1.2. Application à la loi normale

Si on reprend les exemples 4.4 et 4.5, on vient donc de montrer, selon (4.4), pour la loi normale centrée réduite, dans $100 * NC = 95\%$ des cas, la variable aléatoire X est comprise entre -2 et 2. Autrement dit, pour une loi normale centrée réduite il y a 95 chances sur 100 d'observer (sur la valeur tirée aléatoirement) un écart à zéro inférieur à 2.

NC	z	z approché
0	0	0
0.5	0.67449	0.7
0.6	0.841621	0.8
0.7	1.036433	1
0.8	1.281552	1.3
0.9	1.644854	1.6
0.95	1.959964	2
0.99	2.575829	2.6
0.999	3.290527	3.3
1	Inf	Inf

TABLE 4.1. Quelques valeurs de z en fonction du niveau de confiance NC.

Les différentes valeurs de z en fonction du niveau de confiance NC sont données dans le tableau 4.1.

EXERCICE 4.6. Retrouver ces valeurs de z. Voir éléments de correction page 54

En reprenant ce qu'on a dit plus haut, on peut affirmer successivement que pour une loi normale centrée réduite il y a

- 50 chances sur 100 d'observer un écart à zéro inférieur à 0.7.
- 60 chances sur 100 d'observer un écart à zéro inférieur à 0.8.
- 70 chances sur 100 d'observer un écart à zéro inférieur à 1.
- 80 chances sur 100 d'observer un écart à zéro inférieur à 1.3.
- 90 chances sur 100 d'observer un écart à zéro inférieur à 1.6.
- 95 chances sur 100 d'observer un écart à zéro inférieur à 2.
- 99 chances sur 100 d'observer un écart à zéro inférieur à 2.6.
- 99.9 chances sur 100 d'observer un écart à zéro inférieur à 3.3.

On peut aussi remplir ce tableau à "l'envers" : voir le tableau 4.2 page suivante.

Comme précédemmennt, on peut donc affirmer successivement que pour une loi normale centrée réduite il y a

- 68.3 chances sur 100 d'observer un écart à zéro inférieur à 1.
- 95.4 chances sur 100 d'observer un écart à zéro inférieur à 2.
- 99.7 chances sur 100 d'observer un écart à zéro inférieur à 3.
- 99.994 chances sur 100 d'observer un écart à zéro inférieur à 4.
- 99.99994 chances sur 100 d'observer un écart à zéro inférieur à 5.

Voir la figure 4(a) page 38 qui résume tout cela.

z	$100 * NC$	$100 * NC$ approché
0	0	0
1	68.268949	68.3
2	95.449974	95.4
3	99.73002	99.7
4	99.993666	99.994
5	99.999943	99.99994
Inf	100	100

TABLE 4.2. Quelques valeurs du niveau de confiance NC en fonction de z.

REMARQUE 4.7. Voir de nouveau la remarque C.16 page 75 et la justifier !

Si nous transformons cette distribution centrée réduite en une distribution quelconque par translation-dilatation (voir remarque 3.37 page 20), la propriété 4.4, se traduit pour une loi normale de moyenne μ et d'écart-type σ par le fait suivant : il y a $100 * NC$ chances sur 100 d'observer un écart à la moyenne inférieur à z écarts-type.

Soit encore, en utilisant le tableau 4.2 : il y a

- 68.3 chances sur 100 d'observer un écart à la moyenne inférieur à 1 écarts-type.
- 95.4 chances sur 100 d'observer un écart à la moyenne inférieur à 2 écarts-type.
- 99.7 chances sur 100 d'observer un écart à la moyenne inférieur à 3 écarts-type.
- 99.994 chances sur 100 d'observer un écart à la moyenne inférieur à 4 écarts-type.
- 99.99994 chances sur 100 d'observer un écart à la moyenne inférieur à 5 écarts-type.

Voir la figure 4(b) page suivante qui résume tout cela.

4.2. Principe théorique de l'intervalle de confiance

On se donne un niveau de confiance NC, nombre compris entre 0 et 1.

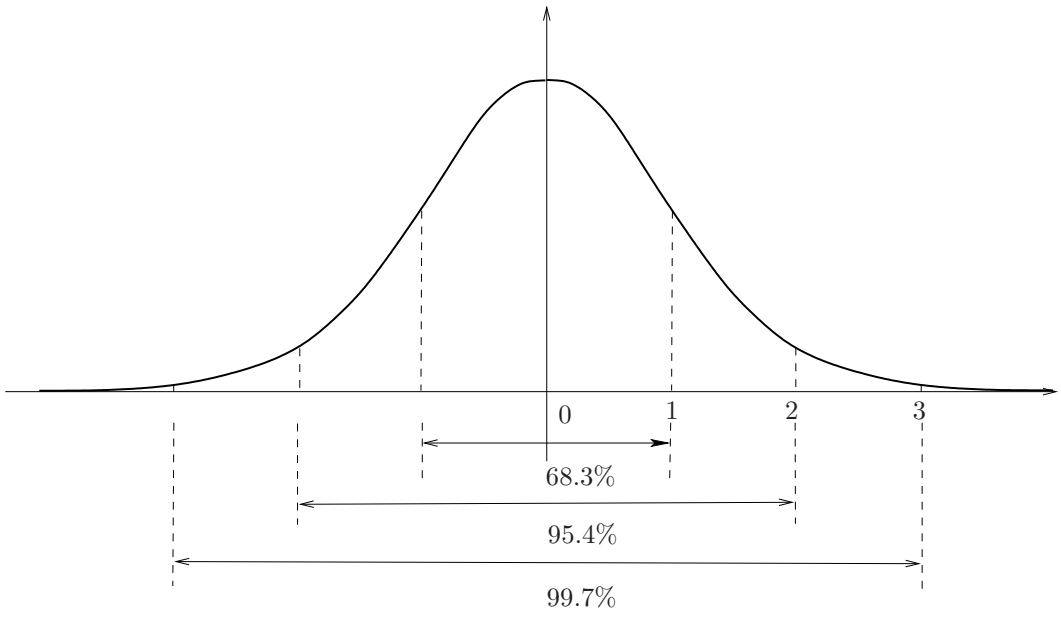
On se réfère à la figure N.5 page 194.

On se donne une variable aléatoire, dont la loi est défini par (au moins) un paramètre Θ. On suppose que l'on connaît un échantillon issu de cette loi, c'est-à-dire, une valeur x prise par X.

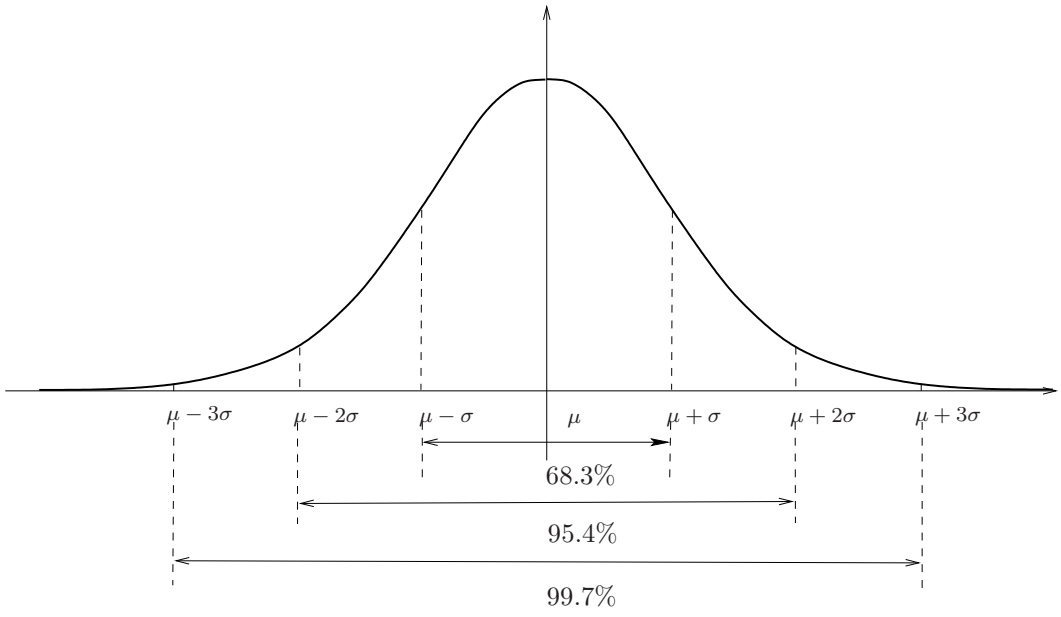
Pour estimer le paramètre inconnu Θ, nous allons construire un *intervalle de confiance*, à partir de θ, la valeur de la statistique définie à partir de x et qui contiendra la valeur inconnue Θ dans $100NC\%$ des cas, au sens suivant : si l'on répétait un grand nombre de constructions identiques (à partir d'autres valeurs d'autres échantillons), la proportion des intervalles de confiance contenant réellement Θ serait de NC. Cela signifie aussi que dans $100(1 - NC)\%$ des cas, l'intervalle ne contiendra pas le paramètre inféré Θ !

Il faut comprendre aussi que si NC augmente et se rapproche de 1, la largeur de l'intervalle de confiance grandira. Pour le cas limite correspondant à $NC = 1$, l'intervalle de confiance sera $\mathbb{R}$, l'ensemble des réel tout entier ! Pour le cas limite correspondant à $NC = 0$, l'intervalle de confiance sera réduit à θ, la valeur expérimentale. Entre ces deux cas limite, plus on voudra "couvrir large", plus on aura de valeur possible, ce qui est légitime. $\diamondsuit$

En général, la valeur de NC est de 0.95 (valeur proche de 1). Dans des domaines sensibles, elle sera beaucoup plus proche de 1 encore.



(a) moyenne $\mu = 0$ et écart-type $\sigma = 1$



(b) moyenne μ et d'écart-type σ

FIGURE 4.4. Deux densités de probabilité normales.

4.3. Intervalle de confiance d'une proportion

4.3.1. Construction à partir d'un exemple

On rappelle qu'en section N.2.2, on a vu que si l'on considérait un grand nombre de tirages d'une variable aléatoire X issue d'une loi binomiale de paramètre n et p, les fréquences d'apparition de chacune des valeurs possibles de succès se rapprochaient de la loi de probabilité.

Plus précisément, on peut aussi générer un certain nombre d'échantillons binomiaux, c'est-à-dire issus loi binomiale de paramètre n et p.

Désormais, on s'intéresse non plus à la variable aléatoire X égale au nombre de succès, mais à la variable aléatoire p_r égale à la proportion de succès X/n.

Nous allons donc créér 10000 échantillons binomiaux correspondant à $n = 1000$ et $p = 0.3$ et tracer ensuite un histogramme en densité avec 25 classes. en procédant comme il suit :

- *Avec Rcmdr :*

  On utilise le menu déroulant "Distributions", l'option "Distributions discrètes" puis "Distribution binomiale", puis "Echantillon d'une distribution binomiale". Dans la fenêtre dialogue, il faut indiquer pour
  - "Nombre d'essais" (valeur de n) : 1000,
  - "Probabilité de succès" (valeur de p) : 0.3,
  - "Nombre d'échantillons" : 10000,
  - "Nombre d'observations" : 1

  Laisser les autres champs tels quels. Un jeu de données est alors créé qui s'appelle par défaut "`EchantillonsBinomiaux`". On introduit alors une nouvelle variable obtenue en divisant `EchantillonsBinomiaux` par 1000. On trace ensuite un histogramme en densité avec 25 classes.
- *Sans Rcmdr :*

  Il faut taper les commandes suivantes :

```
dede<-rbinom(n=10000, size=1000, prob=0.3)/1000
hist(dede,freq=T,breaks=25)
```

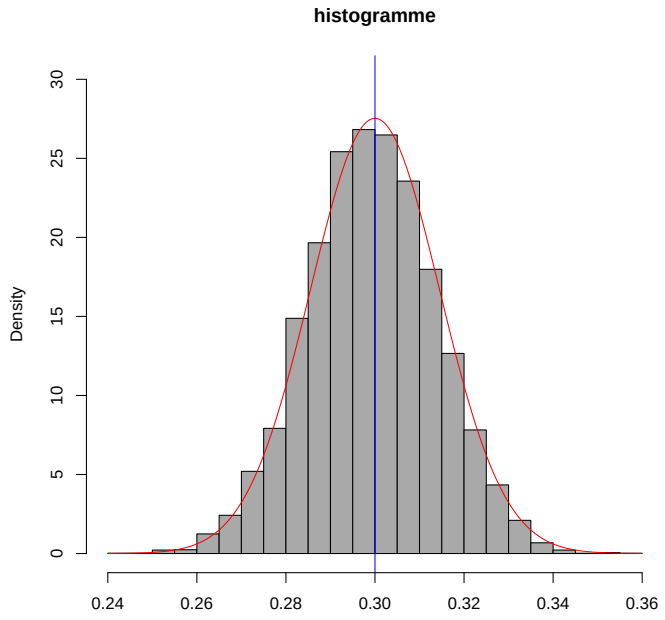


FIGURE 4.5. Histogramme de la distribution d'échantillonnage (10000 tirages) d'une proportion observée sur une loi binomiale de paramètre $n = 1000$ et $p = 0.3$ en densité avec 25 classes

On observe alors un histogramme similaire à celui de la figure 4.5.

On admet que, que par division par n, la proposition 3.23 page 14 donne le résultat suivant :

PROPOSITION 4.8. *L'espérance et l'écart-type de la variable aléatoire p_r correspondant à la proportion de succès d'une variable aléatoire binomiale de paramètres n et p sont respectivement égaux à p et $\sqrt{\dfrac{p(1-p)}{n}}$.*

D'autre part, on constate sur la figure 4.5 page précédente que
- l'histogramme a l'allure d'une densité normale (c'est-à-dire, "en cloche"). Voir la courbe rouge associée à la loi normale de moyenne $\mu = p = 0.3$ et d'écart-type $\sigma = \sqrt{p(1-p)/n} = 0.01449138$, proche de l'histogramme ;
- la distribution de proportions observées est centrée autour de la véritable valeur du paramètre $p = 0.3$. En moyenne la proportion d'un échantillon redonne la probabilité de succès. Voir la ligne d'abscisse $p = 0.3$ en bleu sur la figure.

Rappelons aussi le résultat de la section N.2.3 page 192. Ainsi, on a constaté expérimentalement que

$$p_r \rightsquigarrow \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \tag{4.17}$$

Ce résultat admis est valable pour peu que

$$n \text{ soit "assez grand".} \tag{4.18}$$

On déduit donc (grâce à la remarque 3.37 page 20) que :

$$\frac{p_r - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1) \tag{4.19}$$

On se donne maintenant un niveau de confiance NC dans $[0,1]$. On utilisant le cas 2 page 35, on cherche un nombre z vérifiant (4.10). Ainsi, (4.4) et (4.9) est vraies :

$$P\left(-z \leq \frac{p_r - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z\right) = NC. \tag{4.20}$$

Autrement dit (voir figure 4.3 page 35), dans $100 \times NC\%$ des cas, la variable aléatoire $\frac{p_r - p}{\sqrt{\frac{p(1-p)}{n}}}$ appartient à l'intervalle $[-z, z]$, ce qui se traduit par

$$-z \leq \frac{p_r - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z.$$

C'est donc équivalent (en en prenant l'opposé et en retournant l'inégalité) à

$$-z \leq \frac{p - p_r}{\sqrt{\frac{p(1-p)}{n}}} \leq z,$$

soit encore

$$p_r - z\sqrt{\frac{p(1-p)}{n}} \leq p \leq p_r + z\sqrt{\frac{p(1-p)}{n}}. \tag{4.21}$$

On a donc montré que dans $100 \times NC\%$ des cas, la valeur "théorique" p de la proportion appartient à l'intervalle défini par (4.21), défini à partir de la valeur "expérimentale" p_r. On a donc répondu à l'objectif annoncé en section 4.2 (où ici $\Theta = p$ et $\theta = p_r$). L'inconvient ici est que l'intervalle de confiance contient la valeur p inconnue ! On remplace donc dans cette formule p par p_r, ce qui sera valide si

$$\text{la proportion } p \text{ "n'est pas trop proche" de 0 ou de 1.} \tag{4.22}$$

De façon plus rigoureuse, les hypothèses (4.18) et (4.22) sont équivalente à

$$np \geq 10 \text{ et } n(1-p) \geq 10. \tag{4.23}$$

♢

4.3.2. Théorie

Bref, on a montré que

PROPOSITION 4.9 (Intervalle de confiance "d'une proportion"). *On suppose que l'on réalise un tirage aléatoire d'une loi binomiale de paramètres n et p et que l'on note p_r la proportion de succès observée. L'intervalle de confiance au niveau de confiance NC de la proportion p, sous les hypothèse* (4.18) *et* (4.22) *(ou* (4.23)*), est donné par*

$$\left[p_r - z\sqrt{\frac{p_r(1-p_r)}{n}}, p_r + z\sqrt{\frac{p_r(1-p_r)}{n}}\right] \tag{4.24}$$

Le coefficient multiplicateur z est obtenu sous R de la façon suivante :

- Avec Rcmdr :

  Aller dans le menu "distribution", puis "distribution continue", puis "distribution normale", puis "quantiles normaux" et taper dans le champ "probabilités", le nombre `(1+NC)/2`*, où* `NC` *est le niveau de confiance (laisser les autres champs tels quels).*
- Sans Rcmdr :

  Grâce à la ligne de commande

```
qnorm((1 + NC)/2)
```

Une autre façon de procédérer pour éviter cette approximation est de montrer que, pour tout x dans $[0,1]$, la quantité $x(1-x)$ est majorée par $1/4$. On remplace donc l'intervalle de confiance défini par (4.24) par un intervalle plus large, mais plus "sûr" :

$$\left[p_r - z\frac{1}{2\sqrt{n}}, p_r + z\frac{1}{2\sqrt{n}}\right] \tag{4.25}$$

♢

4.3.3. Avec R

Pour évaluer cet intervalle de confiance dans R, plusieurs solutions :

(1) Une fois le nombre z défini à partir de la proposition 4.9, on peut taper dans "Rgui"

```
pr - z * sqrt(pr * (1 - pr)/n)
```

puis

```
pr + z * sqrt(pr * (1 - pr)/n)
```

Mieux, on pourra taper directement

```
pr + z * c(-1, 1) * sqrt(pr * (1 - pr)/n)
```

(2) Comme dans l'annexe M page 185, vous pouvez récupérer et sourcer la fonction `int.conf.prop.R` qui donne l'intervalle de confiance en fonction
   - de `p` la proportion mesurée,
   - `n` la taille de l'échantillon
   - `NC` le niveau de confiance en tapant

```
int.conf.prop(pr, n, NC)
```

ou encore, avec un ordre quelconque des arguments :

```
int.conf.prop(pr = pr, n = n, NC = NC)
```

REMARQUE 4.10.

(1) Parfois le nombre $\sqrt{\frac{p_r(1-p_r)}{n}}$ est appéle erreur standart de proportion (ou standart error of proportion) et notée SEP, de sorte que l'intervalle de confiance vaut $[pr - z\,SEP, pr + z\,SEP]$ et que sa largeur vaut $2z\,SEP$.

(2) Remarquons aussi, grâce à ce qui a été observé dans la section 4.1.2 que la largeur $2z\,SEP$
   – diminue quand n augmente ;
   – dimimue quand NC diminue.

♢

4.3.4. Applications

Tout contexte binomial pourra être utilisé quand on étudie une "grande population" (en théorie infinie), qui se décompose en deux groupes, de proportions respectives p et $1-p$. Si on prend "au hasard", un individu dans cette population, la probabilité qu'il soit dans le premier groupe est donc p. Si on prend maintenant au "hasard", n individus, on réalise donc une expérience binomiale de paramètre n et p.

En pratique, sur un échantillon de taille n, on observe la proportion expérimentale de succès pr, à partir de laquelle on infére la probabilité théorique p, inconnue, grâce à la détermination de l'intervalle de confiance au niveau NC.

Attention, on rappelle que si l'on construit ainsi un grand nombre d'intervalles de confiance, la proportion effective des intervalles qui contiennent rééllement p sera NC, tandis que la proportion effective des intervalles qui ne contiennent pas p sera $1-NC$! Autrement dit, une construction de l'intervalle de confiance (certes dans un rare nombre de cas) sera mauvaise ! !

Exemple 4.11. Reprenons maintenant les simulations du début de la section 4.3.1.
Créons de nouveau un échantillon binomial correspondant à $n = 1000$ et $p = 0.3$:

- *Avec Rcmdr :*
  On utilise le menu déroulant "Distributions", l'option "Distributions discrètes" puis "Distribution binomiale", puis "Echantillon d'une distribution binomiale". Dans la fenêtre dialogue, il faut indiquer pour
  - "Nombre d'essais" (valeur de n) : 1000,
  - "Probabilité de succès" (valeur de p) : 0.3,
  - "Nombre d'échantillons" : 1,
  - "Nombre d'observations" : 1
- *Sans Rcmdr :*
  Il faut taper les commandes suivantes :

```
dede<-rbinom(n=1, size=1000, prob=0.3)
```

J'obtiens, par exemple, *pour mes valeurs* la valeur suivante :

$$305 \text{ et donc } p_r = \frac{305}{1000} = 0.305. \tag{4.26}$$

Supposons qu'une autre simulation me donne

$$331 \text{ et donc } p_r = \frac{331}{1000} = 0.331. \tag{4.27}$$

Construisons pour ces deux valeurs un intervalle de confiance au niveau $NC = 0.95$. On a successivement

$$z = 1.959964$$

et donc pour la valeur donnée par (4.26) un intervalle de confiance dont les bornes sont

$$\begin{aligned} pr \pm z\sqrt{\frac{pr(1-pr)}{n}} &= 0.305 \pm 1.959964\sqrt{\frac{0.305(1-0.305)}{1000}}, \\ &= 0.305 \pm 0.028536 \end{aligned}$$

et donc un intervalle de confiance

$$[0.2764642, 0.3335358].$$

La valeur du paramètre ($p = 0.3$) est bien, dans ce cas, compris dans l'intervalle de confiance. Il faut bien comprendre que le contraire ne se produit qu'une fois sur 20.

De même, l'intervalle de confiance défini par la valeur (4.27) me donnerait un intervalle de confiance dont les bornes sont

$$pr \pm z\sqrt{\frac{pr(1-pr)}{n}} = 0.331 \pm 1.959964\sqrt{\frac{0.331(1-0.331)}{1000}},$$
$$= 0.331 \pm 0.029166$$

et donc un intervalle de confiance

$$[0.3018341, 0.3601659].$$

On n'est dans les 5 % des cas non chanceux où l'intervalle de confiance ne contient pas le paramètre p ! Ici, la probabilité de sortie de la valeur 331 vaut 0.0028362 plus faible que la probabilité de sortie de la valeur 305, égale à 0.0258145 et qui a donné un intervalle contenant p. Si on faisait une étude complète des cas où on trouve un intervalle ne contenant pas p, on trouverait une probabilité expérimentale approchant 5% !

Remarque 4.12. *L'exemple 4.11 n'est "que" d'ordre pédagogique, puisque la valeur du paramètre inféré est connue, ce qui n'est jamais le cas en pratique, comme dans l'exercice 4.13.*

Exercice 4.13. Quelle représentation de l'enseignant d'éducation physique et sportive (EPS) ont ses collègues d'autres disciplines ? Trois cent trente et un professeurs de collèges et lycées ont accepté de répondre à un questionnaire consacré à ce thème. À part quinze d'entre eux, ils se sont prononcés sur la question : "comparativement aux autres enseignants, l'enseignant d'EPS a une charge de travail :

- moins importante (121 réponses),
- équivalente (185 réponses) ou
- plus importante (10 réponses) ?"

Ici, il n'y a pas véritablement une population clairement définie, sinon une hypothétique population d'enseignants dont une partie pense que... On la "remplace" par un modèle binomial. Nous allons calculer un intervalle de confiance concernant la probabilité qu'un enseignant estime que la charge de travail de l'enseignant d'EPS est moindre. La taille de l'échantillon est

$$n = 121 + 185 + 10 = 316.$$

et la proportion observée dans l'échantillon est donc de

$$pr = \frac{121}{316} = 0.3829.$$

(1) Calculer un intervalle de confiance au niveau $NC = 0.95$ de la proportion.

(2) Calculer un intervalle de confiance au niveau $NC = 0.9$ de la proportion.

(3) Supposons que la proportion observée reste la même mais que la taille de l'échantillon soit de $n = 50$. Calculer un intervalle de confiance de la proportion au niveau $NC = 0.95$.

(4) Et si la taille de l'échantillon est de $n = 5000$?

Voir éléments de correction page 54.

Exercice 4.14. Un questionnaire envoyé par la Caisse Nationale d'Assurance Maladie à un échantillon représentatif de la population française visait à connaître les risques sportifs. Sur 7408 accidents déclarés, 1037 sont le fait de la pratique sportive.

- Calculer le pourcentage que constituent les accidents sportifs.
- Calculer un intervalle de confiance au niveau $NC = 95$ % de cette proportion ? Que pensez-vous de la précision des résultats ? À quoi est-elle due ?

Exercice 4.15. Après une enquête sur un échantillon de 45 étudiants de L3 et M1 APA, on a constaté que 29 d'entre eux avaient, dans leur entourage, une personne présentant un handicap. Estimer, par intervalle de confiance au niveau 0.95, la proportion d'étudiants ayant, dans leur entourage, une personne handicapée.

4.4. Intervalle de confiance d'une moyenne

Nous allons maintenant reproduire, rapidement, ce que nous avons fait en section 4.3.

4.4.1. Construction à partir d'un exemple

On rappelle qu'en section N.2.3, on a vu que si l'on considérait un grand nombre de tirages d'une variable aléatoire X issue d'une loi normale de moyenne μ et d'écart-type σ, l'histogramme obtenu se rapprochait de la de la loi de probabilité continue.

Plus précisément, on peut aussi générer un certain nombre d'échantillon normaux, c'est-à-dire issus de la loi normale.

Nous allons procéder comme dans la section 4.3.1. Nous allons ensuite développer la notion d'intervalle de confiance.

Nous allons supposer que les données sont générées par un modèle normal d'espérance $\mu = 15$ et d'écart-type $\sigma = 3$.

- *Avec Rcmdr :*

  Commençons par générer un échantillon de n=16 valeurs issues de cette loi grâce à au logiciel Rcmdr. Il faut choisir le menu déroulant "Distributions", puis "Distributions continues" puis "Distribution normale" puis "Echantillon d'une distribution normale". Dans la fenêtre de dialogue qui s'ouvre il faut indiquer pour
  - "mu" : 15,
  - "sigma" :3,
  - "Nombre d'échantillons" : 1,
  - "Nombre d'observations" : 16

  Il faut également cocher dans la rubrique "Ajouter au jeu de données" les cases moyennes et écarts-types (et ne pas cocher la rubrique "somme"). Il se crée alors un tableau à 1 ligne et 18 colonnes, les deux dernières indiquant *pour ma machine* une moyenne de 15.02842 (et incidemment un écart-type de 2.250491) :

```
            obs1     obs2     obs3     obs4     obs5     obs6     obs7     obs8
sample1 15.47796 15.92321 18.65116 19.92444 11.60927 15.08626 15.00156 16.14213
            obs9    obs10    obs11    obs12    obs13    obs14    obs15    obs16
sample1 15.15968 11.62515 16.05007 13.53465 14.85276 12.40743 15.64765 13.36134
            mean       sd
sample1 15.02842 2.250491
```

- *Sans Rcmdr :*

  Il faut télécharger et utiliser la fonction '`make.echant.norm`' et taper les commandes suivantes :

```
EchantillonsNormaux<-make.echant.norm(mu=15,sigma=3,li=1,co=16)
```

  La fonction '`rnorm`' de R pourrait suffir, mais il faudrait faire d'autres manipulations pour définir la moyenne. ♢

Comme précédemment nous allons considérer que cet échantillon est un parmi d'autres.

Attention, dans la section 4.3.1, un échantillon consistait en la donnée d'une proportion, créée aléatoirement. Ici, un échantillon consiste en la donnée de 16 nombres. Pour considérer que cet échantillon est un parmi d'autres, nous allons donc maintenant créer un tableau rectangulaire de nombres. Chaque ligne de ce tableau sera un échantillon.

Exercice 4.16.

(1)
- *Avec Rcmdr :*

  En choisissant :
  - "mu" : 15,
  - "sigma" :3,

– "Nombre d'échantillons" : 10000,
– "Nombre d'observations" : 16
générer $p = 10000$ échantillons semblables au précédent.

- *Sans Rcmdr :*
  Taper les commandes suivantes :
  ```
  EchantillonsNormaux<-make.echant.norm(mu=15,sigma=3,li=10000,co=16)
  ```

Vérifiez que vous avez généré $p = 10000$ échantillons semblables au précédent, c'est-à-dire, qu'il se crée alors un tableau à 10000 lignes et 18 colonnes.

(2) Représenter par un histogramme la distribution d'échantillonnage de la moyenne. Calculer la moyenne de ces moyennes et l'écart-type de ces moyennes.

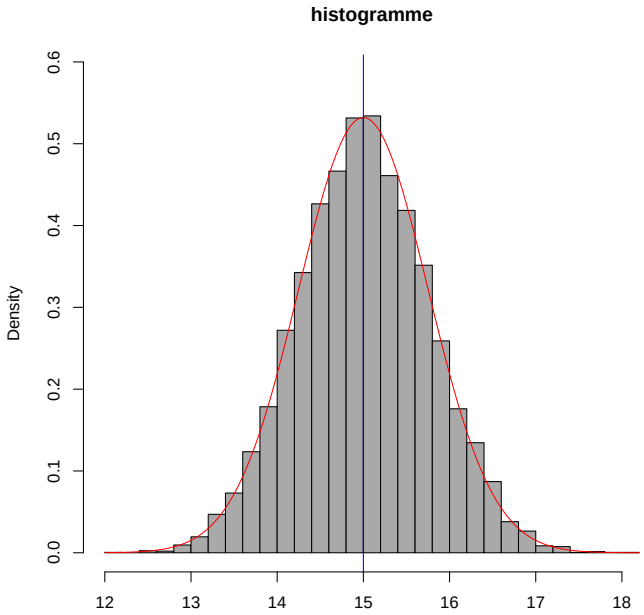


FIGURE 4.6. Histogramme de la distribution des moyennes (échantillon de taille n=16, 10000 tirages) observées sur une loi normale de moyenne $\mu = 15$ et d'écart-type $\sigma = 3$ en densité avec 30 classes. Sur ce graphe se trouve aussi la loi normale de moyenne $\mu = 15$ et d'écart-type $\sigma/\sqrt{n} = 0.75$ et la droite bleue d'abscisse $\mu = 15$.

Vous devriez obtenir pour l'histogramme une figure analogue à celle du graphique 4.6. Cette figure rappelle fortement la figure 4.5 page 39 !

(3) La moyenne et l'écart-type de toutes les moyennes (c'est-à-dire de l'avant dernière colonne de mon échantillon) que j'obtiens sur *mon* échantillon sont

$$m = 15.005754, \tag{4.28a}$$
$$sd = 0.757275 \tag{4.28b}$$

Comparez avec les votres. Comparez avec les nombres suivants (où μ et σ sont les moyenne et écart-type de la loi normale initiale) :

$$\mu = 15, \tag{4.29a}$$

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{4} = 0.75 \tag{4.29b}$$

(4)

Conclusion partielle

Comme précédemment, on *pourrait* conclure que la moyenne des échantillons suit une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$ et en déduire la construction de l'intervalle de confiance de la moyenne de la loi normale, paramètre ici inféré !

Nous allons voir que la réalité est un tout petit peu plus difficile !

(5) Reprendre l'échantillon normal crée précédemment et tracer l'histograme non plus des moyennes mais de la variable suivante

$$t = \sqrt{n}\,\frac{m-\mu}{sd} = 4\,\frac{m-15}{sd} \tag{4.30}$$

où n (ici égal à 16) est la taille de l'échantillon, m est sa moyenne mesurée sd son écart-type *mesuré*.

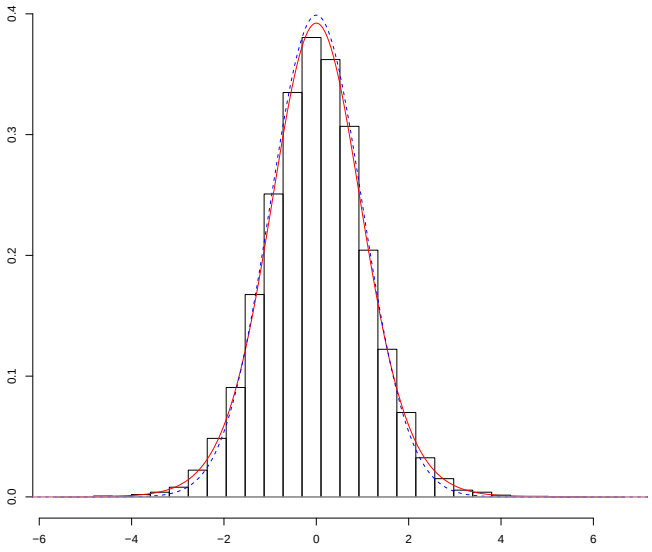


FIGURE 4.7. Histogramme de la distribution des t (définie par (4.30) défini à partir de l'échantillon 'EchantillonsNormaux' crée précédemment en densité avec 30 classes. En dessous se trouvent la loi de student correspondant à ddl=15 (en rouge, trait plein) et la loi normale (en bleu, trait pointillé).

Vous devriez obtenir l'histogramme de la figure 4.7. Sur cette figure, se trouvent aussi la loi de student (voir section 3.5 page 24) correspondant à ddl=15 (en rouge, trait plein) et la loi normale (en bleu, trait pointillé). On constate un léger écart entre la loi de Student et la loi normale. La loi de Student est en théorie plus proche de l'histogramme que la loi normale.

Ainsi, dans $100 \times NC\%$ des cas, on aura (le quantile t étant associé à NC grâce à (4.4))

$$-z \leq \sqrt{n}\,\frac{m-\mu}{sd} \leq z$$

ce qui est équivalent à (en remplaçant z par t)

$$-t \leq \sqrt{n}\,\frac{m-\mu}{sd} \leq t$$

soit encore

$$m - t\frac{sd}{\sqrt{n}} \leq \mu \leq m + t\frac{sd}{\sqrt{n}}$$

On a peut donc construire un intervalle de confiance de la moyenne $\left[m - t\frac{sd}{\sqrt{n}}, m + t\frac{sd}{\sqrt{n}}\right]$, supposant que la moyenne théorique et l'écart-type théorique sont inconnus.

On pourra consulter par exemple `http://fr.wikipedia.org/wiki/Loi_de_Student` ou la page 215 de [**9**].

4.4.2. Théorie

Bref, on admet que

PROPOSITION 4.17 (Intervalle de confiance "d'une moyenne"). *On suppose que l'on réalise un tirage aléatoire d'un échantillon consistant en la donnée de n valeurs issues d'une loi normale. On note respectivement m et sd la moyenne et l'écart-type* estimés *(c'est-à-dire, calculé à partir de l'échantillon). L'intervalle de confiance au niveau de confiance NC de la moyenne est donné par*

$$\left[m - t\frac{sd}{\sqrt{n}}, m + t\frac{sd}{\sqrt{n}}\right] \tag{4.31}$$

Le coefficient multiplicateur t est obtenu sous R *de la façon suivante :*

- Avec Rcmdr :

  Aller dans le menu "distribution", puis "distribution continue", puis "distribution t", puis "quantiles t", taper dans le champ "probabilités", le nombre `(1+NC)/2`*, où* `NC` *est le niveau de confiance, et dans le champ "degrés de liberté" le nombre égal à* $n-1$.
- Sans Rcmdr :

  Grâce à la ligne de commande

```
qt((1 + NC)/2, df = n - 1)
```

4.4.3. Avec R

Pour évaluer cet intervalle de confiance dans R, plusieurs solutions :

(1) Une fois le nombre t défini à partir de la proposition 4.17, on peut taper dans "Rgui"

```
m - t * sd/sqrt(n)
```

puis

```
m + t * sd/sqrt(n)
```

Mieux, on pourra taper directement

```
m + c(-1, 1) * t * sd/sqrt(n)
```

(2) Comme dans l'annexe M page 185, vous pouvez aussi télécharger et sourcer la fonction `int.conf.moy.R` qui détermine l'intervalle de confiance en fonction de

- `mu` : moyenne mesurée
- `sd` : écart-type (déviation standart) mesuré ;
- `n` : la taille de l'échantillon ;
- `NC` : seuil de confiance.

en tapant

```
int.conf.moy(mu = mu, sd = sd, n, NC)
```

ou encore, avec un ordre quelconque des arguments :

```
int.conf.moy(mu = mu, sd = sd, n = n, NC = NC)
```

(3) On peut aussi procéder ainsi si on dispose des données, et pas seulement des statistiques :

- *Avec Rcmdr :*
  Il suffit de choisir le menu déroulant "Statistiques", puis les options "Moyennes" et "t-test univarié". Il faut laisser les champs relatifs aux hypothèses avec les valeurs définies par défaut.
- *Sans Rcmdr :*
  On tape dans "Rgui" : (`x` est le tableau des données)
  ```
  t.test(x, conf.level = NC)
  ```
  ou plus rapidement
  ```
  t.test(x, conf.level = NC)$conf.int
  ```

REMARQUE 4.18. De façon analogue à la remarque 4.10, notons que :

(1) Parfois le nombre $\sqrt{sd}/n$ est appéle erreur standart de la moyenne (ou standart error of the mean) et notée SEM, de sorte que l'intervalle de confiance vaut $[m - t\,SEM, m + t\,SEM]$ et que sa largeur vaut $2t\,SEM$.

(2) Remarquons aussi, grâce à ce qui a été observé dans la section 4.1.2, que la largeur $2t\,SEM$
- diminue quand n augmente ;
- dimimue quand NC diminue.

♢

4.4.4. Applications

EXERCICE 4.19. Pour les tailles d'échantillons (n) et niveaux de confiance (NC) suivants, trouver les valeurs du coefficient multiplicateur t à utiliser pour construire l'intervalle de confiance de la moyenne

(1) $n = 25$, NC =95 % ;

(2) $n = 25$, NC =90 % ;

(3) $n = 15$, NC =99 % ;

(4) $n = 15$, NC =98 % ;

Voir éléments de correction page 55.

EXERCICE 4.20. R. Rollier (M2PPMR) a pris en charge l'entraînement de 39 joueurs de moins de 19 ans de l'ASVEL Rugby. Un test de vitesse (30 mètres avec deux changements de direction) a été passé par tous ces joueurs. Le temps (sec.) a été relevé. Voir le fichier de données '`rollier.txt`' (voir la variable '`Temps`').

(1) Représenter graphiquement les données de temps

(2) Ces valeurs semblent-elles approximativement suivre une loi normale ?

(3) Calculer un intervalle de confiance aux niveaux $NC = 0.95$ et $NC = 0.9$ de la moyenne.

Voir éléments de correction page 55.

Concluons par quelques exercices issus de [**8**].

EXERCICE 4.21. Dans un centre médico-sportif, on a mesuré la taille de 71 footballeurs du département du Rhône. La moyenne est 177 cm pour un écart-type 5.655. Déterminer un intervalle de confiance de la taille moyenne des footballeurs du Rhône.

EXERCICE 4.22. On connait la taille (en m) de 12 basketteurs américains (in *Mondial Basket*, juillet-août 1994).

```
taibask <- c(2.08, 2.01, 2.03, 2.1, 1.98, 2.08, 1.85, 2.03, 2.16,
    2.01, 1.91, 1.88)
```

(1) Donner la moyenne et l'écart-type estimés de la population à partir de l'échantillon.

(2) Donner les intervalles de confiance de la moyenne de la population aux niveaux de confiance 0.95 et 0.99.

(3) On suppose que la taille suit une loi normale dont les paramètres ont été estimés dans la première question. Quelle est la probabilité pour un basketteur d'avoir une taille supérieure à 2.05 m.

4.4.5. Autres intervalles de confiance liés à loi normale

Section facultative.

4.4.5.1. Loi de student pour les grandes valeurs de n et la loi normale.

En fait, la loi de Student tend vers la loi normale centrée réduite quand n tend vers l'infini, comme le montre la figure 4.8.

Ainsi, la remarque faite au point 4 page 46 est justifié *a posteriori* :

Pour de grandes valeurs de *ddl* (supérieures à 30 ou 60, selon la littérature ...), la loi de Student est approché par une loi normale centrée réduite de sorte que on remplace (4.30) par

$$\sqrt{n}\,\frac{m-\mu}{sd} \rightsquigarrow \mathcal{N}(0,1).$$

Ainsi,

- on calcule un nombre z associé au niveau de confiance NC par la loi normale centrée réduite dans (4.4) ;
- on détermine l'intervalle de confiance pour la moyenne μ

$$\left[m - z\frac{sd}{\sqrt{n}}, m + z\frac{sd}{\sqrt{n}}\right] \tag{4.32}$$

Cette approximation, historiquement très utilisée quand les ordinateurs n'existaient pas, n'a plus lieu d'être faite !

On pourra consulter la fiche [**10**].

4.4.5.2. Cas où l'écart-type de la population est connue.

Si le nombre σ est connu, il n'est plus la peine d'introduire la loi de Student ; en effet, dans ce cas, la variable aléatoire $\sigma(m-\mu)/\sigma$ suit une loi normale centrée réduite et l'intervalle de confiance au niveau NC est encore donné par

$$\left[m - z\frac{\sigma}{\sqrt{n}}, m + z\frac{\sigma}{\sqrt{n}}\right] \tag{4.33}$$

4.4.5.3. Intervalles de confiance pour l'écart-type.

On peut aussi donner des intervalles de confiance pour l'écart-type, en supposant que la moyenne est connue ou non. Voir les pages 216-217 de [**9**].

♢

4.5. Exercices supplémentaires

Exercice 4.23. Imaginez une façon de vérifier, *a posteriori*, que dans les simulations menant à la figure 4.5 page 39, la proportion du nombre d'intervalles de confiance contenant le paramètre inféré est bien proche de NC.

Voir éléments de correction page 58.

Exercice 4.24. On s'intéresse à l'expérience aléatoire" jeter 30 fois une pièce de monnaie" où l'on note le nombre de face (qui est considéré comme un succès).

(1) Quelles sont les valeurs des paramètres dans cette expérience ?

(2) J'ai obtenu 18 fois face, soit $pr = 18/30 = 0.6$. Calculer l'intervalle de confiance un niveau de confiance $NC = 0.95$. Contient-il la valeur du paramètre ?

(3) Simuler cette expérience à l'aide du logiciel Rcmdr. Quelle proportion trouvez-vous ? Calculer un intervalle de confiance à $NC = 0.95$. Comprend-il la valeur du paramètre ?

(4) Tapez la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(1, size=30, prob=0.5)
```

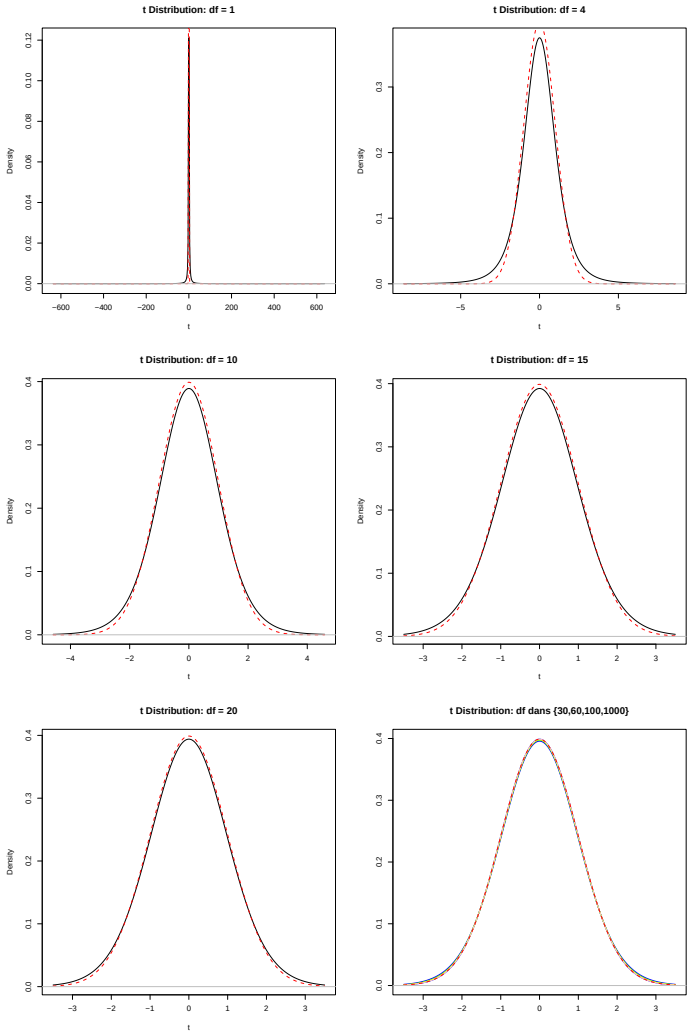


FIGURE 4.8. Tracés des loi de student pour quelques valeurs de l'entier n : n=1, voir graphique numéro 1 ; n=4, voir graphique numéro 2 ; n=10, voir graphique numéro 3. n=15, voir graphique numéro 4. n=20, voir graphique numéro 5. Pour $n \in \{30, 60, 100, 1000\}$, voir dernière figure ; sur l'ensemble des figures, la loi normale centrée réduite est tracé en rouge pointillé.

qui simule en fait 1 fois directement la manipulation précédente. Commentez !

(5) *Questions facultatives*

(a) Tapez la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(20, size=30, prob=0.5)
```

qui simule en fait 20 fois directement la manipulation précédente.
Qu'observez-vous ?

(b) Pour chacun des groupes de commandes donnés, vous les taperez les unes à la suite des autres en tapant sur la touche "enter" après chaque commande.

(i) (A) Récupérez et sourcez la fonction `int.conf.prop.R` et tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
IC<-int.conf.prop(rbinom(1, size=30, prob=0.5)/30,30,0.95);
(0.5>=IC[1])&(IC[2]>=0.5)
```

qui calcule 1 intervalle de confiance et renvoie T (TRUE) s'il contient la valeur du paramètre p et F (FALSE) sinon. Refaites plusieurs fois cette commande (grâce à la flèche ↑ de votre clavier). Commentez.

(B) *Organiser un sondage sur l'ensemble des étudiants !* Conclure !

(ii) Tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
res<-rbinom(25, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=25)
for(i in 1:25) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
(0.5>=IC[,1])&(IC[,2]>=0.5)
```

qui calcule 25 intervalles de confiance et renvoie T (TRUE) si ils contiennent la valeur du paramètre p et F (FALSE) sinon. Commentez.

(iii) Tapez les commandes suivante directement dans la fenêtre de "Rgui"

```
res<-rbinom(1000, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=1000)
for(i in 1:1000) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
100*sum((0.5>=IC[,1])&(IC[,2]>=0.5))/1000
```

qui calcule 1000 intervalles de confiance et qui renvoie directement la poportion d'intervalle de confiance construits qui contiennent la valeur du paramètre inféré.

Qu'osbservez-vous ? Commentez !

(iv) Refaite la même suite de commandes, mais en prenant le nombre de jeters $n = 100$, puis $n = 1e + 05$. Commentez !

Voir éléments de correction page 59.

EXERCICE 4.25. Comme dans l'exercice 4.23, utilisez la fonction et tapez

```
verifie.int.conf.prop(p, n, NC, q)
```

pour

$$n = 100,$$
$$NC = 0.95,$$
$$q = 10000,$$

et

$$p \in \{0.5, 0.1, 0.001, 1e - 05\}.$$

Commentez !

Voir éléments de correction page 60.

EXERCICE 4.26.

Récupérez et sourcez la fonction `test.sondage.R` qui simule le travail des instituts de sondages en période pré-électorale pour un second tour d'élection avec seulement deux candidats (processus très simplifié, car en réalité, on utilise la méthode des quotats, dont on extrapole certains résultats) : on introduit

- `tpop`, la taille de population des électeurs ;
- `techan`, la taille de l'échantillon choisis (tirés aléatoirement dans la population électorale) ;

- `p0`, la probabilité théorique de succès (ici égale à la proportion de voies du gagnant) ;
- `NC`, le niveau de confiance (argument optionnel, égal par défaut à 0.95)

Cette fonction simule une population de taille `tpop`, détermine aléatoirement un groupe d'électeurs votant pour le gagnant de cardinal `tpop`× `p0`, puis réalise un échantillon aléatoire de taille `techan` et en déduit une proportion et un intervalle de confiance au seuil `NC`.

(1) Faites quelques simulations en prenant par exemple

```
 test.sondage(1000, 100, 0.53)
$prop
[1] 0.52

$intc
[1] 0.4220802 0.6179198
 test.sondage(10000, 1000, 0.53)
$prop
[1] 0.514

$intc
[1] 0.4830224 0.5449776
 test.sondage(1e+05, 1000, 0.53)
$prop
[1] 0.528

$intc
[1] 0.4970589 0.5589411
 test.sondage(1e+05, 1000, 0.53, 0.99)
$prop
[1] 0.555

$intc
[1] 0.5145197 0.5954803
 test.sondage(1e+05, 1000, 0.53, 0.999)
$prop
[1] 0.538

$intc
[1] 0.4861227 0.5898773
```

Commentez !

(2)

On s'intéresse au second tour de l'élection présidentielle de 2007, pour laquelle, on a donné dans le tableau 4.3 page ci-contre les résultats officiels (pour l'ensemble de la France) issus de l'url :

`http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/index.html`

Essayez de faire les simulations suivantes (qui risquent de ne pas fonctionner, pour cause de mémoire insuffisante) :

```
 test.sondage(35773578, 1000, 18983138/35773578)
$prop
[1] 0.516
```

	nombre	pourcentage (/Inscrits)
Inscrits	44 472 733	100
Votants	37 342 004	83,97
	nombre	pourcentage (/Votants)
Blancs ou Nuls	1 568 426	4,20
Exprimés	35 773 578	95,80
	nombre	pourcentage (/Exprimés)
M. Nicolas SARKOZY	18 983 138	53,06
Mme Ségolène ROYAL	16 790 440	46,94

TABLE 4.3. Résultats officiels du second tour de la présidentielle 2007

```
$intc
[1] 0.4850261 0.5469739
 test.sondage(35773578, 1000, 18983138/35773578, 0.9)
$prop
[1] 0.542

$intc
[1] 0.4850261 0.5469739
 test.sondage(35773578, 1e+05, 18983138/35773578)
$prop
[1] 0.52849

$intc
[1] 0.5253961 0.5315839
```

Commentez !

(3) Si votre ordinateur ne vous permet pas de faire des simulations pour la France entière, restreniez-vous par exemple à l'Ile-de-France
`http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/011/011.html`
voire à Paris
`http://www.interieur.gouv.fr/sections/a_votre_service/resultats-elections/PR2007/011/075/1175.html`

(4) Refaites les simulations précédentes en imaginant un score proche du résultat du second tour de l'élection présidentielle de 1974 : Giscard : 50,81 % et Mitterand : 49,19 % ; voir
`http://fr.wikipedia.org/wiki/élection_présidentielle_française_de_1974`

```
 test.sondage(26367807, 1000, 13396203/26367807)
$prop
[1] 0.52

$intc
[1] 0.4890351 0.5509649
```

Commentez !

4.6. Éléments de correction

Éléments de correction de l'exercice 4.2

On obtient successivement $q_1 = -0.842$, $q_2 = 0.524$, $q_3 = -Inf$, $q_4 = Inf$, $q_5 = NaN$

Il est normal que les quantiles associés à 0 et 1 valent $-\infty$ et ∞. Le qantile associé à 1.2 n'est pas défini, car cette probabilité n'est pas dans l'intervalle $[0, 1]$!

Éléments de correction de l'exercice 4.3

On obtient successivement $q_1 = -0.183$, $q_2 = 2.549$, $q_3 = -Inf$, $q_4 = Inf$, $q_5 = NaN$

Il est normal que les quantiles associés à 0 et 1 valent $-\infty$ et ∞. Le qantile associé à 1.2 n'est pas défini, car cette probabilité n'est pas dans l'intervalle $[0, 1]$!

Éléments de correction de l'exercice 4.6

On obtient par exemple en ligne de commande

```
qnorm((1 + 0.6)/2)
```

```
[1] 0.8416212
```

et donc le z associé au Niveau de confiance $NC = 0.60$ vaut $z = 0.841621$. Pour les plus avertis, on pourra taper

```
NC <- c(0, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999, 1)
qnorm((1 + NC)/2)
```

```
[1] 0.0000000 0.6744898 0.8416212 1.0364334 1.2815516 1.6448536 1.9599640
[8] 2.5758293 3.2905267       Inf
```

voire

```
NC <- c(0, seq(50, 90, by = 10)/100, 0.95, 0.99, 0.999, 1)
qnorm((1 + NC)/2)
```

Éléments de correction de l'exercice 4.13

En tapant par exemple

```
pr<-121/(121+185+10)
int.conf.prop(pr,316,0.95)
int.conf.prop(pr,316,0.9)
int.conf.prop(pr,50,0.95)
int.conf.prop(pr,5000,0.95)
```

on obtient successivement les intervalles

$$[0.329316, 0.4365068],$$
$$[0.3379327, 0.4278901],$$
$$[0.2481747, 0.5176481],$$
$$[0.3694377, 0.3963851].$$

Pour obtenir le premier intervalle de confiance, on peut aussi déterminer la valeur de z de la proposition 4.9 page 41 :

```
NC <- 0.95
z <- qnorm((1 + NC)/2)
```

et taper ensuite

```
pr <- 121/(121 + 185 + 10)
n <- 121 + 185 + 10
```

puis

```
pr - z * sqrt(pr * (1 - pr)/n)
```

```
[1] 0.329316
```

puis

```
pr + z * sqrt(pr * (1 - pr)/n)
```

```
[1] 0.4365068
```

ou mieux

```
pr + z * c(-1, 1) * sqrt(pr * (1 - pr)/n)
```

```
[1] 0.3293160 0.4365068
```

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.19

On trouve successivement

(1) pour $n = 25$, NC =95 % : t=2.0638986 ;

(2) pour $n = 25$, NC =90 % : t=1.7108821 ;

(3) pour $n = 15$, NC =99 % : t=2.9768427 ;

(4) pour $n = 15$, NC =98 % : t=2.6244941.

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.20

(1) Voir la figure 4.9 qui représente à la fois l'histogramme et le graphe quantile-quantile.

Voir la section E.3.2.5 page 95 pour le graphe quantile-quantile.

On rappelle ici le principe du graphe quantile-quantile.

Le graphe quantile-quantile cherche à confronter l'histogramme des données à une forme prototypique, celle de la loi normale dite aussi courbe en cloche. Il est très important de déterminer si les données suivent approximativement cette forme car les procédures statistiques que nous verrons par la suite sont généralement basées sur cette hypothèses. Lorsque les données suivent la loi normale, les points sont situés exactement sur une droite. Toutefois, un écart est inévitable, l'écart normal étant symbolisé par deux courbes en pointillées sur le graphe.

- *Avec Rcmdr :*
  On trouvera cela dans le menu "graphique".
- *Sans Rcmdr :*
  Il faut (éventuellement d'abord installer) et charger le package car et donc taper
  ```
  library(car)
  ```
  puis on tapera
  ```
  qq.plot(rollier$Temps)
  ```

(2) La courbe est à peu près en "cloche" (sauf peut-être quelques valeurs extrémales) et donc les données peuvent être considérées comme normale.

(3) (a) Donnons pour le niveau de confiance NC les différentes méthodes citées juste après la proposition 4.17 page 47

(i) Cas 1 page 47 :

On calcule le nombre t ainsi :

- *Avec Rcmdr :*
  On calcule d'abord sous "Rgui", le nombre `(1+NC)/2`=0.975. Aller ensuite dans le menu "distribution", puis "distribution continue", puis "distribution t", puis "quantiles t" et rentrer à la place de "probabilités", le résultat `(1+NC)/2`=0.975 et à la place de "degrés de liberté" le nombre égal à $n - 1 = 39$. On obtient $t = 2.024394$

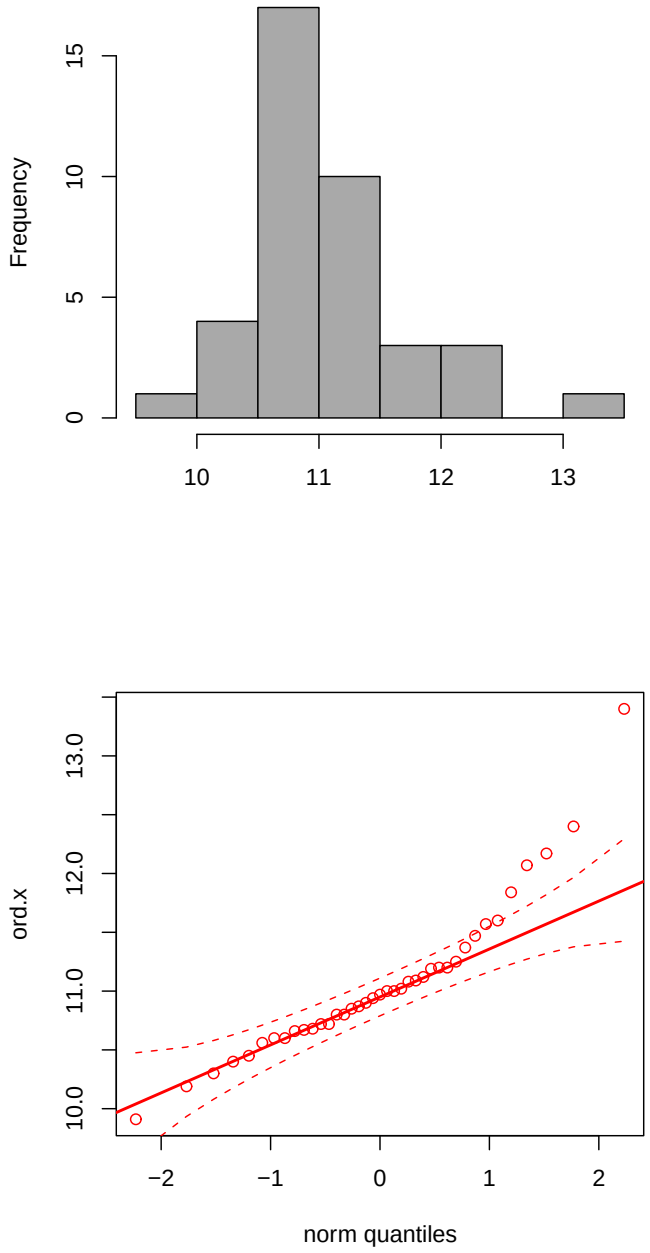


FIGURE 4.9. Histogramme et graphe quantile-quantile de la variable '`Temps`' du fichier '`rollier.txt`'.

- *Sans Rcmdr :*
  Grâce à la ligne de commande
```
 qt((1 + NC)/2, df = n - 1)
[1] 2.024394
```

On calcule ensuite la moyenne et l'écart-type estimé des données ; on obtient

$$m = 11.067436, \quad sd = 0.6517294$$

On peut taper dans "Rgui" directement

```
11.067436+c(-1,1)*2.024394*0.6517294/sqrt(39)
```

ce qui donne

```
[1] 10.85617 11.27870
```

soit un intervalle de confiance

$$[10.85617, 11.278702]. \tag{4.34}$$

Pour $NC = 0.9$, on obtient

$$[10.891489, 11.243382]. \tag{4.35}$$

(ii) Cas 2 page 47 : On tape (après avoir téléchargé et sourcé)

```
int.conf.moy(11.067436,0.651729,39,0.95)
```

ou

```
int.conf.moy(mu=11.067436,sd=0.651729,n=39,NC=0.95)
```

ou encore

```
int.conf.moy(11.067436,0.651729,39)
```

```
[1] 10.85617 11.27870
```

ce qui redonne bien (4.34).

(iii) Cas 3 page 48

- *Avec Rcmdr :*
  Il suffit de choisir le menu déroulant "Statistiques", puis les options "Moyennes" et "t-test univarié". Il faut laisser les champs relatifs aux hypothèses avec les valeurs définies par défaut.
- *Sans Rcmdr :*
  On tape dans "Rgui" :
  ```
  t.test(rollier$Temps, conf.level=0.95)
  ```
  ce qui donne
  ```
          One Sample t-test

  data:  x[, indd]
  t = 106.0503, df = 38, p-value < 2.2e-16
  alternative hypothesis: true mean is not equal to 0
  95 percent confidence interval:
   10.85617 11.27870
  sample estimates:
  mean of x
   11.06744
  ```
  ou encore directement
  ```
  t.test(rollier$Temps, conf.level=0.95)$conf.int
  ```
  ce qui donne
  ```
  [1] 10.85617 11.27870
  attr(,"conf.level")
  [1] 0.95
  ```

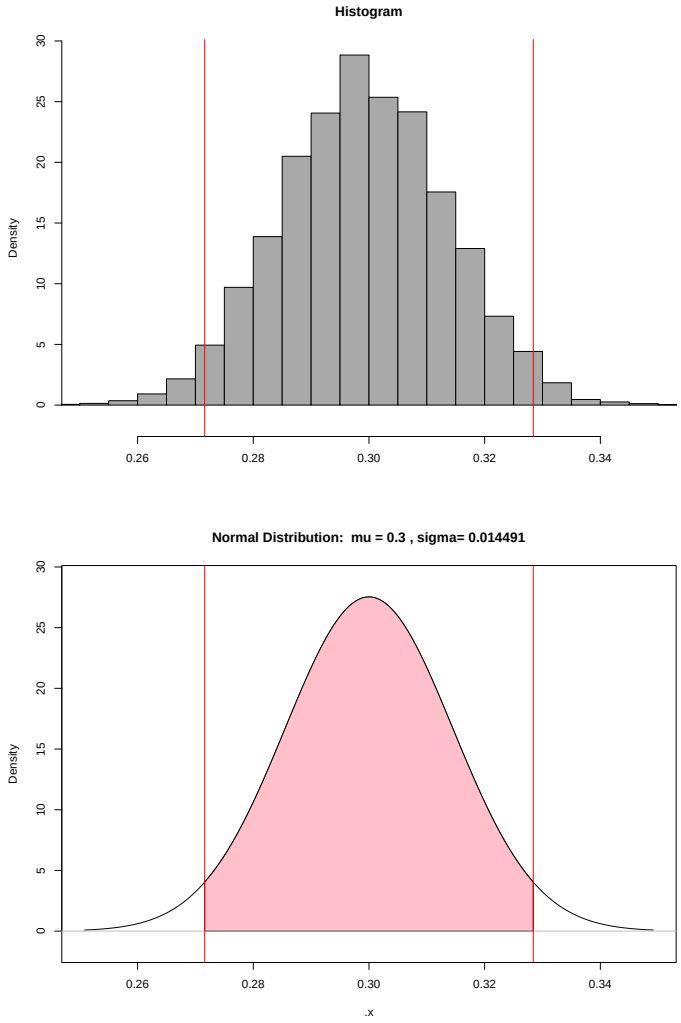


FIGURE 4.10. Histogramme de la distribution d'échantillonnage (10000 tirages) d'une proportion et le graphique de la loi normale de moyenne $m = 0.3$ et d'écart-type $\sigma = 0.014491$ avec les deux droites correspondant aux abscisses 0.271597 et 0.328403

ÉLÉMENTS DE CORRECTION DE L'EXERCICE 4.23

Voir la figure 4.10 qui met en évidence que 95% des données de l'histogrammes des proportions se trouvent dans l'intervalle de confiance NC ici égal à $[p_{\min}, p_{\min}] = [0.271597, 0.328403]$.

Justifions cela expérimentalement en tapant la séquence suivante dans "Rgui" qui dénombre le nombre de proportions dans cet intervalle et qui en détermine le pourcentage (une fois que l'on fait les manipulations page 39) :

- *Avec Rcmdr :*

```
sum((EchantillonsBinomiaux$obs/1000>=0.2715974)&
(EchantillonsBinomiaux$obs/1000<=0.3284026))
```

puis

```
100*sum((EchantillonsBinomiaux$obs/1000>=0.2715974)&
(EchantillonsBinomiaux$obs/1000<=0.3284026))/10000
```

- *Sans Rcmdr :*
  ```
  sum((dede>=0.2715974)&(dede<=0.3284026))
  ```
  puis
  ```
  100*sum((dede>=0.2715974)&(dede<=0.3284026))/10000
  ```

Cela donne *pour mes valeurs* un pourcentage égal à 95.79%, qui est bien proche de 95 % ! Cela est corroboré sur le graphique du haut de la figure 4.10 page ci-contre : on a représenté les deux droites d'absxisses $p_{\min}$ et $p_{\max}$. L'aire de l'histogramme (en densité) entre ces deux valeurs représente aussi 95% de l'aire totale !

Comme dans l'annexe M page 185, vous pouvez récupérer et sourcer la fonction `verifie.int.conf.prop.R` qui simule q intervalles de confiances sur des simulation de la loi binomiale de paramètres n et p et calcule la proportion d'intervalles de confiances qui contiennent réellement p en tapant

```
verifie.int.conf.prop(p, n, NC, q)
```

Par exemple

```
verifie.int.conf.prop(0.3,1000,0.95,10000)
```

donnera une proportion égale à 95.14.

Éléments de correction de l'exercice 4.24

On s'intéresse à l'expérience aléatoire "jeter 30 fois une pièce de monnaie" où l'on note le nombre de face (qui est considéré comme un succès).

(1) Si la pièce est bien équilibrée, on a $p = 0.5$ et $n = 30$.

(2) On obtient l'intervalle de confiance suivant :

$$[0.4246955, 0.7753045],$$

qui contient la valeur de $p = 0.5$.

(3) On procède comme indiqué en cours.

On obtient donc ici 12 succès. L'intervalle de confiance à $NC = 0.95$ est

$$[0.2246955, 0.5753045],$$

qui contient *pour mes valeurs* la valeur de $p = 0.5$.

Refaisons un autre tirage.

On obtient donc ici 9 succès. L'intervalle de confiance à $NC = 0.95$ est

$$[0.1360176, 0.4639824],$$

qui ne contient *pour ces valeurs* pas la valeur de $p = 0.5$. Ici, on sera dans le cas défavorable où l'intervalle de confiance ne contient pas le paramètre inféré !

(4) Si on tape la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(1, size=30, prob=0.5)
```

on observe un tirage binomial de paramètre $n = 30$ et $p = 0.5$.

(5) *Questions facultatives*

(a) De même, la commande suivante directement dans la fenêtre de "Rgui"

```
rbinom(20, size=30, prob=0.5)
```

qui simule en fait 20 tirages binomiaux de paramètre $n = 30$ et $p = 0.5$.

(b) (i) (A)

(B) *Prévoir organisation d'un sondage sur l'ensemble des étudiants !*

(ii) Les commandes suivantes

```
res<-rbinom(25, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=25)
for(i in 1:25) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
(0.5>=IC[,1])&(IC[,2]>=0.5)
```

donnent par exemple

```
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

soit la plupart des intervalle qui contiennent la valeur du paramètre inféré et quelques autres malheureux qui ne le contiennent pas !

(iii) Les commandes suivantes

```
res<-rbinom(1000, size=30, prob=0.5)
IC<-matrix(ncol=2,nrow=1000)
for(i in 1:1000) IC[i,]<-int.conf.prop(res[i]/30,30,0.95)
100*sum((0.5>=IC[,1])&(IC[,2]>=0.5))/1000
```

donnent par exemple

```
[1] 95.1
```

c'est-à-dire une proportion proche de 0.95 !

Comme dans l'exercice 4.23 page 49, vous pouvez récupérer et sourcer la fonction `verifie.int.conf.prop.R` et taper

```
verifie.int.conf.prop(p, n, NC, q)
```

(iv) Avec $n = 100$, puis $n = 1e + 05$, on obtient successivement

```
[1] 94
```

puis

```
[1] 95.7
```

Éléments de correction de l'exercice 4.25

Si on tape

```
verifie.int.conf.prop(p, n, NC, q)
```

pour les différentes valeurs de q, on obtient successivement des proportions

$$94.34,$$
$$92.92,$$
$$9.48,$$
$$0.12.$$

Dans les deux derniers cas, les proportions ne sont pas du tout proche du NC initial. Cela provient du fait que la proportion de succès est trop proche de 0 et que l'hypothèse (4.23) n'est plus valable ; en effet, les quantités np et $n(1-p)$ valent successivement

$$\text{pour } p = 0.5 : 50 \text{ et } 50,$$
$$\text{pour } p = 0.1 : 10 \text{ et } 90,$$
$$\text{pour } p = 0.001 : 0.1 \text{ et } 99.9,$$
$$\text{pour } p = 1e - 05 : 0.001 \text{ et } 99.999.$$

On ne peut donc plus approcher la loi binomiale par la loi normale !

CHAPITRE 5

Tests d'hypothèses

EN COURS DE REDACTION ; ne sera pas traité cette année !
Pour JB :
– prévoir corrigé auto sous R et lien avec probabilité critiques.
– Voir cours M1PPMR.
– Exmple de la thèse....

5.1. Test de Normalité des données

ré-introduire qq.plot

5.2. Test sur la proportion

5.3. Test sur la moyenne

5.4. Autres test

5.5. Tests non paramètriques

CHAPITRE 6

Récapitulatifs des notions essentielles (statistique inférentielle)

Vous trouverez dans ce chapitre l'essentiel (et l'exigible aux examens !) des notions, définitions, propriétés, exercices et manipulations avec R qu'il faut savoir (ou retrouver dans le polycopié de cours) Ces notions sont présentées sous forme de listes, chapitre par chapitre, avec renvois aux points importants.

Compte tenu des modifications mineurs 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 réseaux de l'université.

Statistiques descriptives

Voir les annexes D, E, G, H, I, ainsi que l'annexe récapitulative J et l'ancien projet (révision) K.

Chapitre 3

- Définition de base : Définition 3.1 page 5 ;
- "Simulation" avec R : exercice 3.2 page 5 ;
- Définition 3.4 page 6 ;
- Définition d'une variable aléatoire : définition 3.5 page 7 et exemple 3.6 page 7 ;
- Définition d'une variable aléatoire discrète uniforme : définition 3.7 page 7 ;
- Espérance et écart-type d'une variable aléatoire : définition 3.10 page 10 ;
- Modèle binomial : définitions 3.14 page 12 et 3.15 ;
- Savoir calculer une probabilité binomiale : manipulation page 3.19 page 13 ;
- En faire les graphiques : manipulation 3.20 page 14 ;
- Applications : exercice 3.21 page 14 ;
- Espérance et écart-type binomiaux : proposition 3.23 page 14 et exercice 3.24 page 15 ;
- Probabilités cumulées : définition 3.25 page 15 et manipulation 3.26 page 15.
- Voir aussi définition 3.27 page 15 et la manipulation 3.30 page 16 et l'exercice 3.32 page 16
- Application fondamentale : exercice 3.33 page 16 ;
- Modèle normal : définitions 3.34 page 17 et 3.36 page 20 et remarque 3.37 page 20.
- Espérance et écart-type normaux : proposition 3.39 page 21 ;
- Calculs de probabilités normales : formules (3.18) et (3.22) ; Manipulation 3.40 page 22. Exercices 3.41 page 23 et 3.42 page 23. Application : exercice 3.43 page 24 ;
- Modèle de Student : section 3.5 page 24.

Chapitre 4

- Manipulation 4.1 page 33 ;
- Exercices 4.2 page 34 et 4.3 page 34 ;
- Formule (4.5) et (4.9) ;
- Les deux cas 1 page 35 et 2 page 35 ; exemple 4.4 page 35 et 4.5 page 35.
- Section 4.1.2 page 36 ;

- Section 4.2 page 37 ;
- Intervalle de confiance d'une "proposition" : proposition 4.9 page 41 et manipulations de la section 4.3.3 page 41 ;
- Applications : exemple 4.11 page 42 et exercice 4.13 page 43
- Intervalle de confiance d'une "moyenne" : proposition 4.17 page 47 et manipulations de la section 4.4.3 page 47 ;
- Applications : exercices 4.19 page 48, 4.20 page 48 et 4.21 page 48.

Chapitre 5

Non traité cette année.

ANNEXE A

Installation du logiciel R (et éventuellement du package Rcmd)

A.1. Installation de R pour Windows

Attention, le numéro de la dernière version de R est très souvent réactualisé depuis la compilation de ce document !

(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-2.15.1 for Windows (347 megabytes, 32/64 bit)" *Attention, le numéro de version de R est souvent réactualisé depuis la compilation de ce document !*

(6) Télécharger alors le logiciel d'installation `R-2.15.1-win32.exe` (ou la dernière version en date)

(7) Double-cliquer sur le logiciel `R-2.15.1-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-2.15.1 for Windows" (ou la dernière en date), vous pouvez aussi télécharger les anciennes versions de R, parfois utiles quand les plus récentes peuvent être instables ou présenter un bug non encore résolus ! Voir "Previous releases". Dans cette même rubrique, vous pouvez aussi récupérer des packages das 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 R

Utiliser le menu démarrer, puis Tous les programmes, puis R, puis R-2.15.1 (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-2.15.1`).

En quittant R, ne pas sauvegarder la session !

Ceux qui ont déjà l'habitude d'utiliser les lignes de commandes de R et de travailler sans Rcmdr peuvent naturellement continuer le faire et donc ignorer ce qui suit.

A.3. Installation et chargement du package Rcmdr

Le problème du logiciel R est qu'il s'agit d'un logiciel à langage de commandes, c'est à dire que pour l'utiliser, il faut taper des commandes dans la console, les valider pour obtenir des calculs ou des graphiques. Toutefois, il existe une version interactive employant des menus déroulants qui s'appelle Rcmdr que nous allons employer.

A.3.1. Installation de Rcmdr

Avant de commencer à utiliser ce package, il faut toutefois l'installer. La démarche suivante n'est donc à réaliser qu'une fois :

(1) L'un des menus déroulants de R s'appelle "Packages". Choisir dans ce menu l'option "Installer le(s) package(s)".

(2) Une fenêtre de dialogue s'ouvre qui vous propose un ensemble de site où vous pouvez chercher ce package. Choisir un site situé en France (Lyon devrait figurer sur la liste !).

(3) Une autre fenêtre s'ouvre qui propose une (longue) série de package, il faut choisir Rcmdr. Le téléchargement se produit alors automatiquement sur votre ordinateur.

A.3.2. Utilisation de Rcmdr

Pour utiliser le package Rcmdr (Une fois qu'il est installé...), il suffit d'aller dans le menu déroulant "Package" et de choisir l'option "Charger le package". Il faut choisir dans la liste des packages déjà installés sur votre ordinateur Rcmdr. *Attention*, à la première utilisation, R vous prévient qu'il manque des packages dont Rcmdr a besoin ; il faut répondre "OK" et laisser les champs par défaut ; un grand nombre de packages sont alors téléchargés et installés automatiquement. Pour les fois suivantes, cela ne produit plus ! Une fenêtre s'ouvre alors qui s'appelle "R commander". On utilise alors les menus déroulants pour choisir différentes options (charger un fichier, calculer des statistiques, réaliser des graphiques, ...). Il faut noter que le résultat des actions s'inscrit dans la fenêtre du bas du R commander qui s'appelle "Fenêtre de sortie". En revanche, les graphiques apparaissent dans la fenêtre habituelle de R dite "RGui". Il faut donc jongler entre "RGui" et "Rcommander" (on sy habitue).

Enfin on peut noter que lorsqu'une action est choisie par un menu déroulant, des lignes s'inscrivent dans la fenêtre dite "Fenêtre de script". Il s'agit des commandes réelles que R exécute (et dont on voit le résultat dans la fenêtre de sortie). Il se peut que dans certains cas (très rares), les menus déroulants ne soient pas suffisants, nous entrerons alors directement des commandes soit dans cette fenêtre de script soit dans la fenêtre de "Rgui" pour les exécuter.

On pourra consulter la doc en pdf `Getting-Started-with-the-Rcmdr.pdf` disponible normalement dans votre ordinateur (si vous y avez installé R), à l'adresse habituelle du site de ce cours ou dans l'ordinateur de l'université (dans le répertoire où R est installé, en général `C:\Program Files`) dans le répertoire `\R\R-2.15.1\library\Rcmdr\doc`

ANNEXE B

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 logociels 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 polycopié 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 polycopié 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 '`M2IGAPAS`'.

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 polycopié 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 polycopié 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, dézipez éventuellement (clic droit, "extraire ici") les fichiers de données.

B.3. Démarrage du logiciel R (et éventuellement du package Rcmd)

(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.

(3) Pour ceux qui utilisent Rcmdr, il faut charger l'interface interactive de R en utilisant le menu déroulant "Packages" puis l'option "Charger le package", choisir alors le package Rcmdr (une interface graphique doit alors s'ouvrir).

ANNEXE C

Une toute petite introduction à la statistique descriptive (sans R)

C.1. Introduction

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

C.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 C.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 C.1 ; elles ont été collectée 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 l'url suivante `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 [1]. 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 phénomèmes multivariés (en fait, seul le cas de deux variables sera étudié).

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

C.3. Étude de donnée qualitatives

On s'intéressera au sexe des $N = 8$ étudiants de M1APA (voir le tableau C.1).

C.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 C.1. Déterminer les valeurs de ces statistiques pour l'échantillon des huit étudiants étudié.

Voir éléments de correction page 76.

C.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 C.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 76.

Exercice C.3. On s'intéresse maintenant au sport pratiqué par les huit étudiants déjà étudiés. Il faudra prende 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" ?

C.4. Étude de données quantitatives

On s'intéressera au poids des $N = 8$ étudiants de M1APA (voir le tableau C.1 page précédente).

C.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, ..., n_8. De façon générale, ils seront notés $(n_i)_{1\leq i\leq N}$.

On a donc successivement

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

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

C.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*[2].

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

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

Cela revient à donner la définition suivante :

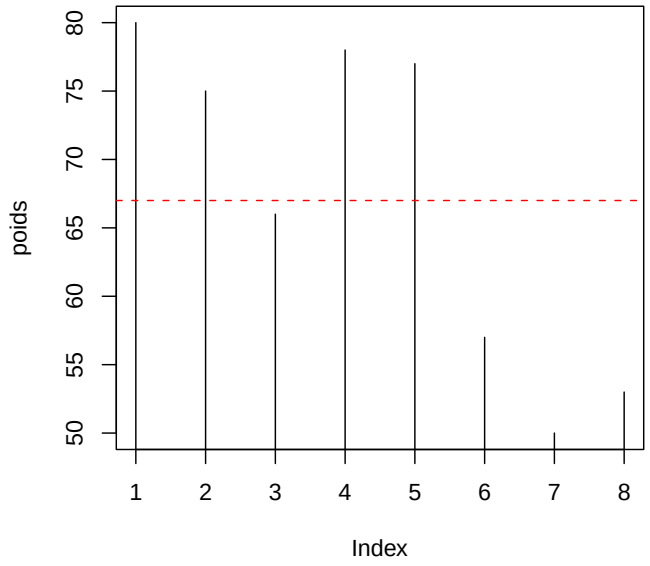
DÉFINITION C.4. La moyenne m des N nombres $(n_i)_{1\leq i\leq N}$ est définie par

$$m = \frac{1}{N}(n_1 + n_2 + \ldots + n_N) = \frac{1}{N}\sum_{i=1}^{N} n_i \tag{C.1}$$

EXEMPLE C.5. La moyenne m des $N = 8$ poids du tableau C.1 page 69 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 extrèmes et correspond au type le plus répandu [**11**]



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

Une notions moins connue est la *la médiane*[3]. La médiane est un nombre qui divise en deux parties la population. On la note Q_2.

Définition C.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 extrèmes que la moyenne.

Exemple C.7. La médiane Q_2 des $N = 8$ poids du tableau C.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 C.8. Si on avait voulu calculer la médiane des $N - 1 = 7$ premiers poids du tableau C.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" [**11**]

C.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 $\min\left(n_{i1\leq i\leq N}\right)$ et $\max\left(n_{i1\leq i\leq N}\right)$.

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

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

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

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

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

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

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

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

Définition C.9. La *variance*, notée σ^2, des N nombres $(n_i)_{1\leq i\leq N}$ est définie par

$$\sigma^2=\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)=\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2 \tag{C.2}$$

Définition C.10. L'*écart-type*, noté σ, des N nombres $(n_i)_{1\leq i\leq N}$ est défini par

$$\sigma=\sqrt{\frac{1}{N}\left((m-n_1)^2+(m-n_2)^2+\ldots+(m-n_N)^2\right)}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(m-n_i)^2} \tag{C.3}$$

Remarque C.11. Attention, on parle quelque fois de l'écart-type estimé :

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

On note aussi l'écart-type par son nom anglais, *sd*, comme standart deviation. Attention, ®détermine la déviation standart et non l'écart-type !

Exemple C.12. La variance σ^2 des $N=8$ poids du tableau C.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

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

EXEMPLE C.13. L'écart-type σ des $N = 8$ poids du tableau C.1 se calcule de la façon suivante. Le détail du calcul est le suivant :

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

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

DÉFINITION C.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 R, le logiciel que vous utiliserez par la suite ! ♢

EXEMPLE C.15. Pour calculer, les trois quartiles des $N = 8$ poids du tableau C.1 :

- On calcule la médiane comme dans l'example C.7 page 72 : $Q_2 = 70.5$.
- 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 :

$$Q_1 = 56,$$
$$Q_2 = 70.5,$$
$$Q_3 = 77.25.$$

♢

C.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 (σ). 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 reflétent totalement les données considérés et les caractérisent donc.

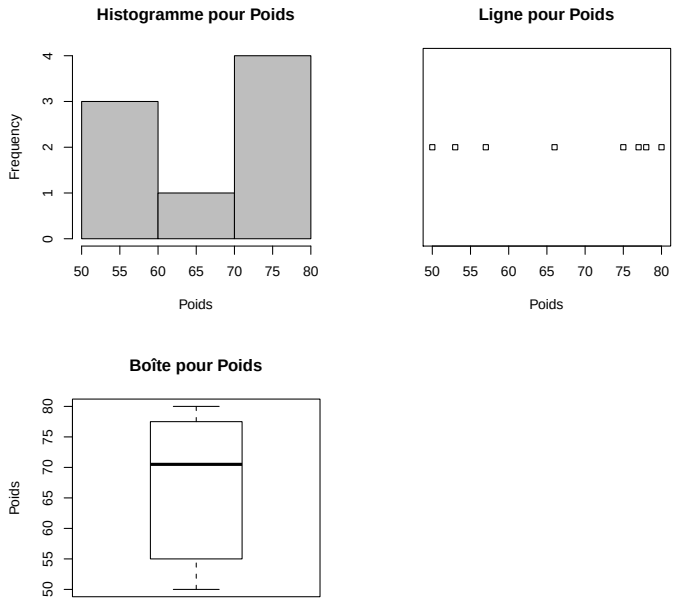
Remarque C.16. En anticipant sur le chapitre 4, on peut monter que si les données suivent une loi normale, alors l'intervalle $[m - \sigma, m + \sigma]$ contient 68.268949% des données.

C.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 boite centrale avec des traits noirs d'ordonnées Q_1, Q_2 et Q_3, en étirant les "moustaches" jusq'aux valeurs minimale et maximale.

On appelle EIQ l'écart nterquartile $EIQ = Q_3 - Q_1$; 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 extrèmes 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 C.17. 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 77.

C.5. Élements de correction

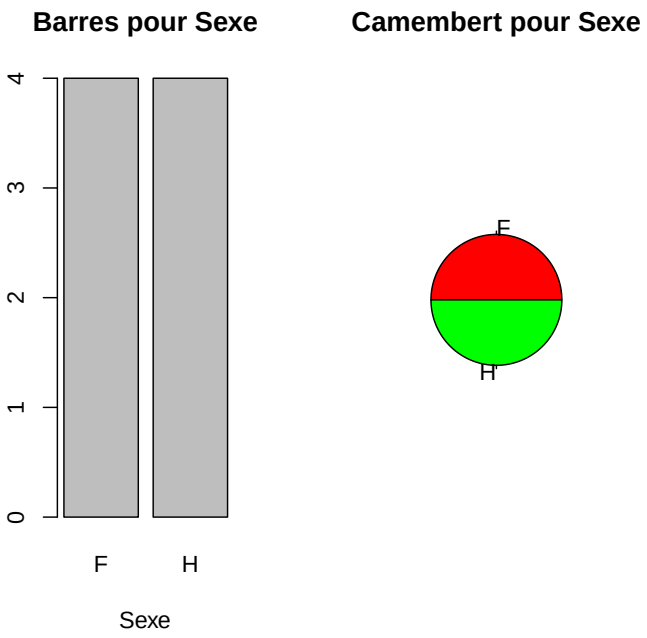
Éléments de correction de l'exercice C.1

Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
F	4	50.000
H	4	50.000

Éléments de correction de l'exercice C.2

(1)

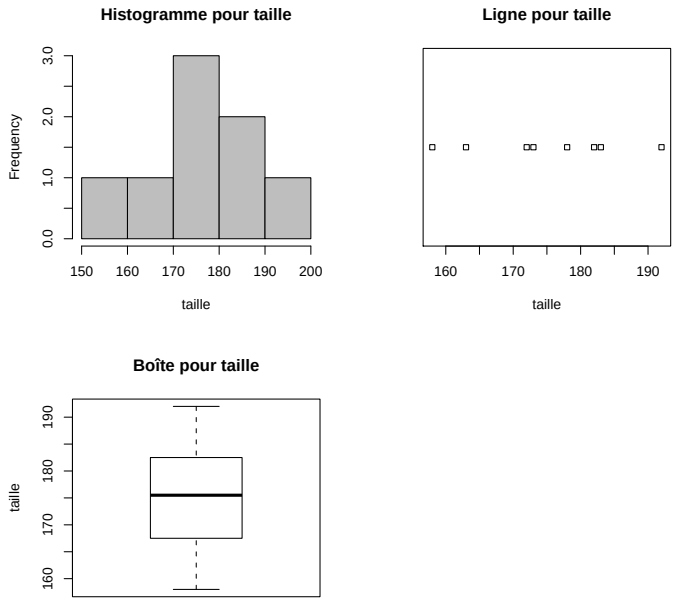


Voir les deux graphiques ci-dessus pour la variable sexe.

(2) À vous de voir ....

Éléments de correction de l'exercice C.17
Les différents résultats déterminés par 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.

ANNEXE D

Données catégorielles

Cette annexe s'inspire fortement du chapitre 3 de [**5**].

Cette annexe sera traitée rapidement comme révision.

Ceux qui n'auront jamais fait de statistiques pourront lire les sections C.1, C.2 et C.3 de l'annexe C, qui ne sera pas traitée en cours.

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.

D.1. La situation concrète : récupération d'un fichier de données

Le jeu de données `DIARRHF.txt` contient les réponses de 38 femmes marathoniennes ayant participé au marathon de Belfast 1986. La question était "Avez-vous lors de marathon déjà éprouvé des problèmes de diarrhées ?". La réponse ne peut être que OUI ou NON et est donc de nature catégorielle. Pour obtenir ce fichier texte, il faut

- Aller sur le site `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html` puis rubrique M2APA ;
- Faire un clic droit à la souris et enregistrer le fichier (ici `DIARRHF.txt`) sur son ordinateur.
- Il est alors possible d'ouvrir avec Wordpad (ou tout autre éditeur de texte) ce fichier. Noter que dans le fichier `DIARRHF.txt` les unités d'observations sont en lignes et les mesures (une seule en l'espèce) en colonnes.

D.2. Importer le jeu de données dans R

L'utilisation de R peut se faire de deux façons :

- Pour ceux qui ont déjà pratiqué R, on utilise la fenêtre de commande de R (Rgui) dans laquelle on rentre directement les lignes de commandes. Cette utilisation est plus difficile, mais, à mon goût plus souple et surtout permet d'automatiser un certain nombre de tâches.
- Pour ceux qui n'ont jamais pratiqué R, on utilise Rcmdr, appelé aussi RCommander (voir section A.3 page 66). Cette utilisation est plus facile, mais, à mon goût moins souple. Il existe des cas où l'on est obligé d'utiliser selon le premier mode.

Naturellement, chacun est libre de faire comme il lui convient !

D.2.1. Avec Rcmdr

Afin d'importer ce ficher dans le logiciel R, et une fois que R et Rcmdr sont ouverts (voir annexe B), il faut suivre les étapes suivantes :

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis un fichier texte ou le presse-papier...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier texte sans extension, c'est-à-dire : '`DIARRHF`'. Laisser les autres champs avec les valeurs choisies par défaut.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier à importer ('`DIARRHF.txt`').

(3) Cliquer alors éventuellement sur le bouton "Visualiser".

Après cette procédure, le jeu de données est actif pour le logiciel R.

REMARQUE D.1. *Attention*, parfois (notamment si vous travaillez sur de vieux sujets d'examens) vous aurez à ouvrir des fichiers xls, format abandonné à cause de problèmes trop nombreux de compatibilité entre R et excel. Néanmoins, la manipulation à faire est presque identique à la précédente :

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis Excel, Access ou dBase...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier xls sans extension.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier xls.

Après cette procédure, le jeu de données est actif pour le logiciel R.

D.2.2. Directement avec Rgui

Il faut d'abord lire le fichier de nom '`DIARRHF.txt`' et l'affecter dans une variable dont le nom sera `DIARRHF`' ; on tapera donc la commande suivante directement dans la fenêtre "Rgui" :

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

ou encore, ce qui est équivalent

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

Voir ensuite ce que donne les commandes

```
DIARRHF
```

ou

```
head(DIARRHF)
```

REMARQUE D.2.

(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 (↑ et ↓) et les modifier en vous aidant des fléches "gauche" et "droite" du clavier (← et →)

(3) Dans ce polycopié,
- Conformément à la présentation de 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 R, les "sorties" (ce qui sera calculé par 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 sous présentés simultanément, vous verrez
```
 cos(2)
[1] -0.4161468
```

REMARQUE D.3. *Attention*, parfois (notamment si vous travaillez sur de vieux sujets d'examens) vous aurez à ouvrir des fichiers xls, format abandonné à cause de problèmes trop nombreux de compatibilité entre R et excel. Il faut charger le package `xlsReadWrite` et donc taper par exemple

```
library(xlsReadWrite)
```

Vous n'avez ensuire plus qu'à taper

```
nom <- -read.xls("nom.xls")
```

où '`nom.xls`' est le nom du fichier xls.

Attention, la commande `library(xlsReadWrite)` charge le package `xlsReadWrite`. Si celui-ci n'est pas installé :

- si vous êtes sur votre propre ordinateur, aller dans le menu déroulant de Rgui "package", puis "installer les packages" et choisir le package `xlsReadWrite`. Cela nécessite une connexion internet Voir aussi la remarque A.2 page 65)
- si vous êtes en salle de TP à Lyon 1, déconnectez-vous de votre session et demandez-moi de l'installer : le CRI a protégé le répertoire de `C:\R` à l'écriture mais j'ai les droit nécessaires.

D.3. Dénombrer les catégories

L' effectif d'une catégorie est le nombre de réponses obtenues correspondant à cette catégorie. On appellera cette opération un dénombrement. On note généralement n l'effectif total.

- *Avec Rcmdr :*

  Afin de calculer les effectifs de chaque catégorie, il faut aller dans le menu déroulant "Statistiques" du Rcommander et choisir les options "Résumés" et "Distributions de fréquences". On obtient dans la fenêtre de sortie le résultat ci-dessous :

  ```
  Non Oui
   12  26
  ```

- *Sans Rcmdr :*

  ```
  summary(diar$PbDigest)
  ```

  ce qui affiche

  ```
  Non Oui
   12  26
  ```

Ainsi, sur 38 femmes, 26 ont déclaré avoir des problèmes digestifs de type diarrhées, ce qui semble énorme.

On peut aussi déterminer les pourcentage en divisant les effectifs par l'effectif total et en multiplant par 100.

Remarque D.4. Pour ceux qui n'utilisent pas Rcmdr, on pourra s'affranchir de taper le nom de la variable contenant les données (ici `diar`) en utilisant un attachement

```
attach(diar)
```

Au lieu de taper

```
summary(diar$PbDigest)
```

ou se contentera de

```
summary(diar)
```

ce qui affiche aussi

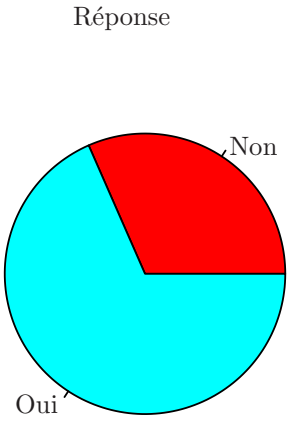
```
Non Oui
 12  26
```

Ne pas oublier de détacher en fin de travail :

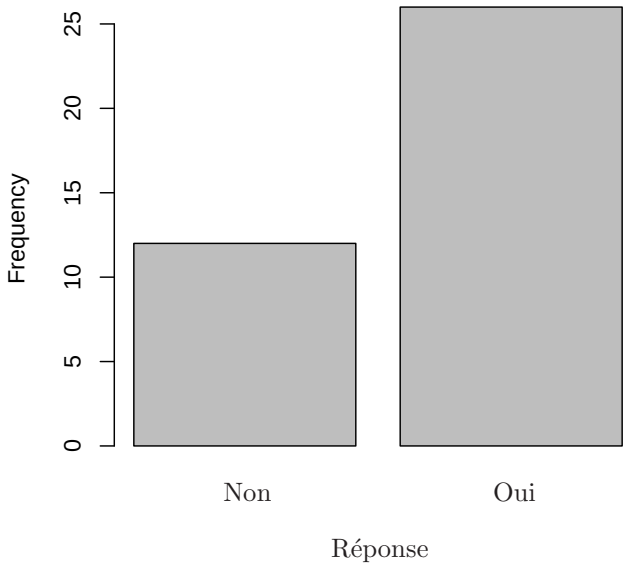
```
detach(diar)
```

D.4. Les graphiques pour les catégories

Il est inutile de produire des représentations graphiques de données avec deux catégories. Nous allons toutefois le faire afin de préparer la voie à des situations plus complexes où le nombre de catégories sera plus élevé.



(a) : camembert



(b) : le graphique en barre

FIGURE D.1. Deux graphiques pour les données '`DIARRHF.txt`'.

Dans le *camembert*, chaque catégorie est représentée par un secteur sur un disque de façon proportionnelle à son effectif.

- *Avec Rcmdr :*

  Pour réaliser un camembert, aller dans le menu déroulant "Graphes" du **RCommander** et choisir l'option "Graphe en camembert". Le graphique est présenté dans la figure du haut de la figure D.1 page précédente.

  Dans le *graphique en barres*, la hauteur de chaque barre montre l'effectif correspondant à la catégorie indiquée à la base de la barre. Pour le réaliser, aller dans le menu déroulant "Graphes" du **RCommander** et choisir l'option "Graphe en barres". Le graphique est présenté dans figure du bas de la figure D.1 page ci-contre.
- *Sans Rcmdr :*

  On tapera :

```
pie(table(diar$PbDigest))
barplot(table(diar$PbDigest))
```

D.5. Extension à l'étude de plus de deux catégories

Nous avons jusqu'à présent étudié des cas ou seules deux catégories existaient. Le fichier `FFHandi.txt` contient la nature du handicap des licenciés (1991) de la Fédération Française Handisport. Nous allons voir que l'analyse se généralise sans changement à plus de deux catégories.

EXERCICE D.5. Utiliser les méthodes précédentes pour décrire la répartition des licenciés dans ces catégories. Que signifie la catégorie "Autres"? N'y a-t-il pas une catégorie dont la présence est surprenante? Quel peut être l'intérêt de connaître également la fréquence de ces différents handicaps dans la population française?

Voyons à présent dans le cadre d'une étude des pratiques sportives des camerounais (de 15 à 75 ans) la généralisation à un nombre plus important de catégories : le sport pratiqué par les sondés. Le jeu de données `cameroun1.txt` comprend les résultats d'un sondage effectué sur 395 individus.

EXERCICE D.6. Analyser la variable APS qui décrit l'activité sportive déclarée par le sondé. En particulier réaliser le graphe en barres et le graphe en camembert. Quel est le nouveau problème qui se pose ici?

Dans cet exercice, le grand nombre de catégories rend difficile l'analyse. Les solutions qui s'offrent à nous sont soit de regrouper les catégories en sous-ensemble qui ont du sens (sport de combat, sport collectif...) soit de créer une catégorie "Autre" qui regroupe les catégories ayant les plus petits effectifs. Afin de visualiser l'ensemble, on peut aussi utiliser un graphe particulier, le graphe en points, qui n'est malheureusement ps disponible dans la version interactive de R.

REMARQUE D.7. (Remarque inutile pour les non-néophytes) Il est possible d'employer R dans une version langage de commandes qui est beaucoup plus puissante. On tape alors des commandes dans la fenêtre de script puis on clique sur le bouton "Soumettre". On peut aussi taper des commandes dans la console (fenêtre "Rgui") et appuyer sur la touche "Enter".

Le résultat s'affiche soit dans la fenêtre de sortie soit dans une fenêtre graphique.

EXERCICE D.8. Bien relire la remarque D.2 page 80, puis soumettez successivement les 4 ordres suivants (attention, ici `cameroun1` est éventuellement à remplacer par le nom de la variable que vous avez choisi lors de l'importation des données, `Dataset` par défaut)

(1) `cameroun1$APS`

(2) `table(cameroun1$APS)`

(3) `dotchart(table(cameroun1$APS))`

(4) `dotchart(sort(table(cameroun1$APS)))`

Que font ces lignes de commandes ? Quelles sont les catégories les plus importantes ? Quelles catégories suggérez-vous de regrouper dans une catégorie "Autres" ?

D.6. Les données ordinales

Le fichier '`sauna.txt`' comprend les réponses de 687 sondés à une enquête de satisfaction concernant les piscines lyonnaises d'hiver. La question portait ici sur l'implantation d'un sauna, était-elle : Trés souhaitée, Souhaité, Indifférent, Pas souhaité, Pas du tout souhaitée ?

Exercice D.9. Analyser ces souhaits en utilisant les méthodes précédemment décrites. Quelle est la nouveauté concernant ces catégories et en quoi les graphes ne sont pas très bons ? Comment regrouper habilement de telles catégories ?

Remarque D.10. Pour modifier l'ordre de catégories, il faut

- Avec Rcmdr, utiliser le menu déroulant "Données", l'option "Gérer les variables dans le jeu de données actif", puis "Réordonner une variable facteur" (en choisir une nouvelle de nom `Implantation_ordonnee`) en choisissant l'ordre :
  - Indifférent $\longrightarrow$ 3
  - Pas du tout souhaitée $\longrightarrow$ 1
  - Pas souhaitée$\longrightarrow$ 2
  - Souhaitée $\longrightarrow$ 4
  - Très souhaitée$\longrightarrow$ 5
- sans Rcmdr, introduire la nouvelle variable `Implantation_ordonnee` en tapant

```
sauna$Implantation_ordonnee <- factor(sauna$Implantation,
levels = c("Pas du tout souhaitée", "Pas souhaitée", "Indifférent",
 "Souhaitée", "Très souhaitée"))
```

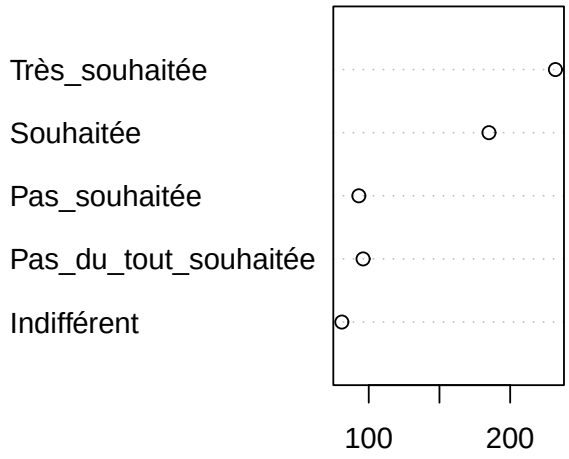
On peut alors, dans les deux cas, vérifier l'odre des niveaux choisi en tapant dans la fenêtre de commande :

```
levels(sauna$Implantation)
levels(sauna$Implantation_ordonnee)
```

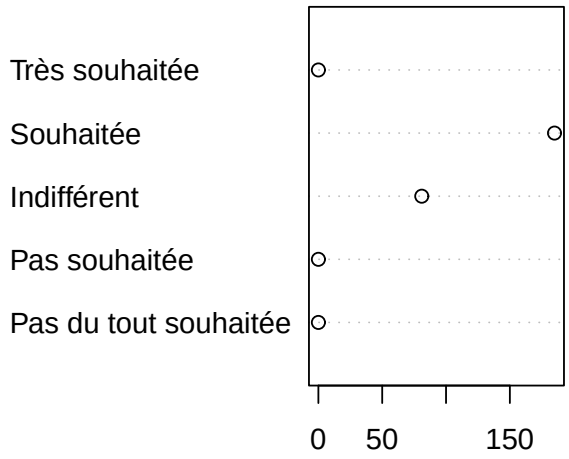
et tracer la ligne de points en tapant la commande :

```
dotchart(table(sauna$Implantation_ordonnee))
```

Voir les deux graphes produit en figure D.2.



(a) : graphe ”non ordonné”



(b) : graphe ”ordonné”

FIGURE D.2. Deux graphe en points pour les données ’`SAUNA.txt`’.

ANNEXE E

Données numériques

Cette annexe s'inspire fortement du chapitre 4 de [**5**].

Cette annexe sera traitée rapidement comme révision.

Ceux qui n'auront jamais fait de statistiques pourront lire les sections C.1, C.2 et C.4 de l'annexe C, qui ne sera pas traitée en cours.

E.1. La situation concrète

Comme dans les sections D.1 et D.2 page 79, récupérer avec ou sans Rcmdr le fichier de données '`STUDENT_H.txt`'.

On s'interessera aux deux données '`Taille`' et '`Age`'.

E.2. Les graphiques pour données numériques

DÉFINITION E.1. Dans l'*histogramme*, l'échelle de mesure est découpée en tranches (souvent d'égales longueurs), le nombre d'unité dans chaque tranche est dénombré et un rectangle prenant pour base la tranche et pour hauteur l'effectif permet alors de représenter où sont concentrées les données.

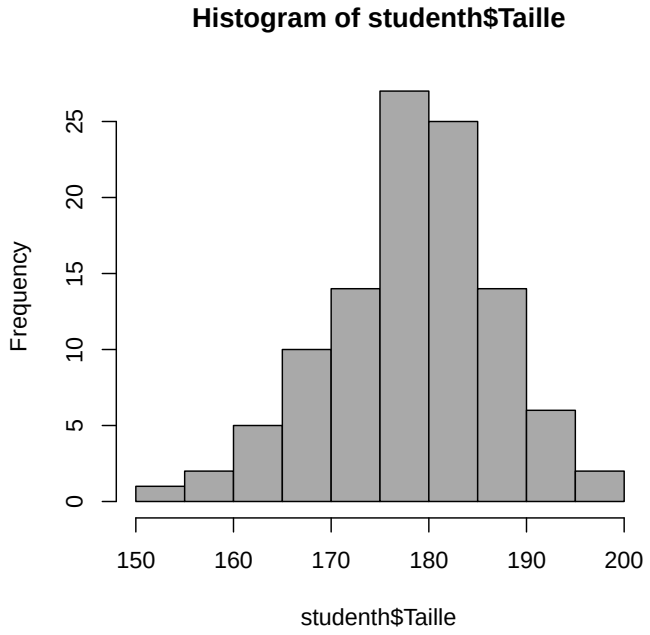


FIGURE E.1. L'histogramme de la variable taille chez les étudiants (données '`STUDENTH.txt`')

Pour réaliser un histogramme,

- *Avec Rcmdr :*
  il faut utiliser le menu déroulant "Graphes" et l'option "Histogramme". On obtient la figure E.1 page précédente.
- *Sans Rcmdr :*
  il faut taper

```
hist(studenth$Taille)
```

Sur cet histogramme, il apparaît clairement que les tranches les plus importantes sont celles allant de 175 à 180 et de 180 à 185.

Le choix des tranches est assez complexe, il est fait automatiquement par le logiciel R mais diverses considérations peuvent conduire à choisir son propre système de tranches [1].

Dans ce graphique, on va chercher à repérer comment sont organisées les données et en particulier (1) s'il existe une valeur typique, c'est-à-dire si les données sont concentrées autour d'une valeur centrale et (2) si les données sont homogènes (proches de la valeur centrale) ou hétérogènes, c'est l'étude de la variabilité

Il existe aussi un autre graphique, appelé boîte de dispersion que nous verrons plus tard dans ce chapitre.

EXERCICE E.2. Le jeu de données `NOUVEL_OBS.txt` contient trois mesures décrivant les numéros récents de l'hebdomadaire *Le Nouvel Observateur* [2] : le numéro concerné `Numéro`, le nombre de pages `Pages` et le nombre de pages de publicité `Publicité`. Importer ce fichier sous le nom `NO`.

(1) Étudier graphiquement cette dernière mesure. Quel peut être le problème concernant cet histogramme ?

(2) Nous allons réaliser une ligne de points. C'est un graphique qui n'a malheureusement pas disponible directement dans le Rcmdr. Taper dans la fenêtre de script la commande :

```
stripchart(NO$Publicité, method = "stack")
```

Voir la ligne de points en figure E.2 page ci-contre.

(3) Calculer une nouvelle variable qui donne le pourcentage de pages de publicité

- *Avec Rcmdr :*
  il faut utiliser le menu déroulant "Données", puis "Gérer les variables dans le jeu de données actifs" puis "Calculer une nouvelle variable".
- *Sans Rcmdr :*
  taper

```
pourcentage <- NO$Publicité/NO$Pages * 100
```

Analyser cette nouvelle variable.

E.3. Les indicateurs statistiques pour les données numériques

E.3.1. La centralité ou l'unité statistique typique

Après avoir repéré visuellement la valeur typique, on calcule des statistiques permettant de quantifier précisément et objectivement la notion de centralité. La *moyenne* est la plus connue.

Notons les données sous la forme y_1, y_2, ..., y_N, soit encore $(n_i)_{1\leq i\leq N}$.

Dans cette section, nous travaillerons sur les données du fichier '`STUDENT_H.txt`'. On étudiera cette fois-ci les âges.

E.3.1.1. La moyenne.

1. une option le permet en partie dans la procédure, il faut indiquer le nombre de classes que l'on souhaite employer.
2. Il n'y a pas de numéros d'été dans ces exemplaires car ils sont assez différents et contiennent moins de pages. Seuls les exemplaires d'au moins 100 pages ont été retenus.

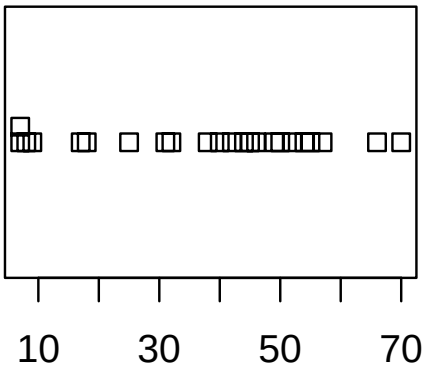


FIGURE E.2. la ligne de points de la variable Publicité (données 'NOUVELOBS.txt').

DÉFINITION E.3. La moyenne est la somme des valeurs divisée par leur nombre. On la notera $\overline{y}$ ou m. On a donc

$$\overline{y} = \frac{1}{N}(n_1 + n_2 + \ldots + n_N) = \frac{1}{N}\sum_{i=1}^{N} n_i$$

Pour calculer différentes statistiques,

- *Avec Rcmdr :*
  on utilise le menu déroulant "Statistiques", puis les options "Résumés" et "Statistiques descriptives".
  On obtient dans la fenêtre de sortie le résultat ci-dessous :

```
     mean       sd    0%    25%    50%     75%   100%
 20.33196 6.069863 16.75 17.917 18.875 20.2915 70.417
```

- *Sans Rcmdr :*

```
summary(studenth$Age)
```

  ou

```
mean(na.omit(studenth$Age))
```

  On pourra aussi regarder ce que donne

```
mean(studenth$Age)
```

On mesure donc une moyenne de 20.33196.

La moyenne est un centre de gravité, ce qui signifie, comme nous le verrons, qu'elle peut être sensible à certaines valeurs particulièrement éloignées. Une autre statistique de centralité, la *médiane*, ne présente pas cet inconvénient.

E.3.1.2. La médiane.

DÉFINITION E.4. La médiane est une valeur choisie pour que la moitié des données lui soit inférieure et l'autre moitié supérieure. On la notera Q_2.

Pour l'obtenir sans Rcmdr, on pourra taper

```
median(na.omit(studenth$Age))
```

ou

```
median(studenth$Age, na.rm = T)
```

La médiane 18.875 est, en l'espèce, très proche de la moyenne, ce n'est pas toujours le cas comme nous le verrons sur d'autres exemples. Avec la médiane, la plus grande des valeurs peut bien doubler, cela ne modifiera pas le résultat. En effet, c'est seulement la place de la valeur qui compte. Si les graphiques montrent clairement l'existence de données extrêmes, on mentionnera systématiquement la valeur de la médiane (avant la moyenne).

EXERCICE E.5. Le fichier '`MONDE84.txt`' contient pour 48 pays cinq mesures socio-démographiques. Nous étudierons seulement la variable pib (Produit Intérieur Brut).

(1) Représenter graphiquement la variable pib et commenter le résultat obtenu

(2) Calculer les deux statistiques de centralité et commenter le résultat. A votre avis, dans quelles conditions cela peut-il se reproduire ?

EXERCICE E.6. Le fichier `EMPREINTE.txt` contient pour 16 pays d'Europe de l'Ouest les valeurs nécessaires au calcul de l'empreinte écologique de ces pays. L'empreinte écologique d'un pays est (une estimation de) la surface totale requise pour produire la nourriture et les fibres qu'il consomme, pour répondre à sa consommation d'énergie, et pour fournir l'espace nécessaire à son infrastructure. La surface mondiale disponible par personne est de 2.3 hectares.

(1) Représenter graphiquement la variable '`total`' qui est l'empreinte écologique totale par personne et commenter le résultat obtenu.

(2) Calculer les deux statistiques de centralité et commenter le résultat.

E.3.2. La dispersion ou l'hétérogénéité des unités statistiques

E.3.2.1. Les valeurs minimum et maximum.

Si la notion de centralité, sous la forme de la moyenne, est devenue "naturelle" pour beaucoup de personnes afin de résumer un jeu de données, elle peut aussi être fortement réductrice. *Au delà de l'individu typique, il y a diversité*. Cette diversité, cette hétérogénéité, qu'on appelle *dispersion* ou *variabilité* en statistique, il faut prendre conscience de son existence, de son importance et il faut la quantifier. La façon la plus simple de le faire est de donner les valeurs *minimum* et *maximum*.

- *Avec Rcmdr :*
  Ces valeurs sont fournies comme précédement.
- *Sans Rcmdr :*
  On pourra taper
  ```
  min(studenth$Age, na.rm = T)
  ```
  et
  ```
  max(studenth$Age, na.rm = T)
  ```
  ou directement
  ```
  range(studenth$Age, na.rm = T)
  ```

Pour les âges des étudiants, la valeur minimum est de 16.75 et la valeur maximum de 70.417. Ces statistiques sont très utiles afin de vérifier qu'il n'y a pas d'erreur dans les mesures [3], et donnent une idée simple de l'hétérogénéité d'un jeu de données. Cependant, elles ont le défaut d'être extrêmement sensibles à l'existence de valeurs extrêmes puisque ce sont justement les valeurs extrêmes !

3. Si on avait obtenu par exemple une valeur inférieure à 10 ou supérieure à 100

E.3.2.2. L'écart-type.

En fait, la statistique la plus utilisée pour décrire la dispersion est l'*écart-type*. L'écart-type vise à mesurer une sorte d'écart moyen entre les valeurs du jeu de données et leur moyenne. *Il mesure donc la façon typique dont l'échantillon s'écarte de l'unité typique.*

Définition E.7. L'écart entre les données et la moyenne, une fois mis au carré, constitue une valeur positive qu'on peut considérer comme une distance. La moyenne de ces distances constitue la variance, qu'on notera Var.

L'écart-type, qui est la racine carrée de la variance, permet de retrouver les unités originales de mesure. On le notera SD.

Définition E.8. Si les valeurs observées sont notées : $y_1, \cdots, y_N$ et si la moyenne est notée $\overline{y}$ la formule employée pour l'écart-type est soit :

$$\sigma = \sqrt{\frac{1}{N}\left((y_1 - \overline{y})^2 + \ldots + (y_N - \overline{y})^2\right)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2},$$

soit plus souvent (et c'est celle qu'utilise R)

$$SD = \sqrt{\frac{1}{N-1}\left((y_1 - \overline{y})^2 + \ldots + (y_N - \overline{y})^2\right)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

En ce qui concerne le jeu de données des âges des étudiants, l'écart-type est $SD = 6.069863$. Pour les tailles, il est de $SD = 8.380252$. Très souvent, pour résumer un jeu de données, on donnera sa moyenne et son écart-type sous la forme $\overline{y} \pm SD$. On veut dire par là que les données les plus classiques sont dans cet intervalle.

Si l'écart-type offre généralement un résumé très efficace de l'hétérogénéité d'un jeu de données, il n'est valide qu'à deux conditions : (1) qu'il n'y ait pas de données extrêmes car l'écart-type est, encore plus que la moyenne, sensible à de tels éléments et (2) que les valeurs soient plus ou moins symétriques autour de la moyenne.

Nous pouvons observer que, dans l'histogramme des tailles (figure E.1 page 87), ces deux conditions étaient ici relativement bien respectées. Dans ce cas, le résumé est donc satisfaisant. En revanche, cela n'est pas valable pour l'histogramme des âges (figure E.3 page suivante). Puisque les deux conditions ci-dessus ne sont pas respectées.

E.3.2.3. Les quartiles.

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

Définition E.9. 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% de 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).

Les quartiles permettent en particulier de voir si les données "s'écartent plus d'un côté que de l'autre par rapport au centre".

Étudions la variable ES (estime de soi) du fichier '`ALIMENTATION.txt`'.

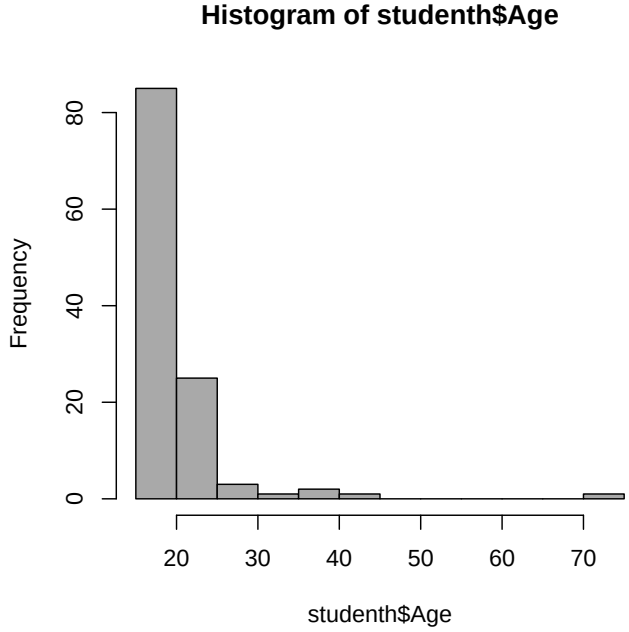


FIGURE E.3. L'histogramme de la variable âges chez les étudiants (données 'STUDENTH.txt').

On a

$$\begin{aligned} Q_1 &= 26, \\ Q_2 &= 29, \\ Q_3 &= 32, \\ Q_3 - Q_2 &= 3, \\ Q_2 - Q_1 &= 3 \end{aligned}$$

La distribution ici est *symétrique* car la distance de Q_1 à Q_2 (3) est égale à celle de Q_2 à Q_3 (3).

Dans cet exemple, on peut mesurer la dispersion avec une seule statistique, l'écart-type, qui est un compromis entre les écarts au centre sur la gauche et ceux sur la droite car ils sont comparables. Dans d'autres cas, lorsqu'il y a forte dissymétrie, ce n'est pas possible. Les quartiles, au prix d'un peu plus de complexité, reflètent alors mieux la situation.

En revanche, pour la distribution des âges du fichier 'STUDENT_H.txt', on a

$$\begin{aligned} Q_1 &= 172.8, \\ Q_2 &= 180, \\ Q_3 &= 185, \\ Q_3 - Q_2 &= 5, \\ Q_2 - Q_1 &= 7.2 \end{aligned}$$

La distribution ici n'est pas *symétrique* car la distance de Q_1 à Q_2 (5) est différente de celle de Q_2 à Q_3 (7.2).

E.3.2.4. La boîte de dispersion.

DÉFINITION E.10. Il existe un graphique très intéressant pour appréhender visuellement ce problème de symétrie : la *boîte de dispersion*. 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 [4], elles apparaissent comme des points que les moustaches ne rejoignent pas.

Pour obtenir ce dessin,

- *Avec Rcmdr :*
  utiliser le menu déroulant "Graphes", puis "Boîte de dispersion".
- *Sans Rcmdr :*
  pour réaliser par exemple la boîte de dispersion de ES du fichier 'ALIMENTATION.txt', on tape

```
boxplot(alimentation$ES, ylab = "ES")
```

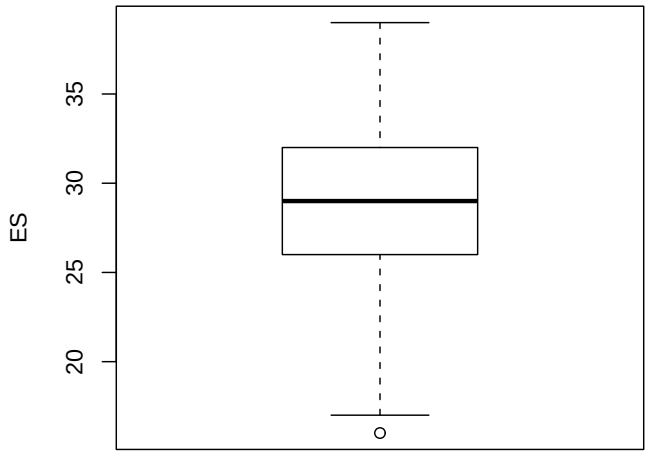


FIGURE E.4. La boîte à moustache de l'histogramme de la variable ES (données 'ALIMENTATION.txt').

Voir la figure E.4.

Dans l'exemple de l'estime de soi (figure E.4), la boîte à moustaches n'indique aucune dissymétrie. Le trait gras est situé au centre dans la boîte. Les deux moustaches sont de tailles équivalentes et il y a, à peine, une donnée extrême.

En revanche, la distribution des âges du fichier 'STUDENT_H.txt', n'est plus symétrique (figure E.5 page suivante), ce qui illustre l'approche par les quartiles.

EXERCICE E.11. Réaliser une boîte à moustaches avec les valeurs de pib du fichier 'MONDE84.txt' afin de bien voir ressortir les données extrêmes.

Les quartiles ont un autre avantage : ils sont peu sensibles aux valeurs extrêmes. C'est pourquoi, il est possible de construire sur leur base un *écart-type modifié* qui donne des résultats similaires à l'écart-type

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

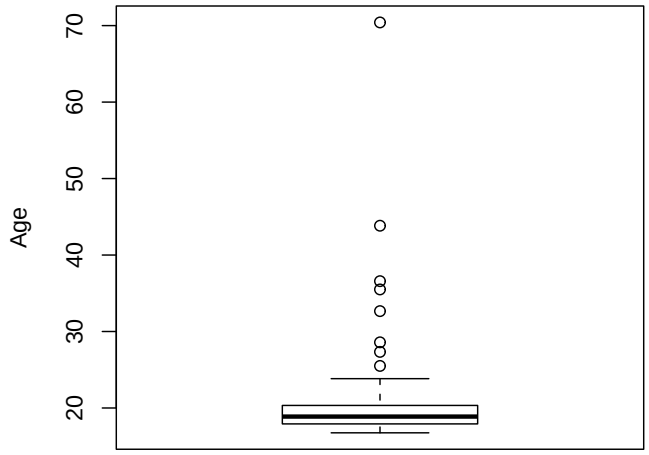


FIGURE E.5. La boîte à moustache de l'histogramme de la variable Age des étudiants.

classique sur des données "bien configurées"[5] et qui écarte sinon les valeurs extrêmes. Cet écart-type modifié est égal à la différence entre les quartiles Q_1 et Q_3 divisé par 1.369 :

$$\widetilde{\sigma} = \frac{Q_3 - Q_1}{1.369}.$$

En revanche, cet écart-type, même modifié, ne saurait refléter la dissymétrie ; seule la donnée des deux quartiles le permet. Il est dans l'exemple de l'estime de soi égal à

$$\widetilde{\sigma} = \frac{Q_3 - Q_1}{1.369} = \frac{32 - 26}{1.369} = 4.382761$$

donc relativement proche de l'écart-type habituel ici à

$$SD = 4.698949$$

Pour réaliser ce calcul,

- *Avec Rcmdr :*
  taper dans la fenêtre de script : `(32-26)/1.369` et cliquer sur le bouton "Soumettre"
- *Sans Rcmdr :*
  taper

```
d <- quantile(alimentation$ES)
```

  puis

```
(d[4] - d[2])/1.369
```

  ou mieux

```
as.numeric((d[4] - d[2])/1.369)
```

5. On fait là allusion à une forme classique, dite de courbe en cloche

REMARQUE E.12. Afin de résoudre le problème des distributions dissymétriques, on emploie souvent la technique des transformations. Il s'agit de trouver une formule mathématique qui transforme la variable dissymétrique en une nouvelle variable plus symétrique. L'analyse statistique est alors plus facile. En revanche, on perd l'avantage des unités de mesures originelles souvent mieux connues. Les transformations les plus connues sont le logarithme, la racine carrée et l'inverse.

EXERCICE E.13. Pour le jeu de données '`TRANSFERTS.txt`',

- *Avec Rcmdr :*
  utiliser le menu déroulant "Données" et l'option "Gérer les variables dans le jeu de données actif" puis "Calculer une nouvelle variable". Dans la fenêtre de dialogue, en tant que variable existante indiquer : "Montant", en tant que nom de la nouvelle variable indiquer : "TransLog" et en tant qu'expression à calculer : "log(Montant+1)".
- *Sans Rcmdr :*
  Sans Rcmdr, évaleur directement "log(Montant+1)".

(1) Réaliser un histogramme sur la nouvelle variable : "Translog". Que constatez-vous ?

(2) Pourquoi avoir utilisé la formule $\log(x+1)$ plutôt que tout simplement $\log(x)$?

E.3.2.5. Le graphe quantile-quantile.

Ce graphe cherche à confronter l'histogramme des données à une forme prototypique, celle de la loi normale dite aussi courbe en cloche. Il est très important de déterminer si les données suivent approximativement cette forme car les procédures statistiques que nous verrons par la suite sont généralement basées sur cette hypothèses. Lorsque les données suivent la loi normale, les points sont situés exactement sur une droite. Toutefois, un écart est inévitable, l'écart normal étant symbolisé par deux courbes en pointillées sur le graphe. On constate sur cet exemple que les données sont globalement compatibles avec l'hypothèse de normalité, sauf en ce qui concerne la valeur extrême.

- *Avec Rcmdr :*
  On trouvera cela dans le menu "graphique".
- *Sans Rcmdr :*
  Il faut (éventuellement d'abord installer) et charger le package car et donc taper

```
library(car)
```

  puis on tapera

```
qq.plot(alimentation$ES)
```

Voir le graphique de la figure E.6 page suivante.

E.3.2.6. la multi-modalité.

EXERCICE E.14. Le fichier '`ROLLERS.txt`' contient le prix de rollers (en Francs 2000) pour 224 pratiquants de la région lyonnaise.

- Réaliser un histogramme de ces données.
- Recommencer en choisissant comme nombre de classes : 15
- Analyser la forme de cet histogramme. Que constatez-vous ?
- Que pensez-vous de la notion de valeur typique dans ce contexte ?

Il arrive que les données ne soient pas regroupées autour d'une unique valeur centrale mais qu'il existe plusieurs groupes. On parle alors de *multimodalité*. Dans ce cas, chaque groupe devrait faire l'objet d'une description quantifiée, mais cette analyse (dite analyse de mélange) est complexe et ne sera pas envisagée. On se contentera de descriptions verbales.

E.4. Dangers des mauvaises affectations ! !

Cette section facultative est réservée à ceux qui travaillent sans Rcdmr.

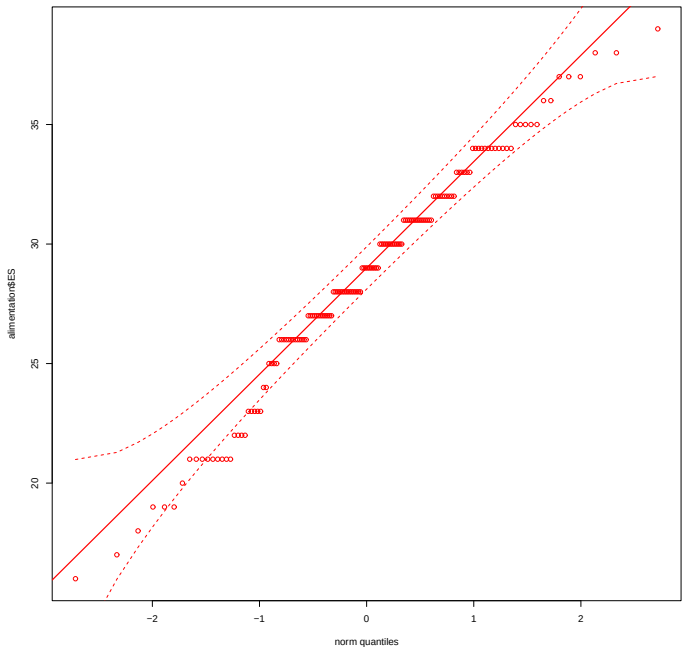


FIGURE E.6. Graphe quantile-quantile sur l'estime de soi.

EXERCICE E.15. *ATTENTION*, comme matlab ou d'autres logiciels du même type, ® 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 ®).

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
```

ANNEXE F

Le jeu " Pierre, Feuille, Ciseaux "

F.1. Introduction

On pourra consulter les trois url suivantes (dont certains passages ont eté repris et adaptés ici) :
`http://fr.wikipedia.org/wiki/Pierre-feuille-ciseaux`
`http://en.wikipedia.org/wiki/Rock-paper-scissors`
`http://de.wikipedia.org/wiki/Schere,_Stein,_Papier`

Le très célèbre jeu " Pierre-feuille-ciseaux ", noté PFC, est un jeu entre deux joueurs, selon les règles suivantes : La pierre bat les ciseaux (en les émoussant), les ciseaux battent la feuille (en la coupant), la feuille bat la pierre (en l'enveloppant). Ainsi chaque coup bat un autre coup, fait match nul lui-même et est battu par le troisième. En théorie, ce jeu est un jeu de hasard pur ; en fait, chacun obéit à une psychologie, inconsciente ou non et les tirages ne sont pas totalement aléatoires. À partir de cela, diverses théories de meilleure attaque se sont développées mais nous adopterons l'hypothèse d'aléat.

F.2. Le jeu à trois, quatre et cinq coups

Il existe de nombreuses variantes régionales ou nationales et appellations, souvent fondées sur trois coups possibles, chacun battant un autre et étant battu par le troisième. Dans certaines de ces variantes, de nouveaux symboles apparaissent : comme le puits battant la pierre ainsi que les ciseaux (en les faisant tomber au fond), et étant battu par la feuille (qui le recouvre). Afin de garder une probabilité de victoire égale entre chaque objet, un cinquième objet a été créé. Il est donc nécessaire qu'il batte deux des quatre objets existants et soit battu par les deux autres. Par exemple, le linuxien Rémi Pannequin [1] propose la toile d'araignée, tout en donnant une manière de produire ce signe à la main (main plate, doigts écartés). Voir `http://linuxfr.org/users/gabygaby/journaux/pierre-feuille-ciseaux-puits-toile-daraigné`. Le mot toile d'araignée pourra être abréger en " toile " . Les règles sont les suivantes : La toile vainc la feuille en l'étouffant et le puits en le bouchant. Elle est vaincue par la pierre qui la crève, et les ciseaux qui la coupent.

Notons que ces trois versions de jeux (à trois coups, PFC, quatre coups, noté PFCPu ou cinq coups, noté PFCPuT) peuvent être formalisées et généralisées de la façon suivante : il faut et il suffit de définir, pour toutes les paires possibles parmi les n coups choisis (ce qui fait donc un total de $p = C_n^2 = n(n-1)/2$), lequel des deux coups bat l'autre. On a donc un total de $p = n(n-1)/2$ règles pour n coups. On peut disposer ces règles dans la partie inférieure par exemple d'un tableau (ou d'une matrice) carrée . Pour chacun des coups numéro i, pour $1 \leq i \leq n$, on note le résultat contre le coup j, pour $1 \leq i \leq n$ et $j < i$, sous la forme $a_{ij} \in \{-1, 1\}$, la valeur 1, noté $+$ correspond à une victoire et la valeur -1, notée $-$ correspond à une défaite. De façon générale, le nombre de jeux différents, correspondant à toutes les règles possibles est donc égal à $q = 2^p = 2^{n(n-1)/2}$. Il est aussi possible de noter cela légèrement différement sous la forme d'une matrice de gain (`http://fr.wikipedia.org/wiki/Théorème_du_minimax_de_von_Neumann`). On considère un des joueurs, noté 1 et on définit le tableau (ou une matrice) carrée $A = (a_{ij})_{1 \leq i,j \leq n}$ de la façon suivante : les indices de lignes correspondent aux choix du premier joueur, tandis que les indices de colonnes correspondent aux choix du second joueur. a_{ij} est égal à 1, noté $+$, 0, si $i = j$ ou -1, noté $-$, selon que l'on ait une victoire,

1. avec son fils de six ans, en 2008 !

un match nul (si $i = j$) ou une défaite. Cette matrice est antisymétrique, puisque $a_{ij} = -aij$. Par exemple, dans le cas de PFC, le tableau est donné par :

	P	F	C
P	0	−	+
F	+	0	−
C	−	+	0

Ici les règles sont

La feuille bat la pierre ; (F.1a)

la pierre bat les ciseaux ; (F.1b)

les ciseaux battent la feuille. (F.1c)

Remarquons que chaque coup bat un des trois autres et est battu par le troisième ; cela peut aussi se traduire par le fait que

Chaque ligne du tableau contient autant de + que de −. (F.2)

Si l'on considère le jeu à quatre objets "Pierre, feuille, ciseaux, puits", noté PFCPu, pour lequel, aux règles précédentes, on adjoint les trois règles suivantes :

Le puits bat la pierre ; (F.3a)

la feuille bat le puits ; (F.3b)

le puits bat les ciseaux, (F.3c)

on aboutit au tableau suivant :

	P	F	C	Pu
P	0	−	+	−
F	+	0	−	+
C	−	+	0	−
Pu	+	−	+	0

On constate que sur ce tableau que (F.2) n'est plus vrai. La deuxième et la quatrième ligne contiennent deux + et un seul − : la feuille et le puits sont donc plus forts que les deux autres objets. Pour avoir un jeu équilibré, la règle (F.2) traduit le fait que, si les choix sont aléatoires, chaque coup a autant de chance de perdre que de gagner contre n'importe lequel des autres coups (distincts). Si on rajoute un objet, noté par exemple T, on peut alors définir totalement le tableau à cinq objets : on commence par remplir la dernière colonne, qui ne peut être que (+, −, +, −, 0), de façon à ce que les quatre premières lignes aient deux signes − et deux signes +. On remplit alors la dernière ligne comme l'opposé de cette colonne, qui est donc nécessairement (−, +, −, +, 0). On aboutit au tableau suivant :

	P	F	C	Pu	T
P	0	−	+	−	+
F	+	0	−	+	−
C	−	+	0	−	+
Pu	+	−	+	0	−
T	−	+	−	+	0

Ainsi, pour ce jeu à cinq coups, les règles sont donc nécessairement données par (F.1) , (F.3) et les 4 règles suivantes (en appelant "toile (d'araigné)" le dernier objet, noté T, comme a choisi le linuxien Rémi Pannequin)

$$\text{La pierre bat la toile ;} \tag{F.4a}$$
$$\text{la toile bat la feuille ;} \tag{F.4b}$$
$$\text{la toile bat le puits ;} \tag{F.4c}$$
$$\text{les ciseaux battent la toile.} \tag{F.4d}$$

Pour $n = 5$, on a donc bien un total de $p = n(n-1)/2 = 10$ règles.

F.3. Généralisation à un nombre de coups quelconque

Le site `http://www.umop.com/rps.htm` propose des règles à 7, 9, 11, 15, 25 et même à 101 coups ! En fait, une condition nécessaire pour que (F.2) ait lieu est bien sûr que

$$n \text{ soit impair.} \tag{F.5}$$

Sans avoir à apprendre par cœur les terrifiantes $p = C_1^2 01 = 5050$ règles de `http://www.umop.com/rps101.htm`, on peut en fait systématiser un jeu à un nombre impair de coup n de la façon suivante. On remarque que les tableaux à 3 et 5 coups obéissent à la loi suivante : la diagonale est formée de 0, la sous-diagonale de $+$, la sur-diagonale de $-$, et ainsi de suite. On peut donc proposer la chose suivante : on remplace chacun des objets par un numéro i dans $\{1, ..., n\}$. La matrice A est donnée alors par

$$A = \begin{pmatrix} 0 & - & + & - & + & \dots \\ + & 0 & - & + & - & \dots \\ - & + & 0 & - & + & \dots \\ + & - & + & 0 & - & \dots \\ & & & & & \\ \vdots & & & \ddots & \ddots & \ddots \\ & & & & & \\ & \dots & - & + & - & + & 0 \end{pmatrix}. \tag{F.6}$$

Ainsi, les éléments a_{ij} de la matrice sont définis par

$$\forall (i,j) \in \{1, ..., n\}, \quad a_{ij} = \begin{cases} 0, & \text{si } i = j, \\ +, & \text{si } ((i > j \text{ et } i - j \text{ est impair}) \text{ ou } (i < j \text{ et } i - j \text{ est pair})), \\ -, & \text{si } ((i > j \text{ et } i - j \text{ est pair}) \text{ ou } (i < j \text{ et } i - j \text{ est impair})). \end{cases} \tag{F.7}$$

Chacune des paires de coup (i, j) est donc définie par la règle suivante :

$$\text{si } i = j, \text{ match nul,} \tag{F.8a}$$
$$\text{si } i > j, \ i \text{ l'emporte sur } j \text{ ssi } i - j \text{ est impair,} \tag{F.8b}$$
$$\text{si } i < j, \ i \text{ l'emporte sur } j \text{ ssi } i - j \text{ est pair.} \tag{F.8c}$$

Dans le cas où $i > j$, on se rappelera de façon mnémotechnique que le plus grand nombre "perd" si la différence est "paire", autrement dit le plus grand nombre gagne si la différence est impaire. On peut aussi écrire cela de la façon équivalente suivante :

$$\text{si } i = j, \text{ match nul,} \tag{F.9a}$$
$$\text{si } i \text{ et } j \text{ sont distincts et n'ont pas la même parité, le plus grand l'emporte ;} \tag{F.9b}$$
$$\text{si } i \text{ et } j \text{ sont distincts et ont la même parité, le plus petit l'emporte.} \tag{F.9c}$$

C'est exactement, ce qui est proposé sur `http://en.wikipedia.org/wiki/Rock-paper-scissors` : "Alternatively, the rankings in rock-paper-scissors-Spock-lizard may be modeled by a comparison of the parity of the two choices. If it is the same (two odd-numbered moves or two even-numbered ones) then the lower number wins, while if they are different (one odd and one even) the higher wins." soit "Sinon, le classement peut être exprimé par une comparaison de la parité des deux choix. Si ce sont les mêmes (deux entiers impairs ou deux impairs), le nombre le plus faible l'emporte, tandis que si elles sont différentes (un impair et un pair), le plus élevé l'emporte."

Par exemple, pour $N = 11$, on a le tableau suivant :

	1	2	3	4	5	6	7	8	9	10	11
1	0	−	+	−	+	−	+	−	+	−	+
2	+	0	−	+	−	+	−	+	−	+	−
3	−	+	0	−	+	−	+	−	+	−	+
4	+	−	+	0	−	+	−	+	−	+	−
5	−	+	−	+	0	−	+	−	+	−	+
6	+	−	+	−	+	0	−	+	−	+	−
7	−	+	−	+	−	+	0	−	+	−	+
8	+	−	+	−	+	−	+	0	−	+	−
9	−	+	−	+	−	+	−	+	0	−	+
10	+	−	+	−	+	−	+	−	+	0	−
11	−	+	−	+	−	+	−	+	−	+	0

On peut montrer qu'avec ce choix, le jeu est bien équilibré, c'est-à-dire que (F.2) a lieu. En effet, la première ligne contient les symboles 0, −, +, ... Puisque n est impair, la première ligne contient donc autant de + que de −. Il en est de même pour les autres lignes.

F.4. Probabilités

En terme de probabilité, si on fait l'hypothèse de tirages aléatoire, chaque coup $i \in \{1, ..., n\}$ est équiprobable et la probablité que la variable alétaoire X égale à la valeur de i est donc donnée par

$$P(X = i) = \frac{1}{n}. \tag{F.10}$$

Si, pour une partie donnée, on s'intéresse cette fois-ci à la variable aléatoire Y égale à 1 (resp. -1) si le joueur 1 gagne (resp. perd) ou 0 s'il y a match nul. Y est donc le gain algébrique du joueur 1, au sens de `http://fr.wikipedia.org/wiki/Théorème_du_minimax_de_von_Neumann`. Si on fait l'hypothèse que chacun des joueurs joue indépendamment de l'autre, les deux valeurs de i et j sont indépendantes. La variable aléatoire (i, j) prend donc ses valeurs de façon équiprobable dans $\{1, ..., n\} \times \{1, ..., n\}$ et la probablité $P((i, j) = (i_0, j_0))$ où (i_0, j_0) est donné, est donc égale à $1/n^2$:

$$P((i, j) = (i_0, j_0)) = \frac{1}{n^2}. \tag{F.11}$$

Le match nul a lieu si $i = j$, de probablilité égale à $\sum_{i0=1}^{n} P((i, j) = (i_0, i_0))$, soit d'après (F.11) , $n/n^2 = 1/n$. De même, la victoire a lieu si et seulement si $Y = 1$. D'après la construction de la matrice de gain, cette probablité est égale au nombre de signe + dans la matrice A multiplié par $1/n^2$. La matrice A contient n symbole égaux à 0 et d'après (F.2) autant de symbole + que de − ; si r est le nombre de ces symboles, on a donc $2r + n = n^2$ et donc $r = (n^2 - n)/2 = n(n - 1)/2$. Ainsi, $P(Y = 1) = n(n - 1)/2/n^2 = (n - 1)/(2n)$. Il

est en de même pour $P(Y=-1)$. Ainsi

$$P(Y=0)=\frac{1}{n}, \tag{F.12a}$$

$$P(Y=1)=\frac{n-1}{2n}, \tag{F.12b}$$

$$P(Y=-1)=\frac{n-1}{2n}. \tag{F.12c}$$

Notons que, d'après (F.12a) , la probabilité de match nul diminue quand n augmente. De prendre n plus grand rend les règles plus nombreuses mais permet donc de diminuer le nombre de matchs nuls ! Cette probabiblité est égale 1/3 pour le jeu PFC et devient égale à 1/5 pour le jeu PFCPuT. Enfin, notons que

$$\mathbb{E}(Y)=0, \tag{F.13}$$

ce qui traduit que ce jeu est un jeu à somme nulle. Cette égalité provient de (F.12) ; en effet, par définition,

$$\mathbb{E}(Y)=0\times P(Y=0)+(1)\times P(Y=1)+(-1)\times P(Y=-1)=\frac{n-1}{2n}(1-1)=0.$$

F.5. Simulations aléatoires

On pourra consulter la fonction `pfc.R` disponible sur
`http://utbmjb.chez-alice.fr/UFRSTAPS/M2APA/fonctionsR/pfc.R`

Grâce à la fonction `pfc.R` on simule un tirage aléatoire correspondant à N parties, avec ici

$$N=5e+05. \tag{F.14}$$

Si on simule le jeu PFC, on a donc

$$n=3, \tag{F.15}$$

et pour le jeu PFCPuT, on a

$$n=5. \tag{F.16}$$

On pourra consulter les deux fichiers suivants, contenant les résultats de ces longues parties sur
`http://utbmjb.chez-alice.fr/UFRSTAPS/M2APA/donneesexamen/pfc3.txt`
`http://utbmjb.chez-alice.fr/UFRSTAPS/M2APA/donneesexamen/pfc5.txt`

Chacun de ces deux fichiers contiennent les données suivantes : `un`, `deux`, `res.un` et `res.deux`, qui contiennent respectivement, les coups du joueur 1 (qui peuvent être "C", "F" et "P" pour 3 ou "C", "F", "P", "Pu" et "T" pour 5) puis, de même, les coups du joueur 2, puis, pour le joueur 1, les résultats de la partie (qui peuvent être "d", "n" et "v", pour défaite, nul, victoire) et enfin, de même le résultat de la partie pour le joueur 2.

On peut dénombrer les matchs nuls, victoires et défaites pour le joueur 1 et en déduire les proportions suivantes, correspondant respectivement aux matchs nuls, victoires et défaites : pour $n=3$, on a

$$pr_{\mathrm{n}}=0.334124, \tag{F.17a}$$

$$pr_{\mathrm{v}}=0.332388, \tag{F.17b}$$

$$pr_{\mathrm{d}}=0.333488, \tag{F.17c}$$

et pour $n=5$, on a

$$pr_{\mathrm{n}}=0.200916, \tag{F.18a}$$

$$pr_{\mathrm{v}}=0.398942, \tag{F.18b}$$

$$pr_{\mathrm{d}}=0.400142, \tag{F.18c}$$

Les résultats donnés par (F.17) sont donc tout à fait conformes aux probabilités données par (F.12) qui valent ici

$$P(Y=0)=0.33333333,$$
$$P(Y=1)=0.33333333,$$
$$P(Y=-1)=0.33333333,$$

et les résultats donnés par (F.18) sont donc tout à fait conformes aux probabilités données par (F.12) qui valent ici

$$P(Y=0)=0.2,$$
$$P(Y=1)=0.4,$$
$$P(Y=-1)=0.4.$$

Notons aussi que l'on peut dénombrer les différents coups joués par chacun des joueurs et en déduire les proportions. Par exemple, pour $n=5$ pour le joueur 1, l'écart maximum entre les probabilités données par (F.11) et les proportions vaut :

$$\varepsilon=0.000488,$$

ce qui est très faible.

On peut aussi comparer les probablilités données par (F.11) et les proportions d'apparition de chacun des couples, dont l'écart maximum vaut

$$\varepsilon=0.00065488889,$$

pour $n=3$ et

$$\varepsilon=0.00041,$$

pour $n=5$.

Notons aussi que l'on peut étudier la corrélation entre les coups des deux joueurs, grâce aux fichiers
`http://utbmjb.chez-alice.fr/UFRSTAPS/M2APA/donneesexamen/pfc3bis.txt`
`http://utbmjb.chez-alice.fr/UFRSTAPS/M2APA/donneesexamen/pfc5bis.txt`
et la fonction `determin.qualiquali.R` qui nous donne ici une taille d'effet égale à

$$w=0.0012214027,$$

pour $n=3$ et

$$w=0.0018461237,$$

pour $n=5$, ce qui est très faible, donc la liaison est très faible. Attention, cependant aux probablités critiques égales à

$$p_c=0.47430142,$$

pour $n=3$ et

$$p_c=0.49197754,$$

pour $n=5$.

Enfin, notons, que, pour $n = 3$, les nombres de match nuls, de victoires et de défaites du joueur 1 sont respectivement égaux à

$$N_{\mathrm{n}} = 167062,$$
$$N_{\mathrm{v}} = 166194,$$
$$N_{\mathrm{d}} = 166744.$$

Ainsi, si on éliminine les match nuls, la proportion de victoire est égale à

$$p = \frac{N_{\mathrm{v}}}{N - N_{\mathrm{n}}} = \frac{166194}{5e + 05 - 167062} = 0.49917402.$$

La fonction `int.conf.prop.R` fournit donc l'intervalle de confiance suivant, au seuil usuel de 95% :

$$[0.49778812, 0.50055992]$$

qui contient bien 1/2. Si on choisit finalement $n = 5$, l'intervalle de confiance au seuil usuel de 95% est :

$$[0.49786324, 0.50063504]$$

qui contient bien 1/2. En fin de partie, le gain du joueur 1 est égal à (pour $n = 5$)

$$g = -600,$$

très faible par rapport à 5e+05.

Reprenons tous les calculs précédents si on choisit maintenant N toujours défini par (F.14) mais, par exemple,

$$n = 11. \tag{F.19}$$

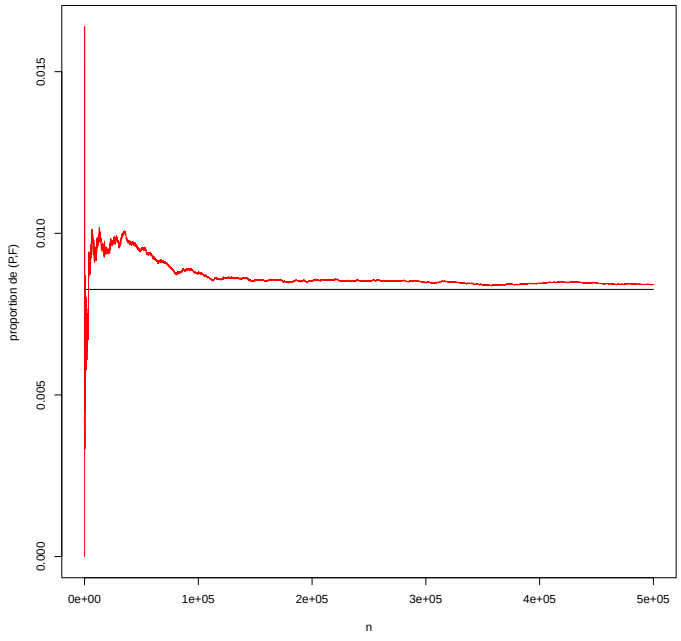


FIGURE F.1. Proportions d'apparition d'un couple donné en fonction du nombre de parties.

On peut dénombrer les matchs nuls, victoires et défaites pour le joueur 1 et en déduire les proportions suivantes, correspondant respectivement aux matchs nuls, victoires et défaites : pour $n = 11$, on a

$$pr_{\mathrm{n}} = 0.091038,$$
$$pr_{\mathrm{v}} = 0.45511,$$
$$pr_{\mathrm{d}} = 0.453852,$$

ce qui est tout à fait conforme aux probabilités données par (F.12) qui valent ici

$$P(Y = 0) = 0.090909091,$$
$$P(Y = 1) = 0.45454545,$$
$$P(Y = -1) = 0.45454545.$$

On peut aussi comparer les probablilités données par (F.11) et les proportions d'apparition de chacun des couples, dont l'écart maximum vaut

$$\varepsilon = 0.00035153719.$$

On pourra aussi consulter le graphique F.1 .

ANNEXE G

Croisement de deux variables quantitatives

Cette annexe s'inspire fortement du document [**1**] et du chapitre 9 de [**5**].
Cette annexe sera traitée rapidement comme révision.

G.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 [1] 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 H page 125) ou vers l'analyse de variance (une numérique et une catégorielle, voir chapitre I page 135).

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 [**12**].

G.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 existe mathématique existe entre X et Y, il existe donc une fonction f telle que en théorie

$$Y = f(X) \tag{G.1}$$

ce qui se traduira donc par

$$\forall i \in \{1, \ldots, n\}, \quad y_i \approx f(x_i). \tag{G.2}$$

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* :

$$f(X) = aX + b \tag{G.3}$$

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 (G.2) au *sens des moindres carrés*, c'est-à-dire trouver une fonction f parmi un ensemble de fonctions

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

données qui minimise l'expression :

$$\sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{G.4}$$

Plus petite sera cette quantité, meilleure sera l'approximation (G.2). Dans le cas de ce chapitre, on est dans l'hypothèse (G.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 :

$$\text{trouver } (a, b) \text{ qui minimise } S = \sum_{i=1}^{n} (ax_i + b - y_i)^2. \tag{G.5}$$

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

$$\forall (a, b) \in \mathbb{R}^2, \quad \sum_{i=1}^{n} (a_0 x_i + b_0 - y_i)^2 \leq \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{G.6}$$

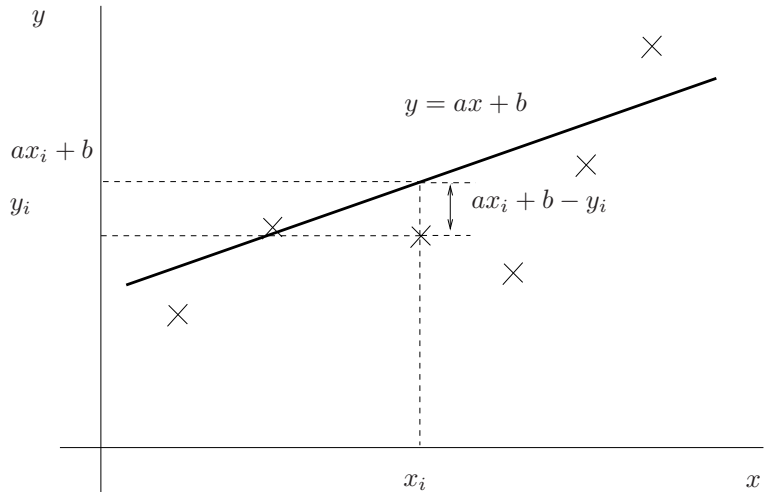


FIGURE G.1. le principe de la droite de régression linéaire

Voir la figure G.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`). Mais R sait déterminer ces coefficients, par la suite notés a et b.

Voir par exemple la figure G.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 [**12**].

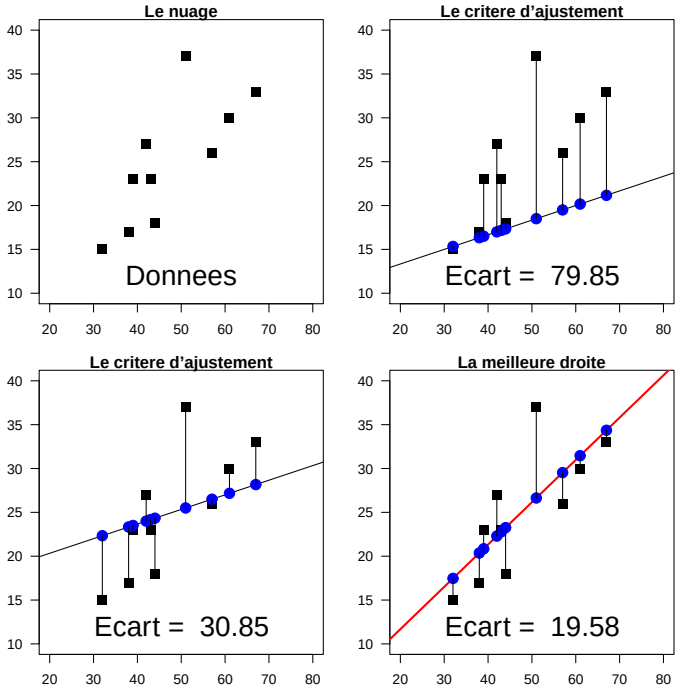


FIGURE G.2. la droite de régression linéaire

G.3. La significativité pratique de la liaison

EXEMPLE G.1. Avant de commencer à quantifer, 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 G.3 montre quatre situations possibles avec deux groupes représentées 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.

DÉFINITION G.2. On définit le *coefficient de corrélation linéaire*[2] 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 (G.1)-(G.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 (G.2) sera d'autant meilleure. Autrement dit, plus $|r|$ est proche de 1, plus l'écart S défini par (G.5) est proche de 0. Si $|r| = 1$, alors $S = 0$ et le points sont alignés.

Cohen dans [**13**] 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

$$\text{si } |r| \begin{cases} < r_1, & \text{la significativité pratique de la liaison linéaire est faible,} \\ \in [r_1, r_2[\,, & \text{la significativité pratique de la liaison linéaire est moyenne,} \\ \in [r_2, r_3[\,, & \text{la significativité pratique de la liaison linéaire est forte,} \\ > r_3, & \text{la significativité pratique de la liaison linéaire est très forte} \end{cases} \tag{G.7}$$

2. il sera plus exacte de dire affine.

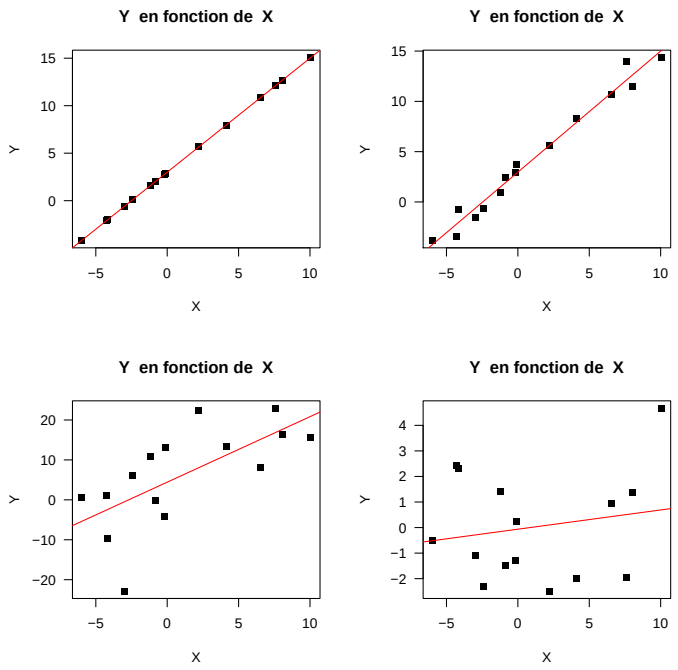


FIGURE G.3. Quatre situations concernant quatre nuages de points

EXEMPLE G.3. Donnons les différentes valeurs des r pour l'exemple G.1 page précédente qui sont dans l'ordre des graphiques de la figure G.3

$$r_1 = 1,$$
$$r_2 = 0.989064,$$
$$r_3 = 0.669144,$$
$$r_4 = 0.180478.$$

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

G.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*[3].

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 corélation 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 corélation 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. ♢

DÉFINITION G.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 G.5.

G.5. Avec R

L'exemple G.5 et l'exercice G.12 de la cette section sont issus de [**4**].

EXEMPLE G.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) :

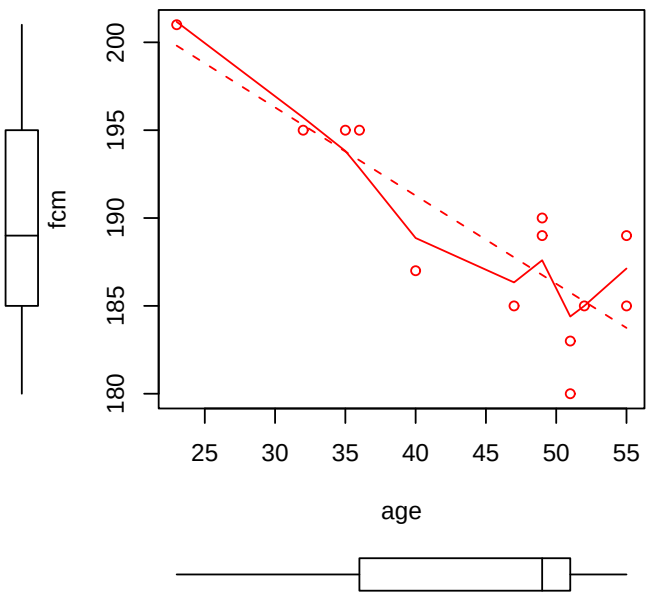
$$\text{fcm} = 220 - \text{âge}. \tag{G.8}$$

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

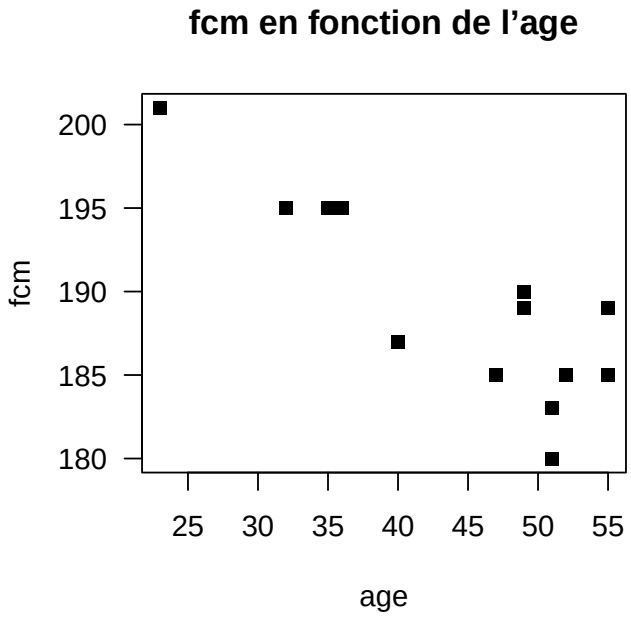
REMARQUE G.6. Pour toute la suite de ce cours, nous allons utiliser des fonctions. Lire (ou relire) pour cela l'annexe M page 185.

(1) La description d'une liaison entre deux variables numériques commence par la réprésentation 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.

- *Avec Rcmdr :*
  Il faut utiliser le menu déroulant "Graphes" du Rcmdr, puis l'option "Nuage de points". X sera ici la variable âge alors que Y sera la variable fcm. On obtiendra la figure du haut de la figure G.4 page suivante.
- *Sans Rcmdr :*
  La commande
  ```
  plot(coureurs$age, coureurs$fcm)
  ```
  donnera une figure ressemblant à celle de la figure du bas de G.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 G.4 page suivante.
  Si on veut obtenir la figure du haut, il faudra charger le package `car` en tapant par exemple



(a)



(b)

FIGURE G.4. Nuages de points $(x_i, y_i)_{1 \leq i \leq n}$ pour les données '`coureurs.txt`'

```
library(car)
```
puis
```
scatterplot(coureurs$age, coureurs$fcm)
```
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.

- *Avec Rcmdr :*
  On utilise le menu déroulant "Statistiques", et les options "Ajustement de modèles" et "Régression linéaire". Comme variable de réponse, on choisit la variable Y, ici fcm, et comme variable explicative la variable X, ici âge. 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 pente a se lit en face de `age` et l'ordonnée à l'origine b se lit en face de `(Intercept)`. On a donc ici

$$a = -0.5019, \tag{G.9a}$$
$$b = 211.3537. \tag{G.9b}$$

- *Sans Rcmdr :*
  On tape
```
coefficients(lm(coureurs$fcm ~ coureurs$age))
```
  et on voit apparaître
```
(Intercept)         age
211.3537051  -0.5019099
```

  Remarque G.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

$$a = -0.5019, \tag{G.10a}$$

$$b = 211.3537. \tag{G.10b}$$

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).

- *Avec Rcmdr :*
  La probabilité critique p_c se lit en face de `p-value`; on a donc

$$p_c = 0.000236 \tag{G.11}$$

- *Sans Rcmdr :*
  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

$$p_c = 0.000236 \tag{G.12}$$

- *Avec Rcmdr :*
  Pour le calcul de r, on a deux solutions

– On note que r^2 est donné par le nombre en face de `Multiple R-Squared` :

$$r^2 = 0.7218 \tag{G.13}$$

Pour déterminer r, il faut se rappeler que $\sqrt{r^2} = |r|$ et que le signe de r est égal à celui de a, soit

$$r = \text{signe}(a)\sqrt{r^2}.$$

Ici, on a donc

$$r = -\sqrt{r^2} = -0.8496$$

On a donc finalement

$$r = -0.8496 \tag{G.14}$$

– On peut aussi directement utiliser le menu déroulant "Statistiques" de Rcommander, puis les options "Résumés" et "Tests de corrélation". On sélectionne les deux variables qui nous intéressent (ici forcément age et fcm).
On obtient sur les dernières lignes de la fenêtre de sortie le résultat en dessous de cor :

```
        Pearson's product-moment correlation

data:  coureurs$age and coureurs$fcm
t = -5.3426, df = 11, p-value = 0.0002365
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9540024 -0.5614398
sample estimates:
       cor
-0.8496042
```

On retrouve donc bien la valeur donnée par (G.14).

Remarque G.8. *Attention* Si dans le meu déroulant les deux variables ne sont pas l'une derrière l'autre, comme ici, il faut appayer sur la touche "Ctrl" du clavier pour pouvoir les selectionner !

- *Sans Rcmdr :*
Pour obtenir r, on tape

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

Remarque G.9. *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(coureurs$age) & !is.na(coureurs$fcm)
r <- cor(coureurs$age[indc], coureurs$fcm[indc])
```

On a donc

$$r = -0.8496 \tag{G.15}$$

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

On peut utiliser la fonction `determin.quantiquanti` (voir annexe M page 185) 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(coureurs$age, coureurs$fcm)
```

ce qui donne les 4 valeurs (naturellement identiques à celle de (G.9), (G.11),(G.14))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(coureurs$age, coureurs$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(coureurs$age, coureurs$fcm)
determin.quantiquanti(coureurs$age, coureurs$fcm, fig = T)
determin.quantiquanti(coureurs$age, coureurs$fcm, fig = T, echo = F)
res <- determin.quantiquanti(coureurs$age, coureurs$fcm, echo = T)
res
```

Pour obtenir une figure analogue à celle de la figure G.5, il suffit de taper directement

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

Remarque G.10. Constater en tapant

```
determin.quantiquanti(coureurs$fcm, coureurs$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 G.11. En accord avec la remarque G.10, si on tape

```
lmd <- lm(fcm ~ age, data = coureurs)
lmdinv <- lm(age ~ fcm, data = coureurs)
par(mfrow = c(2, 1))
plot(coureurs$age, coureurs$fcm, pch = 15, las = 1, main = "fcm en fonction de l'age")
abline(lmd, col = "red")
```

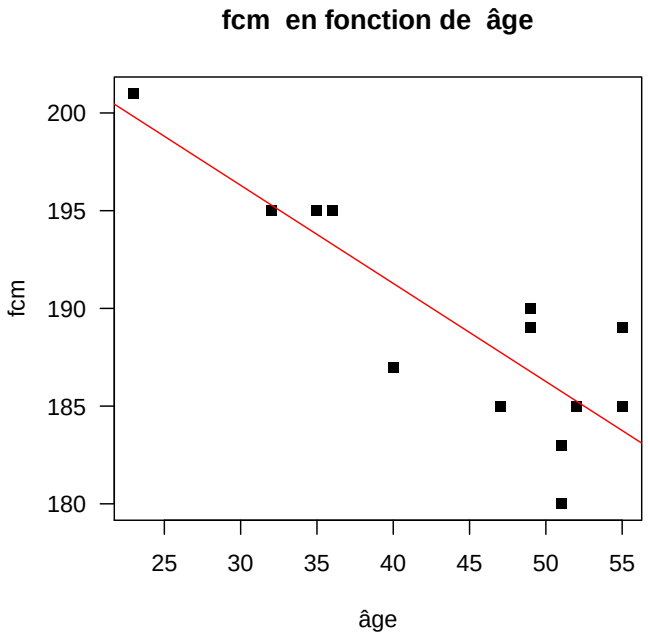


FIGURE G.5. Le nuage de point et la droite de régression pour les données 'coureurs.txt'

```
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 G.6 et G.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 G.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 G.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)
```

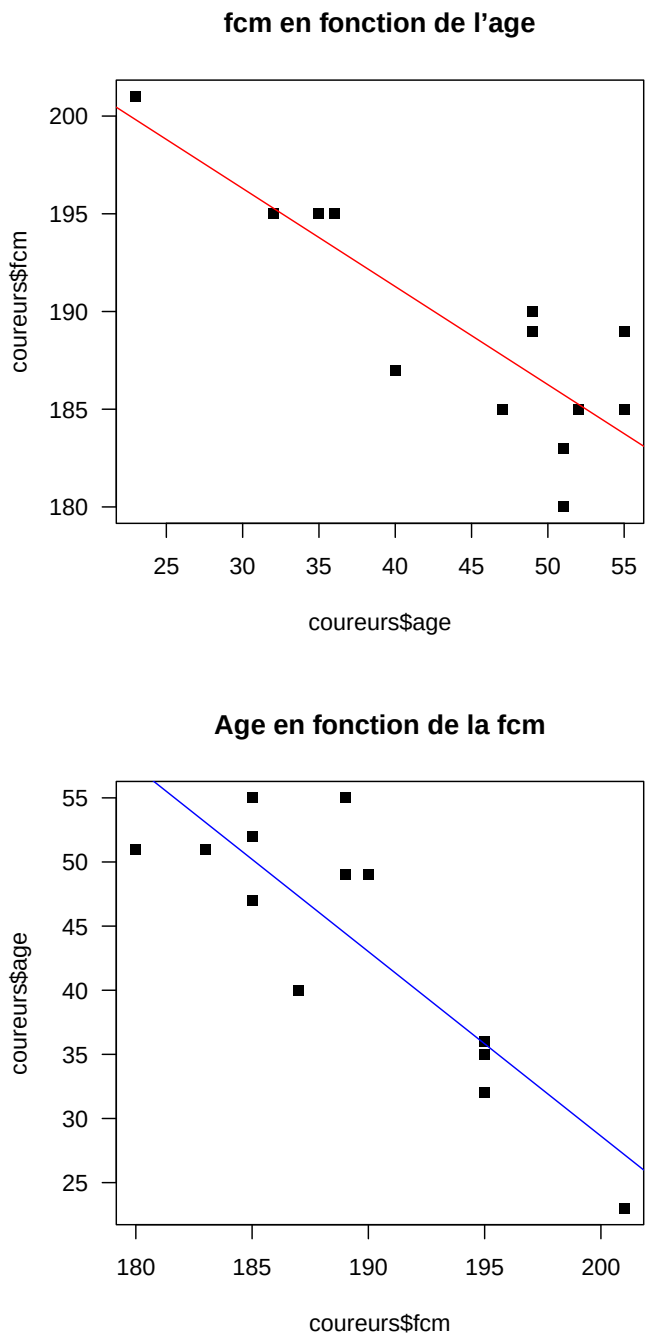


FIGURE G.6. Le nuage de point et les deux droites de régression pour les données '`coureurs.txt`'

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

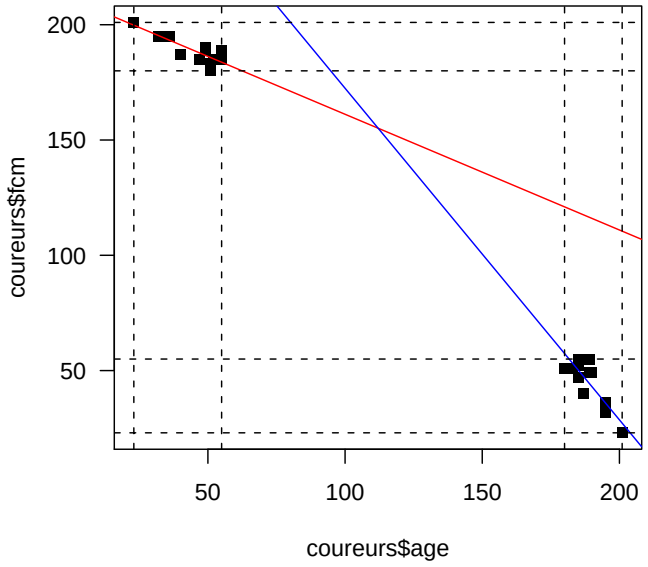


FIGURE G.7. Le nuage de point et les deux droites de régression sur le même graphe pour les données 'coureurs.txt'

```
points(age, fcmtheo, pch = 20, col = "blue")
segments(age, fcmtheo, age, fcm)
```

Voir figure G.9.

EXERCICE G.12. Charger le fichier `L3APA06.txt` à l'URL habituelle.

(1) Définissez une nouvelle variable égale à l'IMC : Indice de masse Corporelle, dont on rappelle la définition

$$\text{IMC} = \frac{\text{poids}}{\text{taille}^2}, \tag{G.16}$$

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

(2) Étudier les relations ('poids','taille'), ('poids','IMC') et ('taille','IMC').

Voir éléments de correction page 120

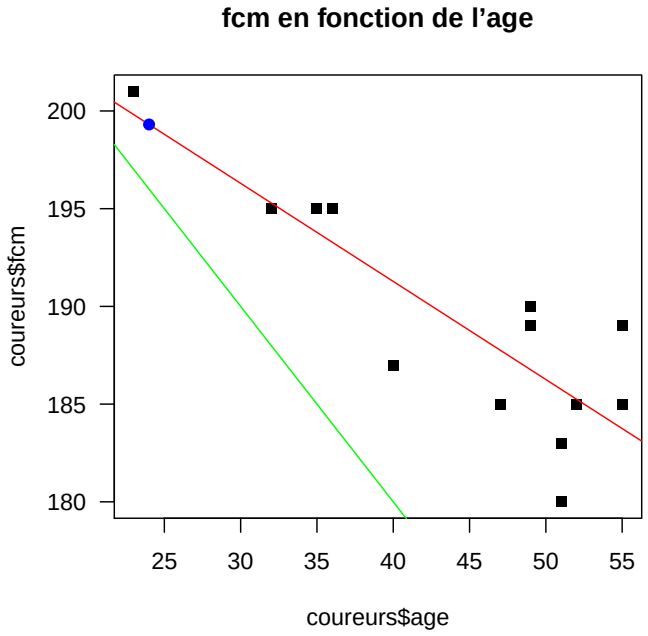


FIGURE G.8. Le nuage de point, la droite de régression avec le point de coordonnées $(24, 199.3079)$ et la droite d'Astran (fcm=220-âge) en vert pour les données '`coureurs.txt`'

G.6. Sur les dangers de la régression linéaire abusive : exemple d'Anscombe

On consultera l'annexe L.

G.7. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE G.12

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

$$\text{IMC} = \frac{\text{poids}}{(\text{taille}/100)^2}, \tag{G.17}$$

(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 G.10 page 122. Sur cette figure, les points semblent assez bien alignés.
- Confirmons cela grâce à R.
  Les résultats donnés par 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

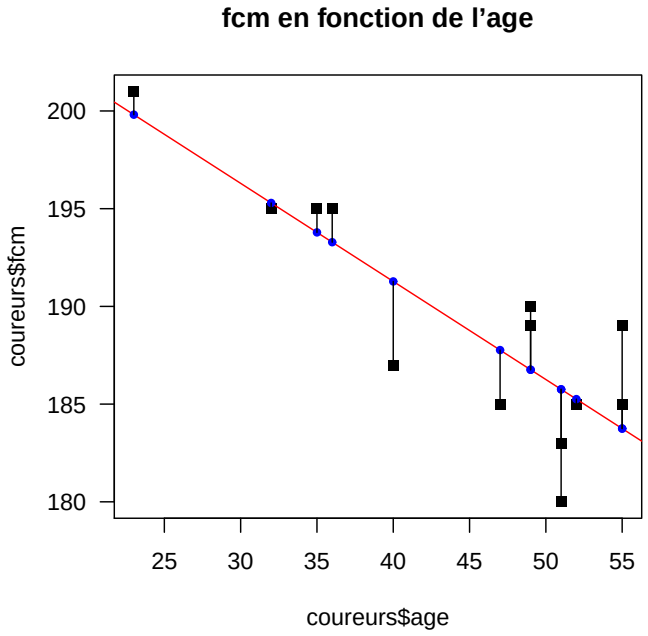


FIGURE G.9. Le nuage de point, la droite de régression et les fmc prédites pour les données '`coureurs.txt`'

On compare la valeur absolue de la corrélation linéaire r =0.891533 aux seuils de Cohen (0.1,0.3,0.5) (voir [**13**]) 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	**très forte**
significativité statistique	**oui**

- 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 G.11 page suivante. Sur cette figure, les points semblent très bien alignés.
- Confirmons cela grâce à R.
  Les résultats donnés par 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

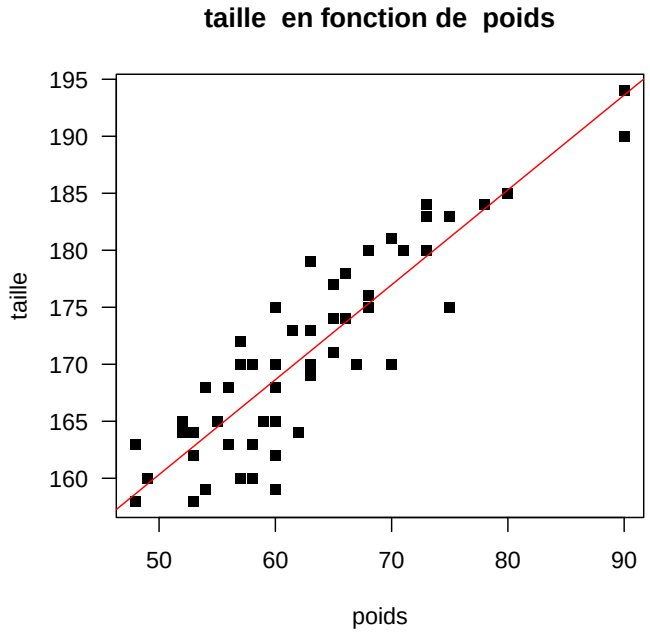


FIGURE G.10. Le nuage de point et la droite de régression

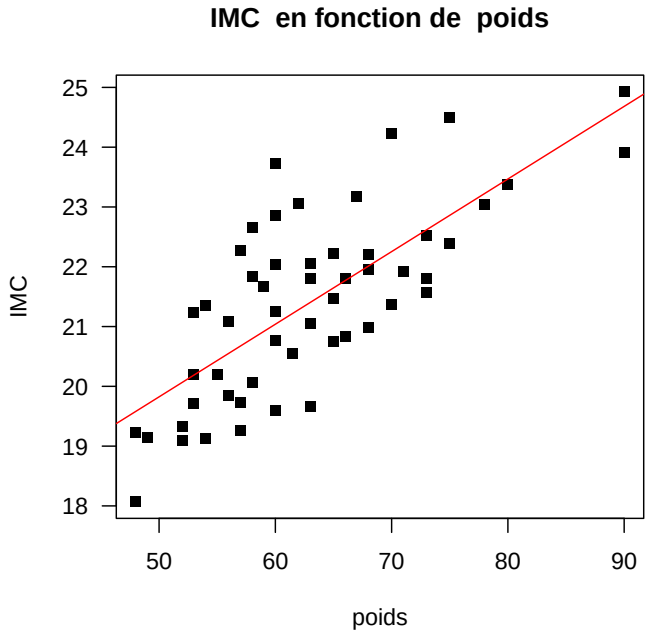


FIGURE G.11. Le nuage de point et la droite de régression

On compare la valeur absolue de la corrélation linéaire r =0.742738 aux seuils de Cohen (0.1,0.3,0.5) (voir [**13**]) 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	**très forte**
significativité statistique	**oui**

- 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 G.12. Sur cette figure, les points semblent très bien alignés.

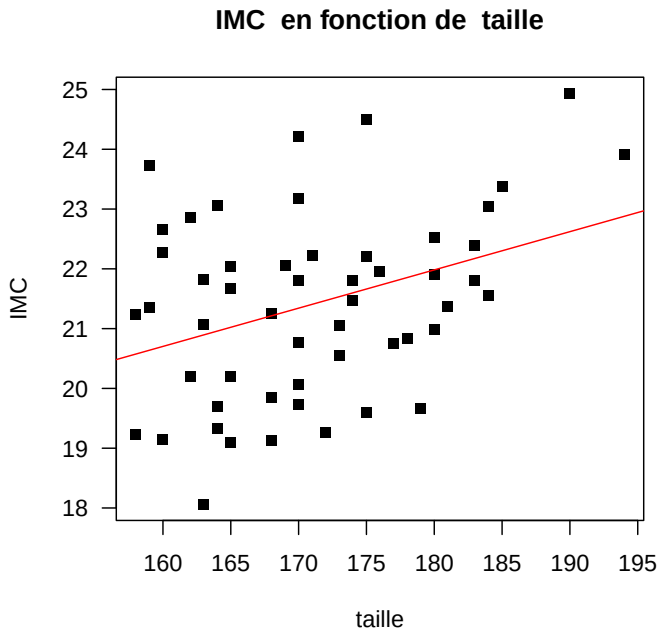


FIGURE G.12. Le nuage de point et la droite de régression

- Confirmons cela grâce à R.
  Les résultats donnés par R sont les suivants :

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

On compare la valeur absolue de la corrélation linéaire r =0.365269 aux seuils de Cohen (0.1,0.3,0.5) (voir [**13**]) 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	**forte**
significativité statistique	**oui**

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

ANNEXE H

Croisement de deux variables qualitatives

Cette annexe s'inspire fortement du document [**2**] et du chapitre 7 de [**5**].
Cette annexe sera traitée rapidement comme révision.

H.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.

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

TABLE H.1. Tableau de contingence

On obtient le tableau de contingence H.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 catégaries 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 H.1. De façon plus générale, on construit la table de contingence théorique sous l'hypothèse de l'indépendances 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 H.2 où

$$\widehat{n}_{ij} = \frac{n_{i.} n_{.j}}{n} \tag{H.1}$$

Remarquons *a posteriori* que les marges de la table des contingence théorique (H.1) sont bien celle de la table expérimentale : On a en effet les sommes en lignes

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

et les sommes en colonnes

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

Preuve facultative. la première hypothèse se traduit par l'existence d'un nombre α_i (qui ne dépend que la ligne i) tel que

$$\forall (i,j), \quad \widehat{n}_{ij} = \alpha_i n_{.j} \tag{H.2}$$

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

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

et donc

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

En reportant dans (H.2), on obtient

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

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

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

Table H.2. Tableau de contingence théorique

Définition H.2. On introduit enfin le χ^2 qui mesure l'écart entre la table théorique et la table observée défini par

$$\chi^2 = \sum_{ij} \frac{(\widehat{n}_{ij} - n_{ij})^2}{\widehat{n}_{ij}} \tag{H.3}$$

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

H.3. La significativité pratique de la liaison

Définition H.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

$$V = \sqrt{\frac{\chi^2}{n \min(p-1, q-1)}}. \tag{H.4}$$

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 [**13**] définie par

$$w = \sqrt{\frac{\chi^2}{n}}. \tag{H.5}$$

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

$$\text{si } w \begin{cases} < w_1, & \text{la significativité pratique de la liaison est faible,} \\ \in [w_1, w_2[, & \text{la significativité pratique de la liaison est moyenne,} \\ \in [w_2, w_3[, & \text{la significativité pratique de la liaison est forte,} \\ > w_3, & \text{la significativité pratique de la liaison est très forte} \end{cases} \tag{H.6}$$

H.4. La significativité statistique de la liaison

Comme dans la section G.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 H.5.

H.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 enquétés et en colonnes leurs réponses aux différentes questions. Nous avons conservé dans un tableau les résultats aux deux questions : classe d'age (attention, la variable age est quantitative, mais mise sous forme de classes, elle devient qualitative) et 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

- *Avec Rcmdr :*
  On utilise le menu déroulant "Statistiques" avec les options "Tables de contingence" et "Tableau à double entrée". Il faut indiquer la variable dont les catégories constitueront les lignes (ici age) et celles qui seront les colonnes (ici logement).
- *Sans Rcmdr :*
  Il faut taper la commande suivante :
  ```
  table(hebergement$age, hebergement$logement)
  ```

ou

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

	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 H.3. table de contingence du type d'hébergement en fonction des classes d'âges

On obtient le tableau H.3.

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

- *Avec Rcmdr :*
  on utilise de nouveau le menu déroulant "Statistiques" avec les options "Tables de contingence" et "Tableau à double entrée". Il faut indiquer la variable dont les catégories constitueront les lignes (ici age) et celles qui seront les colonnes (ici logement). il faut de plus cocher la case "Imprimer les fréquences attendues".
- *Sans Rcmdr :*
  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 H.4. table de contingence théorique du type d'hébergement en fonction des classes d'âges

On obtient le tableau H.4. Remarquons qu'il comprend $p = 9$ lignes et $q = 2$ colonnes.

(3) Pour obtenir le χ^2,

- *Avec Rcmdr :*
  On refait comme ci-dessus 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
```

- *Sans Rcmdr :*
  - 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

$$\chi^2 = 32.510687 \tag{H.7}$$

Remarque H.4. Seule cette dernière étape est totalement nécessaire !

(4) Pour calculer le coefficient de Cramer défini par (H.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 :

- *Avec Rcmdr :*
  On peut visualiser le tableau de données et constater qu'il contient $n = 591$ lignes. On peut visualiser le tableau `tab` et constater que $p = 9$ et que $q = 2$.
  Mieux, on peut procéder comme indiqué ci-dessous.
- *Sans Rcmdr :*
  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

$$n = 591,$$
$$p = 9,$$
$$q = 2$$

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

$$V = 0.2345413 \tag{H.8}$$

(5) On procède de même pour la taille d'effet (H.5) :

$$w = \sqrt{\frac{\chi^2}{n}},$$
$$= \sqrt{\frac{32.510687}{591}}$$

et donc

$$w = 0.2345413 \tag{H.9}$$

Ici, $w = V$! Pourquoi ?

Le V est proche de zéro donc les deux variables ne sont pas liées. Les seuils de Cohen données par (H.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

$$p_c = 7.5e-05 \tag{H.10}$$

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 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 contigence, un carré dont la surface lui est proportionnel.

Voir la figure H.1 page suivante.

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 H.1 a été crée 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 M) disponible sur le site et qui fournit directement les valeurs de χ^2, V, w et p_c :

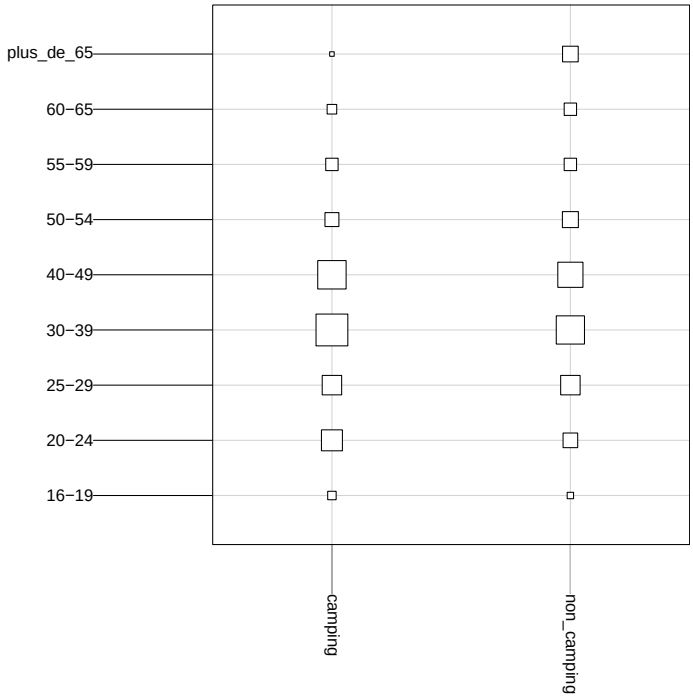


FIGURE H.1. Le graphique illustrant la table de contingence des données du fichier '`hebergement.txt`'.

```
 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 (H.7), (H.8), (H.9) et (H.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 H.5. Considérons le jeu de données portant sur 592 étudiants (extrait de [**14**]) 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 probalité critique entre les deux variables la couleur des cheveux et la couleur des yeux. Conclure

Voir éléments de correction page 133

H.6. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE H.5

La table de contingence donne

```
        y
x        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 H.5. Table de contigence des couleurs de cheveux en fonction de celle des yeux

Ces résultats sont présentés dans le tableau H.5. Avec les notations précédentes, on a

$$p = 4,$$
$$q = 4,$$
$$n = 592$$

On obtient grâce à R,

$$\chi^2 = 138.289842,$$
$$p_c = 2.32529e - 25.$$

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 χ^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 [**13**] (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.

ANNEXE I

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

Cette annexe s'inspire fortement du document [**3**] et du chapitre 5 de [**5**].
Cette annexe sera traitée rapidement comme révision.

I.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 reliée au groupe ? C'est cette deuxième formulation qui va expliquer l'organisation informatique des données.

I.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 [1].

I.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

- *Avec Rcmdr :*
  Utiliser le menu déroulant "Graphes", l'option "Boite de dispersion", sélectionner la variable '`notes3TDbis`' (par défaut) mais il faut de plus cliquer sur le bouton "Graphe par groupes" afin d'indiquer que l'on souhaite réaliser grâce à la variable '`groupes`', une boîte de dispersion pour chaque type de groupe.
- *Sans Rcmdr :*
  Taper la ligne de commande

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

```
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)
```

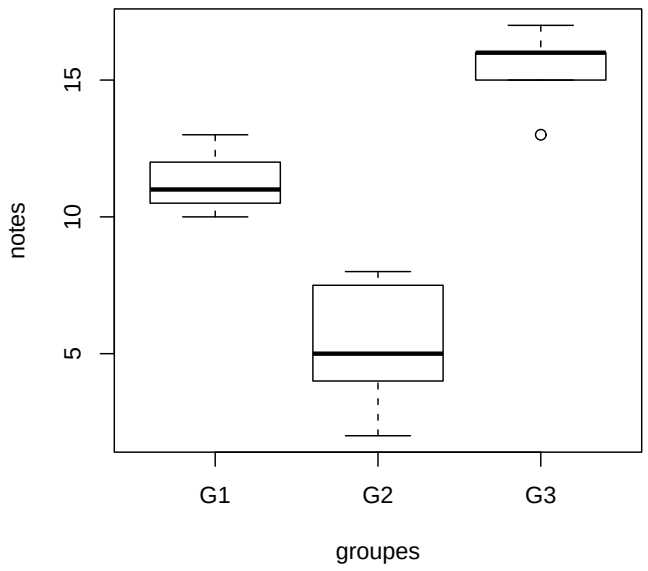


FIGURE I.1. La collection de boites de dispersion des notes par groupe (données 'notes3TDbis.txt')

Que voit-on sur le graphique I.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 I.2 page suivante) Pour obtenir ce dessin, il faut taper dans la fenêtre "Rgui"

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

I.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,

- *Avec Rcmdr :*

  Il faut utiliser le menu déroulant "Statistiques", puis "Résumés", puis "Statistiques descriptives". On sélectionne la variable (par défaut) 'notes', mais il faut de plus cliquer sur le bouton "Résumer par groupe" et indiquer la variable 'groupes'. On obtient alors dans le fenêtre de sortie le résultat ci-dessous.

```
        mean       sd 0%  25% 50%  75% 100% n
G1 11.333333 1.527525 10 10.5  11 12.0   13 3
```

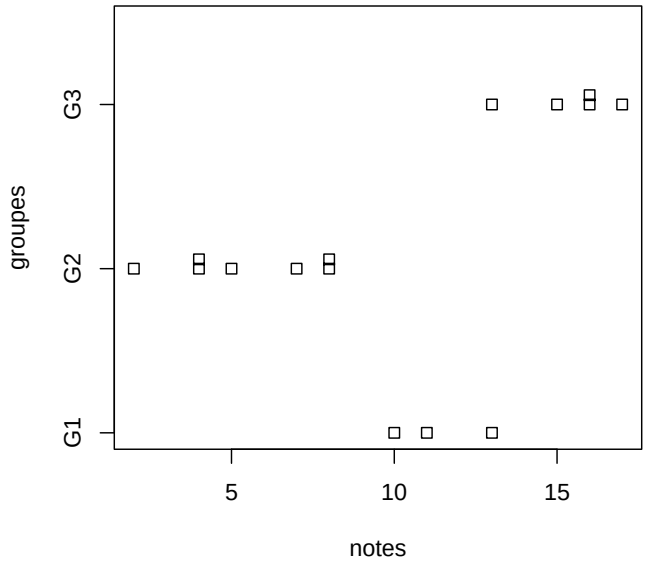


FIGURE I.2. La collection de lignes de points des notes par groupes (données 'notes3TDbis.txt')

```
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
```

- *Sans Rcmdr :*

  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

  ```
  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 synthaxe utilisée, voir l'annexe M page 185)

```
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.

I.2.3. La significativité pratique de la liaison

Exemple I.1. Avant de commencer à quantifer, il faut d'abord comprendre dans quelles situations on considère qu'une liaison est intense. Le graphique I.3 page suivante montre quatre situations possibles avec deux groupes représentées par des collections de lignes de points empilés.

– 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 connait 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.

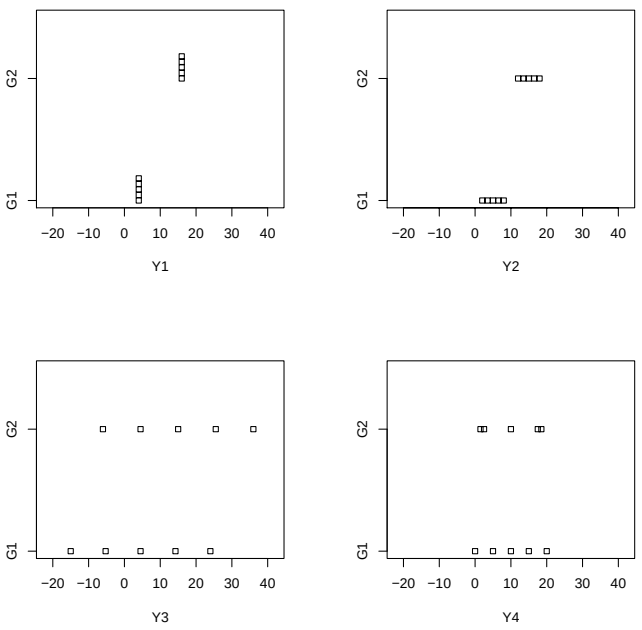


FIGURE I.3. Quatre situations concernant deux groupes avec 5 mesures numériques pour chacun

- 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.

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 I.4 page suivante 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 I.4 page suivante, 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

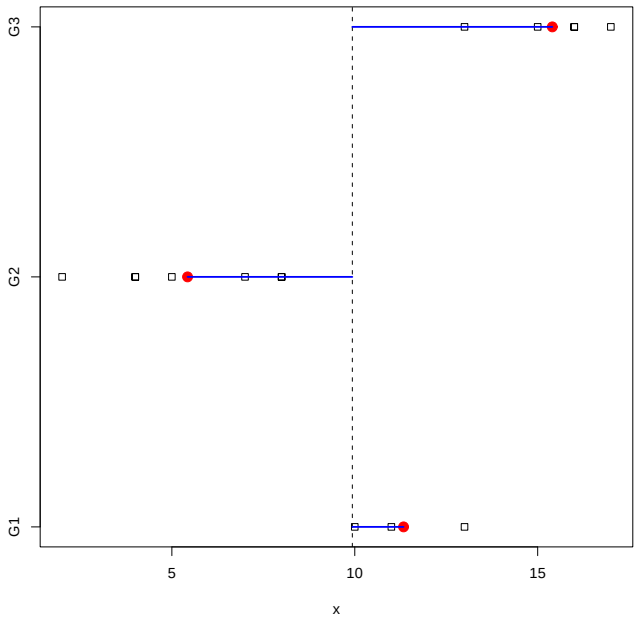


FIGURE I.4. La collection de lignes de points des notes par groupes (données '`notes3TDbis.txt`') et les moyennes des groupes.

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_{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 ($j = 1, \cdots, G$) :

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

♢

Définition I.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 :

$$\mathrm{RC} = \frac{SC_{\text{Inter-Groupes}}}{SC_{\text{Totale}}} \tag{I.2}$$

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 I.3. Donnons les différentes valeurs des RC pour l'exemple I.1 page 138 qui sont dans l'ordre des graphiques de la figure I.3 page 139

$$\begin{aligned}\mathrm{RC}_1 &= 1,\\ \mathrm{RC}_2 &= 0.847458,\\ \mathrm{RC}_3 &= 0.118357,\\ \mathrm{RC}_4 &= 0.\end{aligned}$$

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 (I.1), on a l'expression exacte du RC :

$$\mathrm{RC} = \frac{\sum_{j=1}^{G} n_j (M_j - M)^2}{\sum_{j=1}^{G}\sum_{i=1}^{n_j} (y_{ij} - M)^2}. \tag{I.3}$$

♢

Pour calculer cette quantité,

- *Avec Rcmdr :*

  il faut utiliser le menu déroulant "Statistiques", utiliser l'option "Ajustement de modèle" et "Modèle linéaire". Dans la cellule vide à gauche, indiquer la variable numérique, c'est-à-dire ici pour les données du fichier '`notes3TDbis.txt`', '`notes`' et dans la cellule de droite [2] la variable qualitative (ou catégorielle) : '`groupes`'. 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 *
```

2. Il est possible de le faire en cliquant le nom de ces variables dans la liste au dessus. Toutefois, Attention, pour les versions de R antérieures à 7.2, lorsque la variable catégorielle est choisie pour la cellule de droite, elle est accompagné entre crochets de l'expression `factor`. Il faut absolument éliminer cette mention sinon il y a un bug !. Ce bug a été corrigé dans les versions suivantes de R.

```
---
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
```

- *Sans Rcmdr :*
  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

$$\mathrm{RC} = 0.867085 = 86.71\%$$

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 I.2.4).

En ce qui concerne la significativité pratique, une grille d'interprétation qualitative a été proposée par Cohen [**13**] 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

$$\text{si RC} \begin{cases} < RC_1, & \text{la significativité pratique de la liaison est faible,} \\ \in [RC_1, RC_2[, & \text{la significativité pratique de la liaison est moyenne,} \\ \in [RC_2, RC_3[, & \text{la significativité pratique de la liaison est forte,} \\ > RC_3, & \text{la significativité pratique de la liaison est très forte} \end{cases} \tag{I.4}$$

I.2.4. La significativité statistique de la liaison

Comme dans la section G.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,

- *Avec Rcmdr :*

  On procède comme page 141 et on regarde les deux dernières lignes, en face de `p-value`. dans le cas des notes des étudiants, on a

  ```
  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
  ```

- *Sans Rcmdr :*

  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

$$p_c = 5.51369e-06 = 0.00055\%, \tag{I.5}$$

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
```

I.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
```

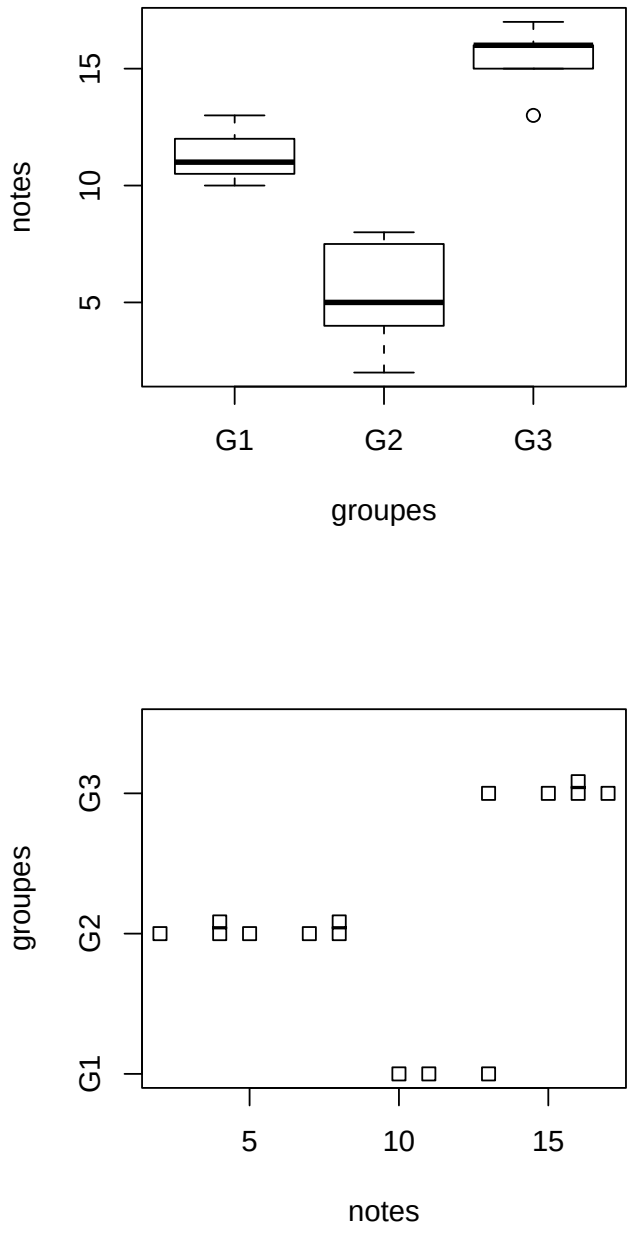


FIGURE I.5. Les collections de boites de dispersion et des lignes de points des notes par couleurs de cheveux (données '`notes3TDbis.txt`')

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 I.5 page ci-contre.

I.4. Quelques exercices

Exercice I.4.

	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 I.1. Notes de statistique selon la couleur de cheveux des étudiantes

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 I.1.

Voir le fichier '`blondes.txt`'.

Est-ce que la note de statistique est reliée à la couleur des cheveux ?

Voir éléments de correction page 145

Exercice I.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 147

Exercice I.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 menstruel. 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 ?

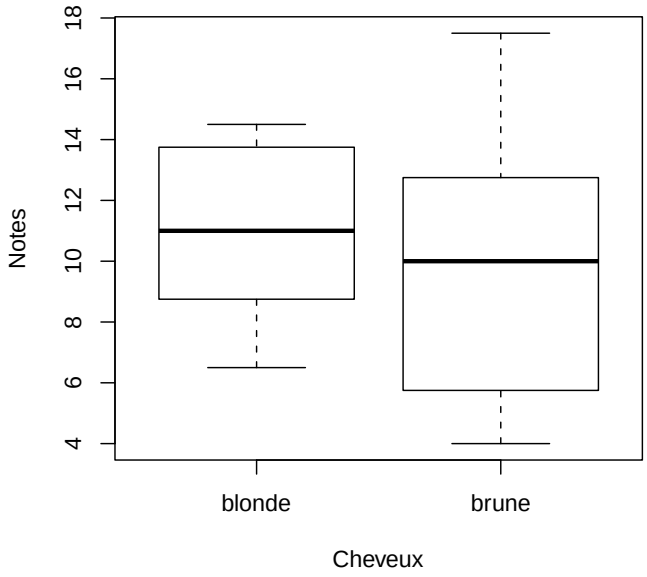
Voir éléments de correction page 149

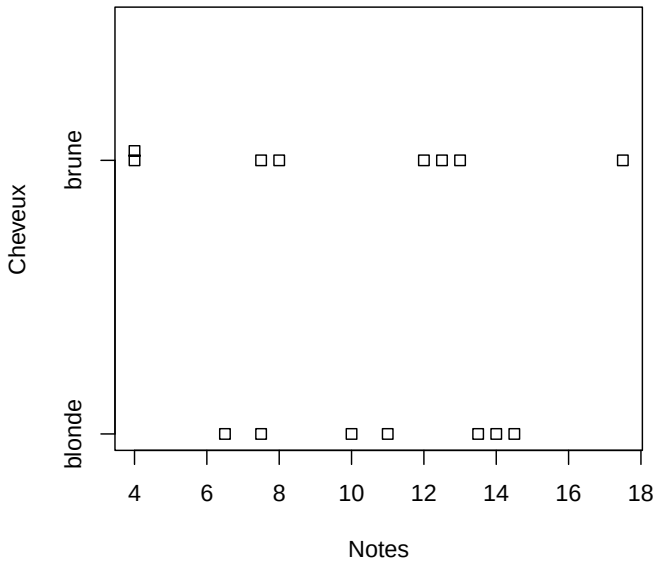
I.5. Éléments de correction

Éléments de correction de l'exercice I.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 R, on obtient les statistiques par groupes données dans le tableau suivant ;

On rappelle que, dans ce tableau :

- le nombre noté `0%` est le quartile à 0 % (c'est le minimum) ;

	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

- 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 R.

Les autres résultats donnés par 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 [**13**]) 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	**moyenne**
significativité statistique	**non**

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

ÉLÉMENTS DE CORRECTION DE L'EXERCICE I.5

- *Avec Rcmdr :*

En faisant la manipulation décrite précédemment, on obtient

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.5345 -3.5345  0.4741  3.4741  6.4569

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.552      3.714   0.956    0.356
notes[, ind.travail[2]]   2.991      1.650   1.813    0.093 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.589 on 13 degrees of freedom
Multiple R-squared: 0.2018,     Adjusted R-squared: 0.1404
F-statistic: 3.286 on 1 and 13 DF,  p-value: 0.093
```

et donc une probabilité critique égale apparemment à 0.093 et un RC apparemment égal à 0.2018. En fait, ici R considère la deuxième variable (dont les valeurs dont '1',...) comme numérique et la probablité critique renvoyée est celle du croisement de deux variables numériques !

Il faut rendre la variable '`groupes`' factorielle en allant dans "Données", puis "Gérer les variables...", puis "convertir var. numériques en facteur". En procédant comme précédemment, on obtient alors

```
Call:
lm(formula = notes[, ind.travail[1]] ~ as.factor(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 ***
as.factor(notes[, ind.travail[2]])2  -5.905      1.345  -4.390  0.00088 ***
as.factor(notes[, ind.travail[2]])3   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
```

et donc une probabilité critique égale cette fois-ci à 5.514e-06 et un RC égal à 0.8671.

- *Sans Rcmdr :*

  Attention, si on tape sans prendre de précaution

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

  la probabilité critique vaudrait

$$p_c = 0.09300072, \tag{I.6}$$

à comparer avec (I.5).

En fait, ici R considère la deuxième variable (dont les valeurs dont '1',...) comme numérique et la probablité critique renvoyée est celle du croisement de deux variables numériques ! En effet, si on utilise la fonction `determin.quantiquanti` du chapitre G 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 (I.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 (I.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 I.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 I.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 I.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 maitenant le rapport de corrélation et la probabilité critique en tapant

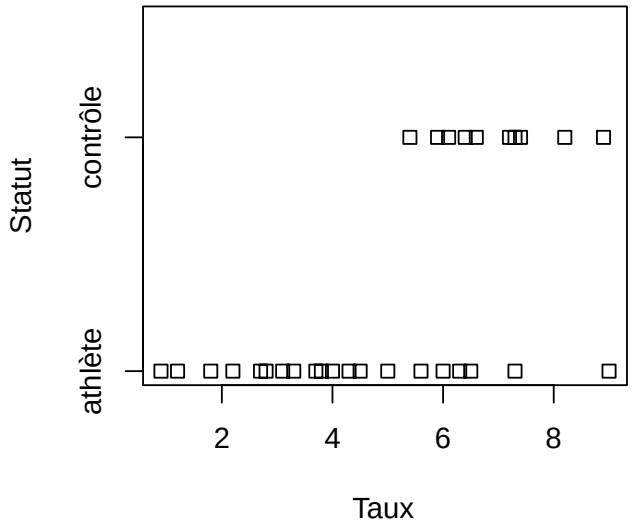
```
determin.qualiquanti(TESTOSTERONE$Taux,TESTOSTERONE$Statut)
```

```
$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
```

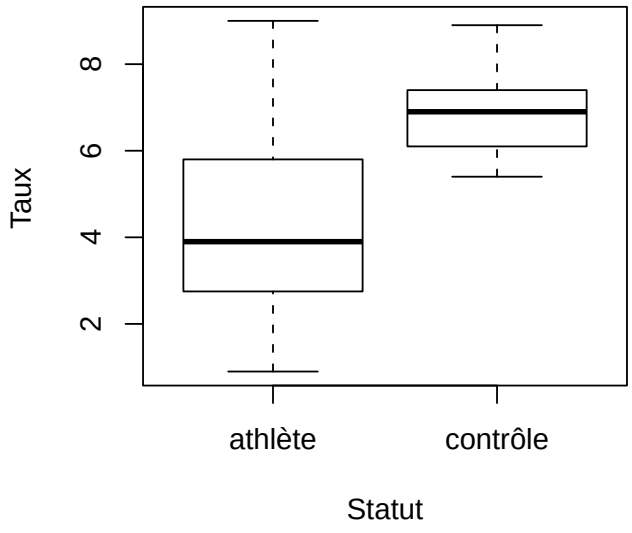
On obtient donc

$$\mathrm{RC} = 0.344999, \quad p_c = 0.000643707.$$

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 I.6. Deux collections pour les données '`TESTOSTERONE.txt`'.

ANNEXE J

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.

Vous trouverez à partir de la section J.1.1 page suivante les manipulations avec Rcmdr et à partir de la section J.3.1 page 154 les manipulations directement avec "Rgui".

J.1. Analyse univariée (avec Rcmdr)

J.1.1. Importer des données (avec Rcmdr)

Voir section D.1 page 79.

J.1.2. Avec Rcmdr

Pour lire dans R le contenu du fichier '`nom.txt`' :

(1) Dans le menu déroulant "Données" de Rcmdr, choisir l'option "Importer des données" puis "Depuis un fichier texte ou le presse-papier...". Dans la fenêtre de dialogue qui s'ouvre, donner un nom au jeu de données (à la place de Dataset, choisi par défaut), le nom du fichier texte sans extension, c'est-à-dire : '`nom`'. Laisser les autres champs avec les valeurs choisies par défaut.

(2) Employer la fenêtre qui s'ouvre alors pour retrouver le fichier à importer ('`nom.txt`').

(3) Cliquer alors éventuellement sur le bouton "Visualiser".

J.1.3. Variable qualitative (ou catégorielle) (avec Rcmdr)

Voir sections D.3 page 81 et D.4 page 81.

J.1.3.1. Indicateurs.

Pour afficher les différentes fréquences et les pourcentages, il faut aller dans le menu déroulant "Statistiques" du Rcommander et choisir les options "Résumés" et "Distributions de fréquences".

J.1.3.2. Graphiques.

(1) Peu de données :
- Camembert (tourte) :
  Aller dans le menu déroulant "Graphes" de Rcmdr et choisir l'option "Graphe en camembert".
- Diagramme en barre
  Aller dans le menu déroulant "Graphes" de Rcmdr et choisir l'option "Graphe en barres".

(2) Beaucoup de données :
- Diagramme de Cléveland : il faut taper dans la fenêtre de "Rgui"
  ```
  dotchart(table(nom$variableY))
  ```
- Diagramme de Cléveland (classé) : il faut taper dans la fenêtre de "Rgui"
  ```
  dotchart(sort(table(nom$variableY)))
  ```

J.1.4. Variable quantitative (ou numérique) (avec Rcmdr)

Voir sections E.3 page 88 et E.2 page 87

J.1.4.1. Indicateurs.

On utilise le menu déroulant "Statistiques", puis les options "Résumés" et "Statistiques descriptives".

J.1.4.2. Graphiques.

(1) Peu de données :
- Ligne de point (avec empilement) : il faut taper dans la fenêtre de "Rgui"
  ```
  stripchart(nom$variableY, method = "stack")
  ```

(2) Beaucoup de données :
- Histogramme : Il faut utiliser le menu déroulant "Graphes" et l'option "Histogramme".
- Boîte de dispersion : Il faut utiliser le menu déroulant "Graphes", puis "Boîte de dispersion".

J.2. Analyse bivariée (avec Rcmdr)

On veut croiser les variables '`variableX`' et '`variableY`' du data frame '`nom`'.

J.2.1. Quantitatif × quantitatif (avec Rcmdr)

Voir chapitre G page 107.

- 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 (G.7) page 109) ;
  (3) p_c à la valeur seuil de 0.05 (voir définition G.4 page 111).
- Pour obtenir en plus la figure (nuage de point et droite de corrélation) :
  `determin.quantiquanti(nom$variableX, nom$variableY, fig = T)`

J.2.2. Qualitatif × qualitatif (avec Rcmdr)

Voir chapitre H page 125.

- Après avoir téléchargé la fonction '`determin.qualiquali`', il faut la sourcer.
- Pour obtenir le coefficient χ^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 est proche de 1, les variables sont liées.
  (2) w aux seuils de Cohen $(0.1, 0.3, 0.5)$ (voir équation (H.5) page 128) ;
  (3) p_c à la valeur seuil de 0.05 (voir définition G.4 page 111).

J.2.3. Qualitatif × quantitatif (avec Rcmdr)

Voir chapitre I page 135.

- 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 est proche de 1, les variables sont liées.
  (2) RC aux seuils de Cohen $(0.01, 0.05, 0.15)$ (voir équation (I.4) page 143) ;
  (3) p_c à la valeur seuil de 0.05 (voir section I.2.4 page 143).
- Pour obtenir en plus la figure (boîtes de dispersion et lignes de point par groupes) :
  `determin.qualiquanti(nom$variableX, nom$variableY, fig = T)`

J.3. Analyse univariée

J.3.1. Importer des données

Voir sections D.1 page 79 et D.2 page 79.

- 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 totalités 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)
  ```

J.3.2. Variable qualitative (ou catégorielle)

Voir sections D.3 page 81 et D.4 page 81.
On suppose que la commande '`read.table`' a déjà été tapée !

J.3.2.1. 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)
  ```

J.3.2.2. 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 :
   - Diagramme de Cléveland
     ```
     dotchart(table(nom$variableY))
     ```

- Diagramme de Cléveland (classé)
  `dotchart(sort(table(nom$variableY)))`

J.3.3. Variable quantitative (ou numérique)

Voir sections E.2 page 87 et E.3 page 88.
On suppose que la commande '`read.table`' a déjà été tapée !

J.3.3.1. 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)`

J.3.3.2. 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)`

J.4. Analyse bivariée

On veut croiser les variables '`variableX`' et '`variableY`' du data frame '`nom`'.

J.4.1. Quantitatif × quantitatif

Voir chapitre G page 107.
- 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 (G.7) page 109) ;
  (3) p_c à la valeur seuil de 0.05 (voir définition G.4 page 111).
- Pour obtenir en plus la figure (nuage de point et droite de corrélation) :
  `determin.quantiquanti(nom$variableX, nom$variableY, fig = T)`

J.4.2. Qualitatif × qualitatif

Voir chapitre H page 125.
- Après avoir téléchargé la fonction '`determin.qualiquali`', il faut la sourcer.
- Pour obtenir le coefficient χ^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 est proche de 1, les variables sont liées.
  (2) w aux seuils de Cohen $(0.1, 0.3, 0.5)$ (voir équation (H.5) page 128) ;
  (3) p_c à la valeur seuil de 0.05 (voir définition G.4 page 111).

J.4.3. Qualitatif × quantitatif

Voir chapitre I page 135.

- 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 est proche de 1, les variables sont liées.
  (2) RC aux seuils de Cohen $(0.01, 0.05, 0.15)$ (voir équation (I.4) page 143) ;
  (3) p_c à la valeur seuil de 0.05 (voir section I.2.4 page 143).
- Pour obtenir en plus la figure (boîtes de dispersion et lignes de point par groupes) :
  `determin.qualiquanti(nom$variableX, nom$variableY, fig = T)`

ANNEXE K

Projet

Ce projet ne sera plus traité en cours, contrairement aux années précédentes, mais pourra tenir lieu de révisions à faire chez vous !

K.1. Quelques définitions

K.1.1. OVE : oeuvre des villages d'enfants

Voir par exemple `http://www.ove.asso.fr/`.
Voir aussi le polycopié de cours [**15**].

K.1.2. SESSAD : Service d'Éducation Spéciale et de Soins à Domicile

Extrait de `|http://scolaritepartenariat.chez-alice.fr/page75.htm` :
"Les SESSAD sont devenus, dans le secteur médico-éducatif, la structure privilégiée de l'aide à l'intégration scolaire (1)...

L'éducation nationale a obligation d'ouvrir des classes ou même des écoles dans les établissements spécialisés, IME (Institut médico-éducatif) ou IR (Institut de rééducation), voire dans les hôpitaux de jour. Ce fonctionnement répond à un droit légitime des enfants handicapés, il a été bien cadré par les textes dès la Loi de 75 et ses premières circulaires d'application, et il ne pose pas de problème d'ordre administratif (2). Mais on peut concevoir également que ce soient les personnels du secteur médico-éducatif qui viennent travailler dans ou avec une école, auprès d'un enfant en intégration scolaire . C'est précisément le SESSAD : Service d'éducation spécialisée et de soins à domicile. Un SESSAD, pour être bref, c'est un établissement ou une partie d'un établissement, qui devient mobile et qui va travailler "à domicile"... Précisons d'emblée, pour éviter tout malentendu, que le terme de "domicile", dont l'utilisation pourrait prêter à confusion, marque essentiellement la différence d'avec l'établissement spécialisé. Le domicile, en l'occurrence, ce sont les lieux où l'enfant vit et où il exerce ordinairement ses activités. (3) "

K.1.3. ITEP : Les Instituts Thérapeutiques, Educatifs et Pédagogiques

Extrait de `http://daniel.calin.free.fr/itep.html`
"Les Instituts Thérapeutiques, Educatifs et Pédagogiques (ITEP) sont des établissements médico-éducatifs qui ont pour vocation d'accueillir des enfants ou des adolescents présentant des troubles du comportement importants, sans pathologie psychotique ni déficience intellectuelle. Ce sont les anciens Instituts de Rééducation (IR), ou Instituts de Rééducation Psychothérapeutique (IRP), réformés par le décret n° 2005-11 du 6 janvier 2005. L'accueil se fait en internat ou demi-pension. L'enseignement est dispensé soit dans l'établissement par des enseignants spécialisés, soit en intégration dans des classes, ordinaires ou spécialisées, d'établissements scolaires proches."

K.1.4. IME : Les Instituts Médico-Éducatifs

Extrait de `http://daniel.calin.free.fr/ime.html`
"Les IME sont des établissements médico-éducatifs qui accueillent les enfants et adolescents atteints de déficience mentale. Ils sont régis par l'annexe XXIV au décret n° 89-798 du 27 octobre 1989 et la circulaire

n° 89-17 du 30 octobre 1989. Ils regroupent les anciens IMP et IMPro. Les IME ont souvent été au départ des fondations caritatives, généralement à l'initiative de familles bourgeoises touchées par le handicap mental. Même s'ils sont désormais à financement quasi exclusivement public, après agrément par les DDASS, la grande majorité des IME restent à gestion associative. Ils sont différenciés par degrés de gravité de la déficience du public accueilli. La plupart disposent d'un internat, mais l'accueil en demi-pension est de plus en plus souvent pratiqué."

K.1.5. SAISP : Service d'Accompagnement à l'Insertion Sociale et Professionnelle

Chechez sur la toile !

K.1.6. CHRS : Les centres d'hébergement et de réinsertion sociale

Extrait de `http://www.adai13.asso.fr/fiches/log/log_chrs.htm`

"Les CHRS, Centres d'Hébergement et de Réinsertion Sociale", ont pour mission d'assurer l'accueil, l'hébergement, l'accompagnement et l'insertion sociale des personnes en recherche d'hébergement ou de logement, afin de leur permettre de retrouver une autonomie personnelle et sociale. Pour cela, elles bénéficient d'aide éducative et d'activités d'insertion professionnelles.

De part ces missions les CHRS interviennent dans différentes instances de décisions concernant les politiques sociales locales.

Les CHRS font aussi partie du "dispositif hivernal d'accueil d'urgence", c'est-à-dire que leurs capacités d'accueil d'urgence devraient être augmentées en période hivernale."

K.1.7. Indicateurs utilisés et principe du projet

Consultez le document `cb2005_arrete261004_annexe3.rtf`, issu de `http://www.actif-online.com/fichiers/texteLegislatif/cb2005_arrete261004_annexe3.rtf`, relatif à l'arrêté du 26 octobre 2004 sur les indicateurs.

Le principe de ce projet est de retrouver par vous-même une partie des taleaux résumant les indicateurs statistiques définis précédemment ; consulter la note d'information N° DGAS/5B/2006/102 du 6 mars 2006 relative aux valeurs moyennes pour 2003 des indicateurs fixé par l'arrêté du 26 octobre 2004 (voir document `DGAS5B2006102.pdf`, issu de `http://www.fhf.fr/file.php?tb=dos_article&at=id_article&px=fic1&id=1666`).

Vous aurez aussi à réaliser des analyses uni et bi-variées à partir des données.

K.2. Travail à fournir

K.2.1. Importation des données

Toutes les données se trouvent à l'URL habituelle.

(1) Consulter le fichier d'exemple `BiviersSessad_brutes.pdf` (il contient les données du SESSAD OVE à Biviers) et vérifier qu'il contient bien les indicateurs (définis dans la section K.1.7) à identifier :
   - A11-A115 : age
   - A1/A2/A : sexe
   - C1/C2 : durée du séjour
   - D1-D8 : temps actif mobilisable
   - ES11-ES99 : répartition de la population suivant leurs déficiences, hors ITEP
   - G11-G22 idem mais pour les ITEP
   - E1/E2 : heures de formation
   - F1/F2 : nombre de travailleurs handicapés
   - W : capacité autorisée
   - H1-H6 : qualification

- K1-L2 : compte administratif

(2) Importer avec R les données du fichier `indicIGAPA.txt`, qui contient pour 29 établissements du réseau OVE, les l'indicateurs de l'arrêté du 26 octobre 2004.

Attention, pour importer les données, vous procéderez comme indiqué dans la section D.2 ; cependant, il faut prendre au garde au fait que dans ce fichier, le séparateur utilisé est le point virgule ";". Ainsi,

- *Avec Rcmdr :*
  on procèdera comme dans la section D.2.1, mais il faudra spécifier dans "séparateur de champs", "Autre" et tapez point-virgule : ";"
  Finir ensuite comme d'habitude.
- *Sans Rcmdr :*
  Vous utiliserez la commande vue en section D.2.2, mais en spécifiant la valeur de l'argument optionnel 'sep' égal à ";" (par défaut il vaut l'espace " ") en tapant :
  ```
  indicIGAPA <- read.table("indicIGAPA.txt", h = T, sep = ";")
  ```

Visualisez-le pour constater qu'il contient les indicateurs précédents pour toutes les structures.

Vérifier qu'il contient bien les données du fichier `BiviersSessad_brutes.pdf`

K.2.2. Analyses univariées

(1) Analysez dans un premier temps
  (a) La capacité autorisée ('`W`') ;
  (b) Le nombre total d'individu ('`A`') ;
  (c) Le nombre de sorties ('`C1`') ;
  (d) Le nombre de jours cumulés des sortants ('`C2`') ;
  (e) Le nombre d'heures de formation réalisées ('`E1`') ;
  (f) Le département ('`departement`') ;

  Attention, pour le département, il faudra convertir cette donnée vue numérique par R en données factorielle ! Pour cela, on procédera ainsi :

  - *Avec Rcmdr :*
    Il faut aller dans "Données", puis "Gérer les variables...", puis "convertir var. numériques en facteur".
  - *Sans Rcmdr :*
    Il faudra créer une nouvelle variable '`departement`' en tapant
    ```
    departement <- as.factor(indicIGAPA$departement)
    ```

  (g) La catégorie ('`categorie`').

(2) Dans un second temps, vous en déduiserez l'analyse de :
  (a) La durée moyenne de séjours ('`C2/C1`') ;
  (b) Le pourcentage d'homme ('`100*A1/A`') ;
  (c) Le pourcentage de moins de 15 ans ('`100*(A11+A12+A13+A14+A15)/A`').

  Attention, il vous faudra créer de nouvelles variables ! Voir l'exercice E.2 page 88.

K.2.3. Analyses bivariées

(1) Relation entre le pourcentage de moins de 15 ans et la durée moyenne de séjours.

(2) Relation entre le département et la catégorie d'établissement

(3) Analyse diverses en fonction de la catégorie d'établissement

K.3. Quelques éléments de correction

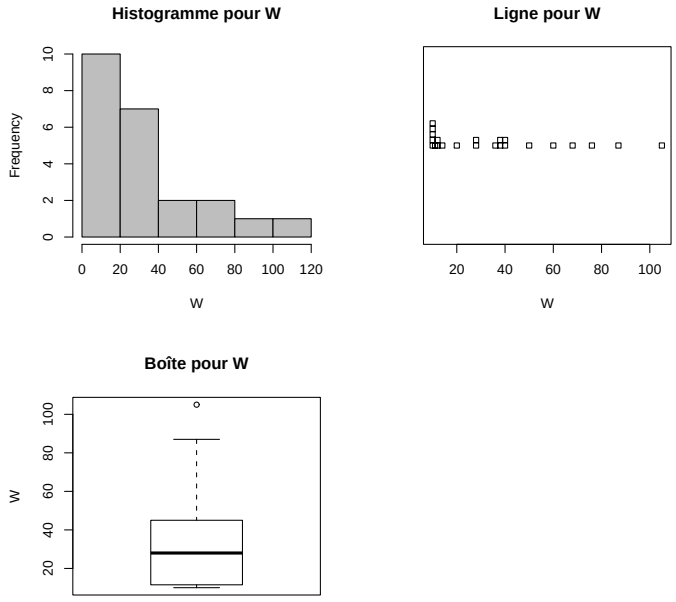
K.3.1. Analyses univariées

(1) (a) • On étudie la variable quantitative (ou numérique) 'W'. Pour les manipulations avec R, on renvoie donc aux sections E.2 et E.3 et aux sections récapitulatives J.1.1 et J.1.4 du document de cours.

• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	35.347826
écart-type	27.64484
Q_1 (quartile à 25 %)	11.5
médiane	28
Q_3 (quartile à 75 %)	45
minimum	10
maximum	105
nombre	29

•

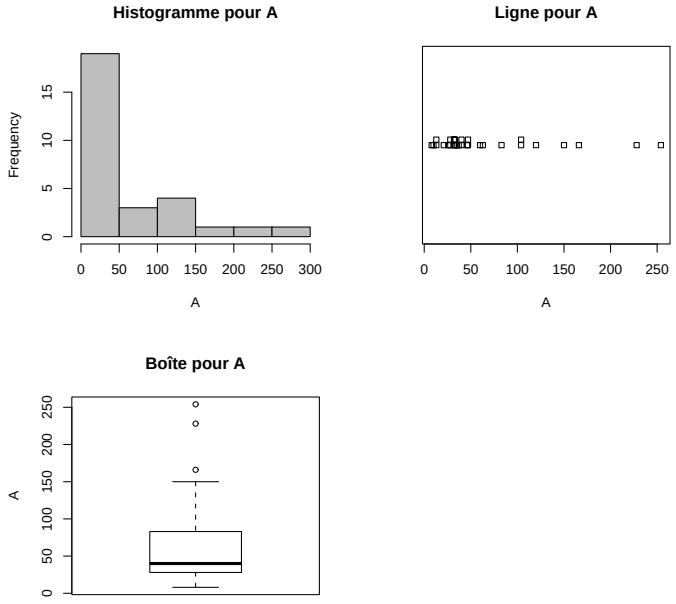


Voir les trois graphiques ci-dessus pour la variable 'W'.

(b) • On étudie la variable quantitative (ou numérique) '`A`'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	65.551724
écart-type	63.294429
Q_1 (quartile à 25 %)	28
médiane	40
Q_3 (quartile à 75 %)	83
minimum	8
maximum	254
nombre	29

•

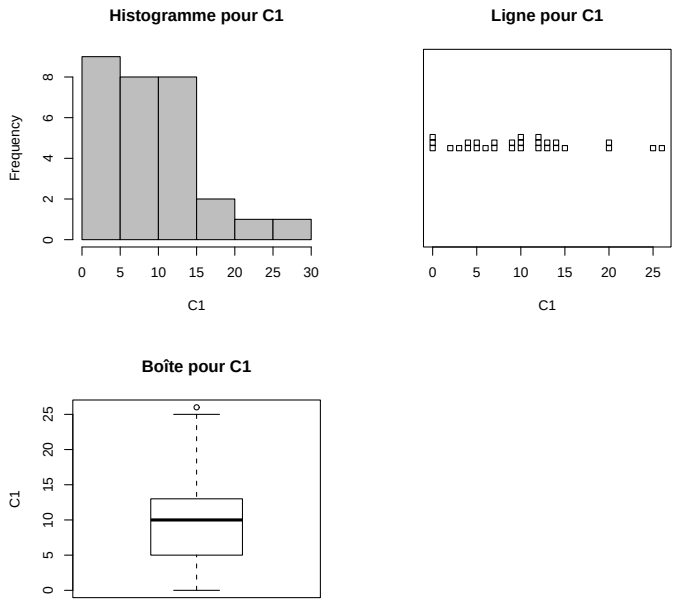


Voir les trois graphiques ci-dessus pour la variable '`A`'.

(c) • On étudie la variable quantitative (ou numérique) '`C1`'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	9.896552
écart-type	6.914503
Q_1 (quartile à 25 %)	5
médiane	10
Q_3 (quartile à 75 %)	13
minimum	0
maximum	26
nombre	29

•

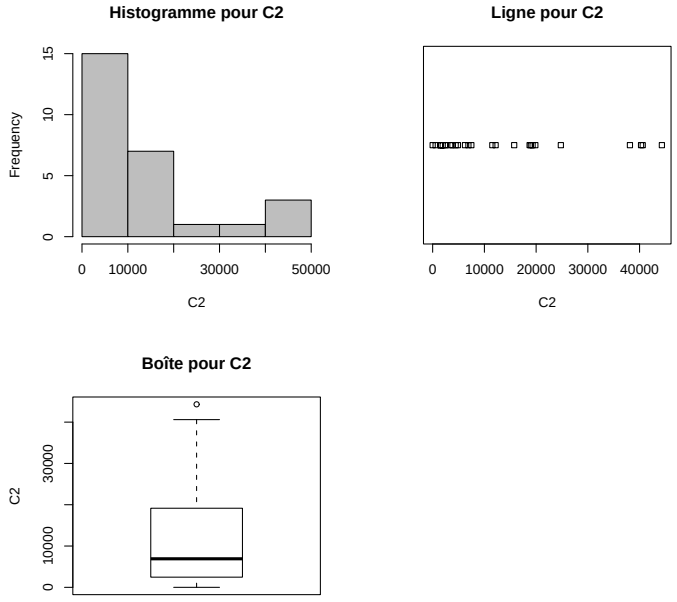


Voir les trois graphiques ci-dessus pour la variable '`C1`'.

(d) • On étudie la variable quantitative (ou numérique) 'C2'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	13104.962963
écart-type	13743.668032
Q_1 (quartile à 25 %)	2464
médiane	6910
Q_3 (quartile à 75 %)	19151.5
minimum	0
maximum	44311
nombre	29

•

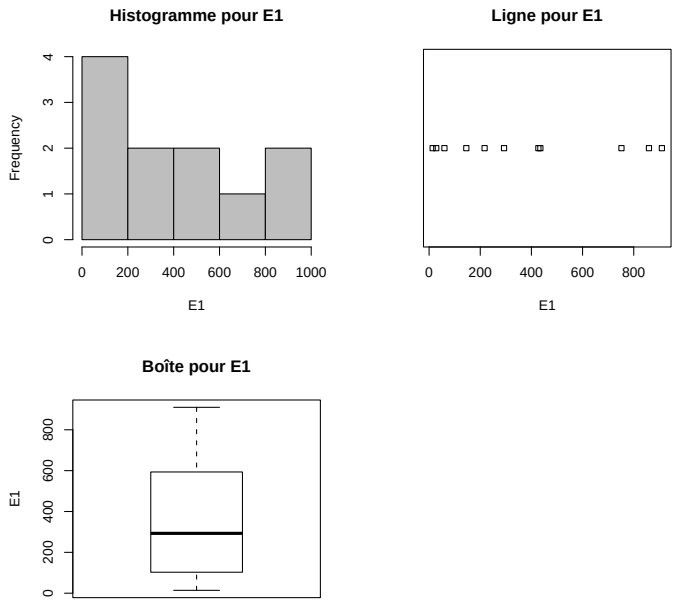


Voir les trois graphiques ci-dessus pour la variable 'C2'.

(e) • On étudie la variable quantitative (ou numérique) 'E1'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	376.479091
écart-type	332.225059
Q_1 (quartile à 25 %)	102.75
médiane	293
Q_3 (quartile à 75 %)	593.5
minimum	14
maximum	910.02
nombre	29

•

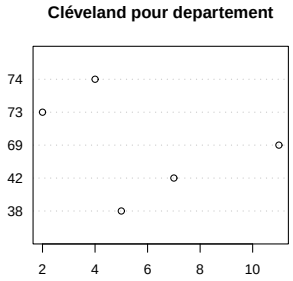


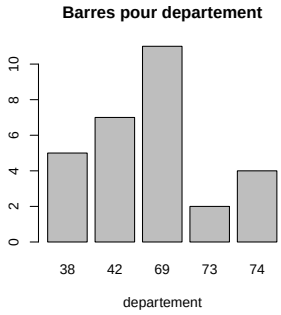
Voir les trois graphiques ci-dessus pour la variable 'E1'.

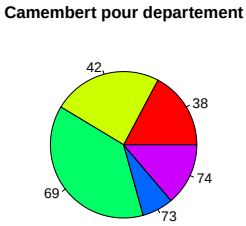
(f) • On étudie la variable qualitative (ou catégorielle) '`departement`'.
• Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
38	5	17.241
42	7	24.138
69	11	37.931
73	2	6.897
74	4	13.793

•





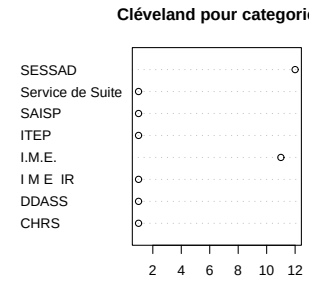


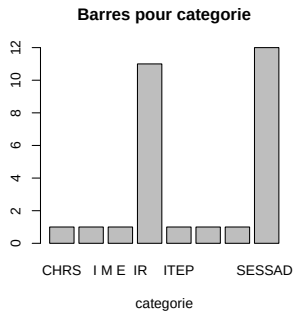
Voir les trois graphiques ci-dessus pour la variable '`departement`'.

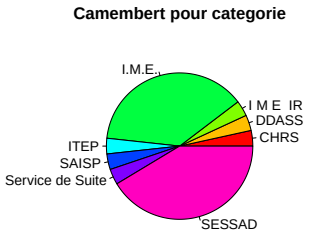
(g)
- On étudie la variable qualitative (ou catégorielle) '`categorie`'.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
CHRS	1	3.448
DDASS	1	3.448
I M E IR	1	3.448
I.M.E.	11	37.931
ITEP	1	3.448
SAISP	1	3.448
Service de Suite	1	3.448
SESSAD	12	41.379

-





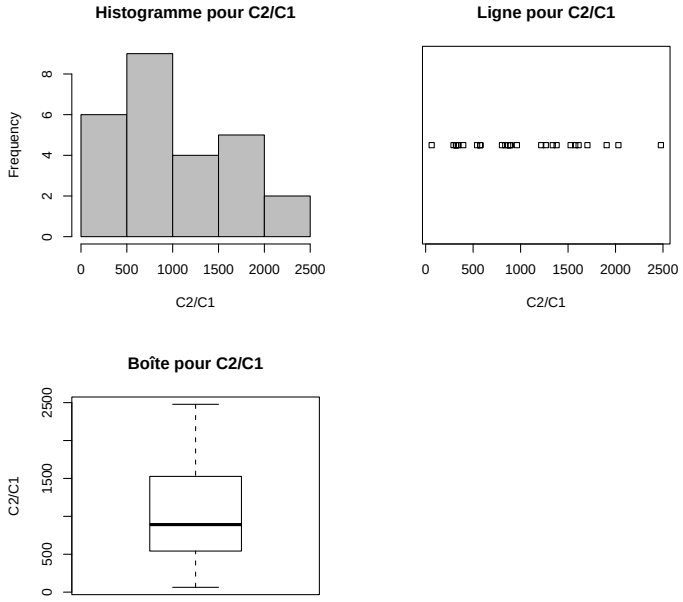


Voir les trois graphiques ci-dessus pour la variable '`categorie`'.

(2) (a) • On étudie la variable quantitative (ou numérique) '`C2/C1`'.
• Les différents résultats déterminés par sont donnés dans le tableau suivant

noms	valeurs
moyenne	1027.768537
écart-type	622.012872
Q_1 (quartile à 25 %)	550.24359
médiane	891.128571
Q_3 (quartile à 75 %)	1490.063187
minimum	63.333333
maximum	2477.7
nombre	29

•

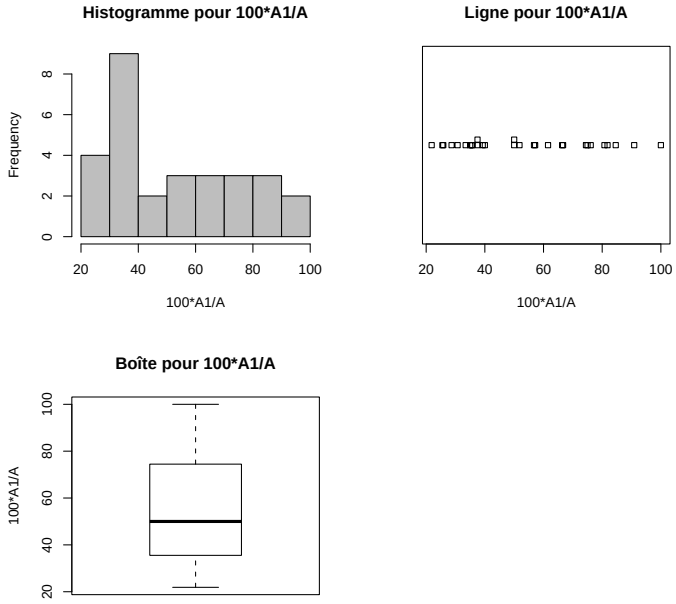


Voir les trois graphiques ci-dessus pour la variable '`C2/C1`'.

(b) • On étudie la variable quantitative (ou numérique) '`100*A1/A`'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	53.467095
écart-type	22.28414
Q_1 (quartile à 25 %)	35.526316
médiane	50
Q_3 (quartile à 75 %)	74.468085
minimum	21.875
maximum	100
nombre	29

•

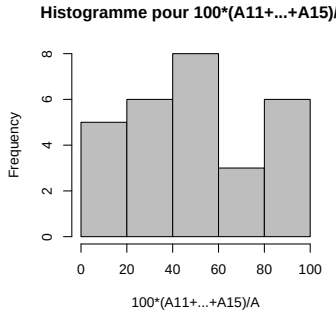


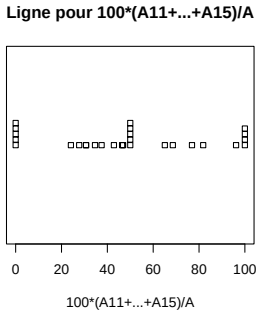
Voir les trois graphiques ci-dessus pour la variable '`100*A1/A`'.

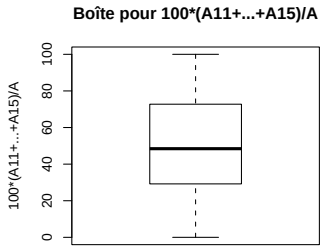
(c) • On étudie la variable quantitative (ou numérique) '`100*(A11+...+A15)/A`'.
• Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	48.572844
écart-type	32.920522
Q_1 (quartile à 25 %)	29.954027
médiane	48.4375
Q_3 (quartile à 75 %)	70.659341
minimum	0
maximum	100
nombre	29

•



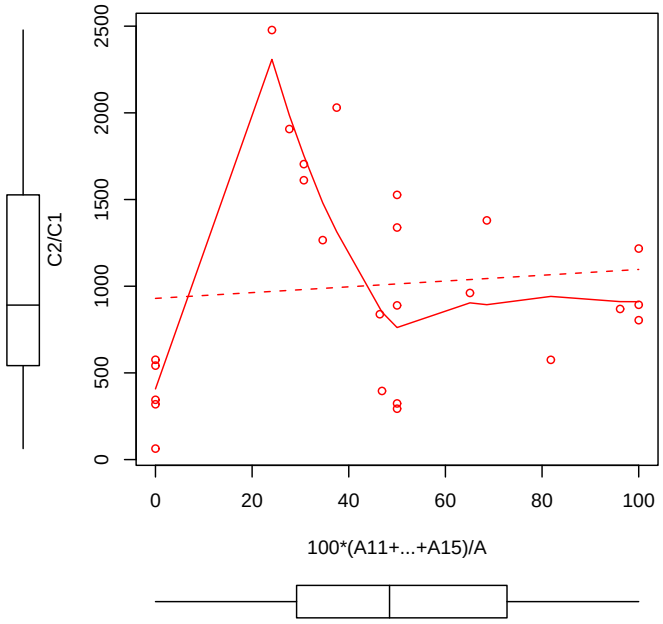




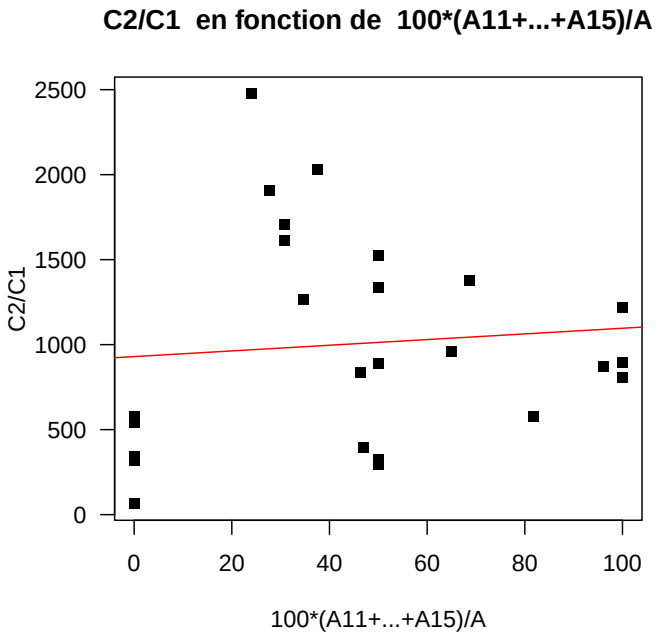
Voir les trois graphiques ci-dessus pour la variable '`100*(A11+...+A15)/A`'.

K.3.2. Analyses bivariées

(1)
- On étudie le croisement de la variable quantitative (ou numérique) '`100*(A11+...+A15)/A`' et de la variable quantitative (ou numérique) '`C2/C1`'. Pour les manipulations avec R, on renvoie donc à la section G.5 du document de cours.
- Voir la figure ci-dessous.
  - *Avec Rcmdr :*



  - *Sans Rcmdr :*



Sur cette figure, les points semblent peu alignés.

- Confirmons cela grâce à R.
  Les résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
pente a	1.668168
ordonnée à l'origine b	929.753528
corrélation linéaire r	0.087175
probabilité critique p_c	0.678617

  On compare la valeur absolue de la corrélation linéaire r =0.087175 aux seuils de Cohen (0.1,0.3,0.5) (voir [**13**]) et la probabilité critique p_c=0.678617 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	**faible**
significativité statistique	**non**

- On peut donc affirmer il existe une relation faible entre les variables '`100*(A11+...+A15)/A`' et '`C2/C1`'.

(2)
- On étudie le croisement de la variable qualitative (ou catégorielle) '`departement`' et de la variable qualitative (ou catégorielle) '`categorie`'. Pour les manipulations avec R, on renvoie donc à la section H.5 du document de cours.
- La table de contingence déterminée par R est donnée dans le tableau suivant

	CHRS	DDASS	I M E IR	I.M.E.	ITEP	SAISP	Service de Suite	SESSAD
38	0	0	0	1	1	1	0	2
42	0	0	0	4	0	0	0	3
69	1	1	1	4	0	0	1	3
73	0	0	0	0	0	0	0	2
74	0	0	0	2	0	0	0	2

Les autres résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
χ^2	20.437525
coefficient de Cramer V	0.419745
taille d'effet w	0.839489
probabilité critique p_c	0.847997

On compare la taille d'effet w=0.839489 aux seuils de Cohen (0.1,0.3,0.5) (voir [**13**]) et la probabilité critique p_c=0.847997 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	**très forte**
significativité statistique	**non**

- On peut donc affirmer qu'il existe une relation entre les variables 'departement' et 'categorie', cepandant l'absence de significativité statistique (probabelement due aux faibles nombres d'établissement nous invite à nous méfier de cette relation.

ANNEXE L

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 L.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 correlations linéaires et les probabilités critiques.

(3) Conclure.

Corrigé

Éléments de correction de l'exercice L.1
Cet exemple pédagogique a été mis au point par Anscombe [**16**] 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 R, on renvoie donc à la section G.5 du document de cours.

On a indiqué en figures L.1 page 176 et L.2 page 177, les quatres nuages de points et les droites de régression linéaire (obtenues avec Rcmdr). On a indiqué en figure L.3 page 178 et L.4 page 179, 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 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 [**13**]) 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	**très forte**
significativité statistique	**oui**

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 L.2. Les données étudiées ici, créées de façon pédagogiques par Anscombe, sont en fait déjà présentes dans R ! Il suffit de taper dans 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 crées 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 figures L.5 page 180 et L.6 page 181, les quatres nuages de points et les droites de régression linéaire (obtenues avec Rcmdr). On a indiqué en figure L.7 page 182 et L.8 page 183, 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`

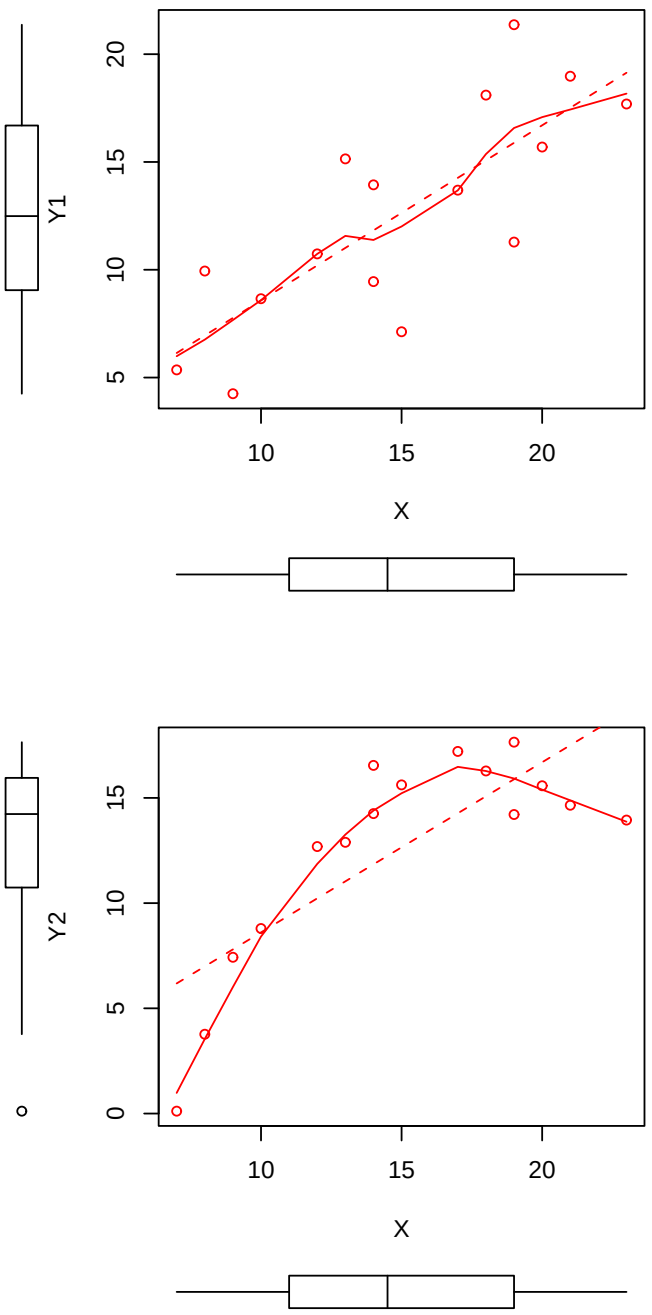


FIGURE L.1. Les deux premiers nuages de points et les droites de régression linéaire (avec Rcmdr).

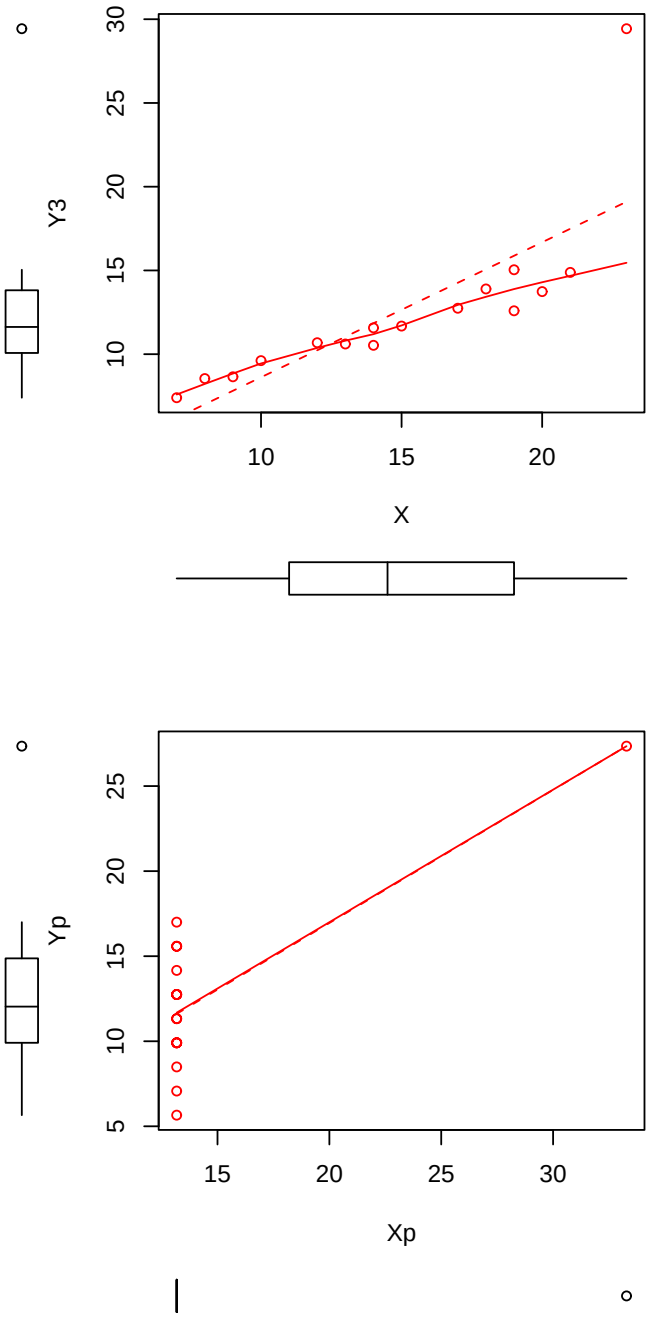


FIGURE L.2. Les deux derniers nuages de points et les droites de régression linéaire (avec Rcmdr).

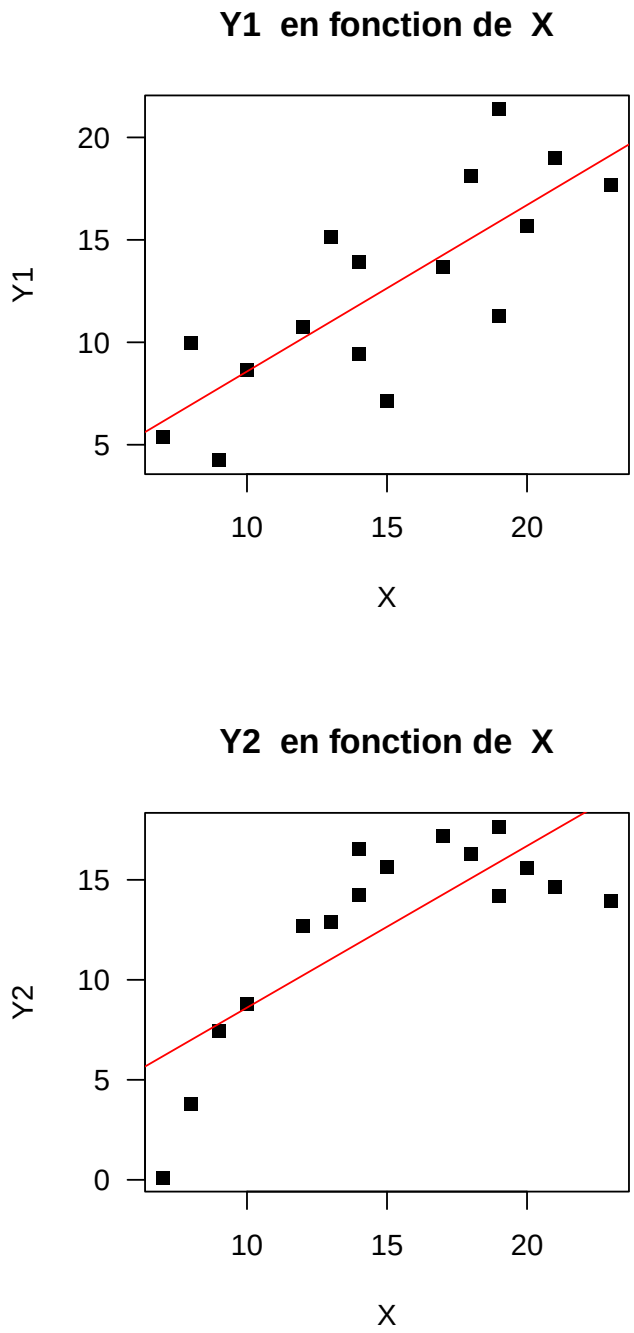


FIGURE L.3. Les deux premiers nuages de points et les droites de régression linéaire.

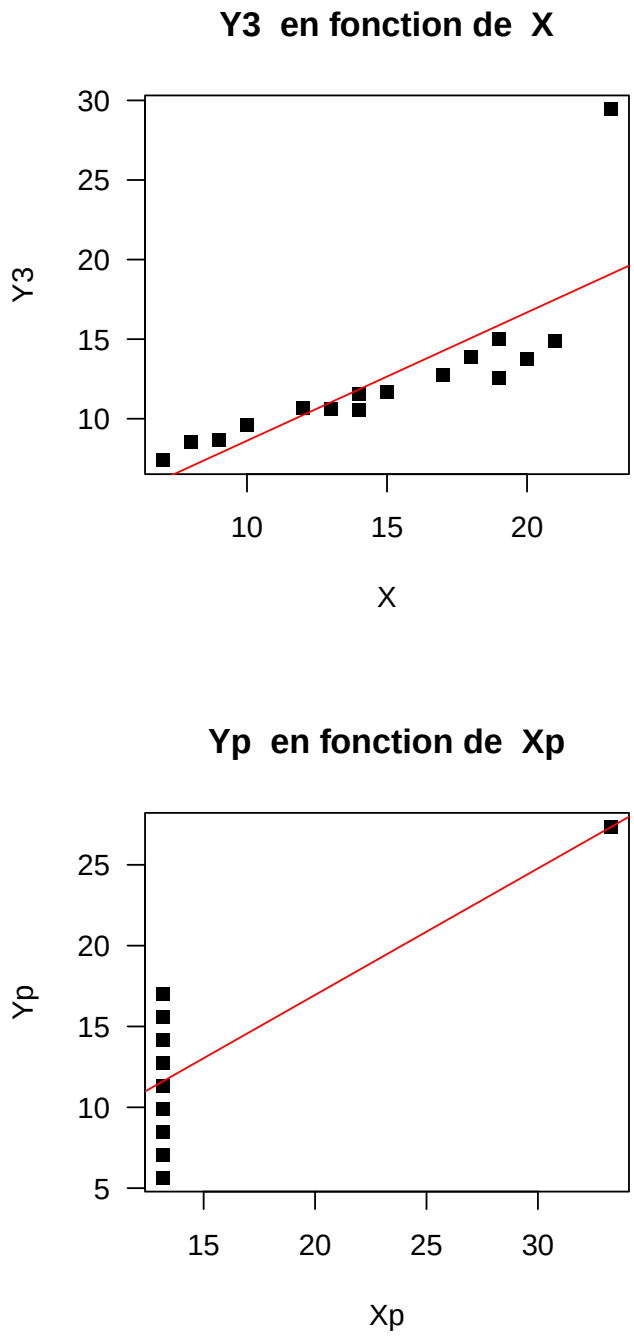


FIGURE L.4. Les deux derniers nuages de points et les droites de régression linéaire.

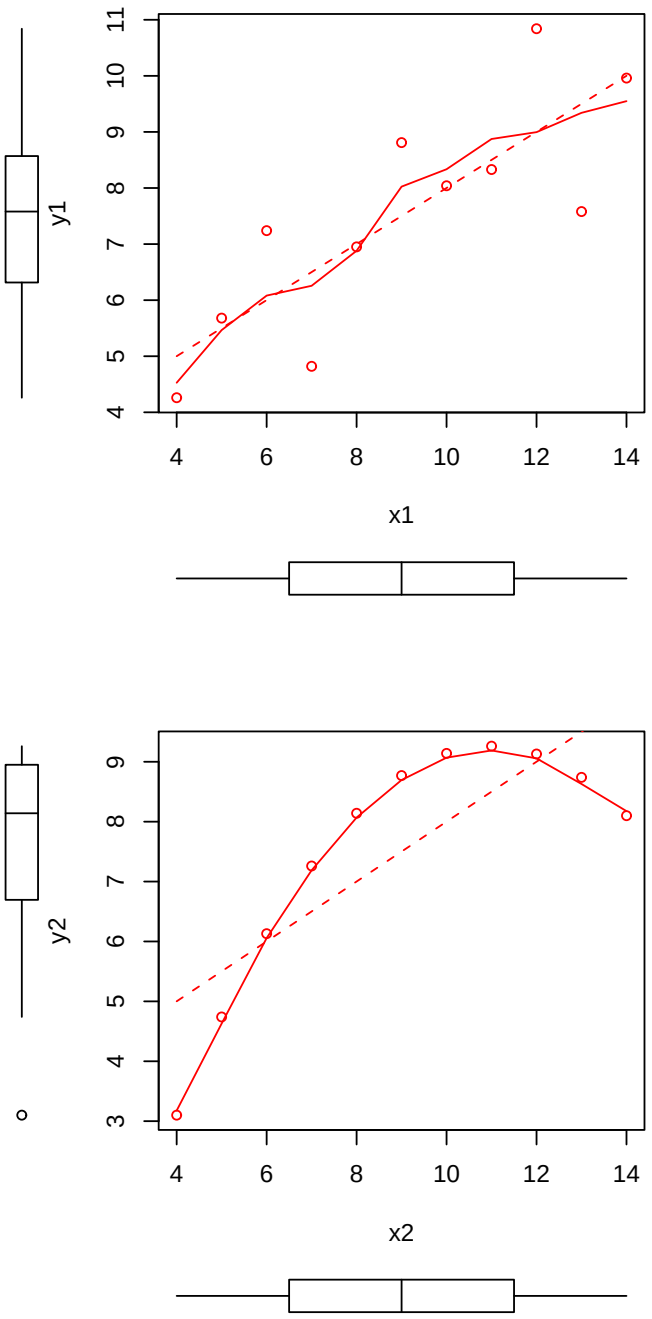


FIGURE L.5. Les deux premiers nuages de points et les droites de régression linéaire (avec Rcmdr) pour les données '`anscombe`'.

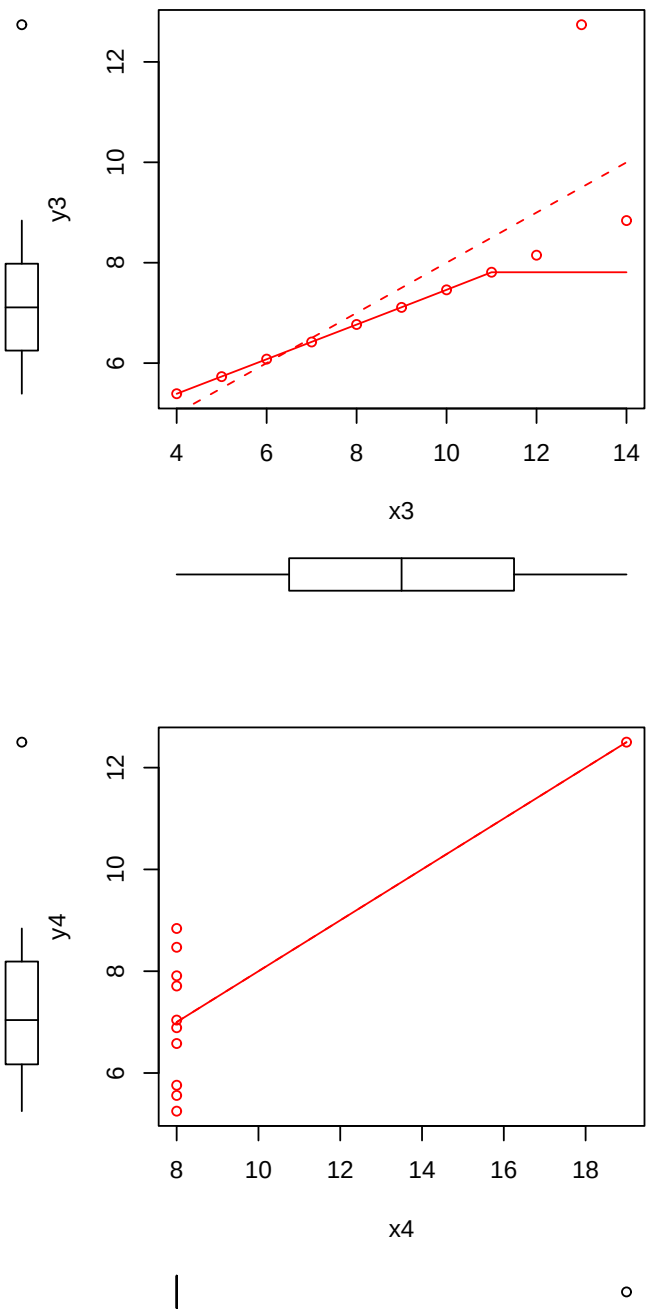


FIGURE L.6. Les deux derniers nuages de points et les droites de régression linéaire (avec Rcmdr) pour les données '`anscombe`'.

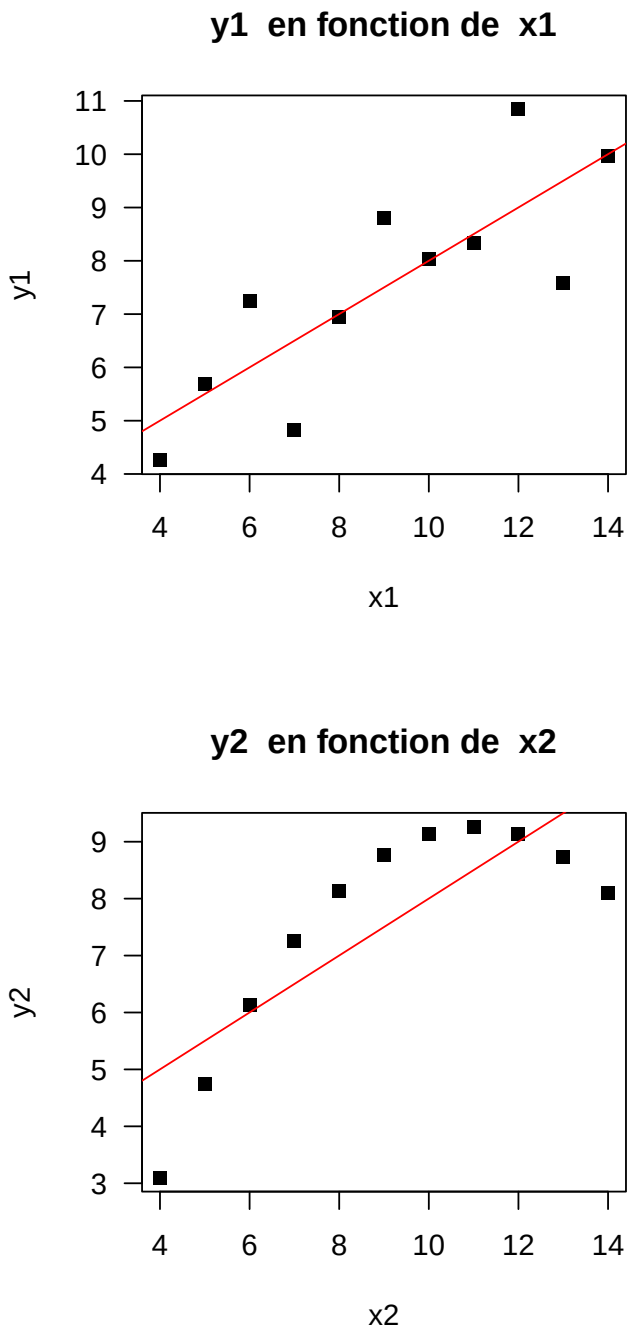


FIGURE L.7. Les deux premiers nuages de points et les droites de régression linéaire pour les données 'anscombe'.

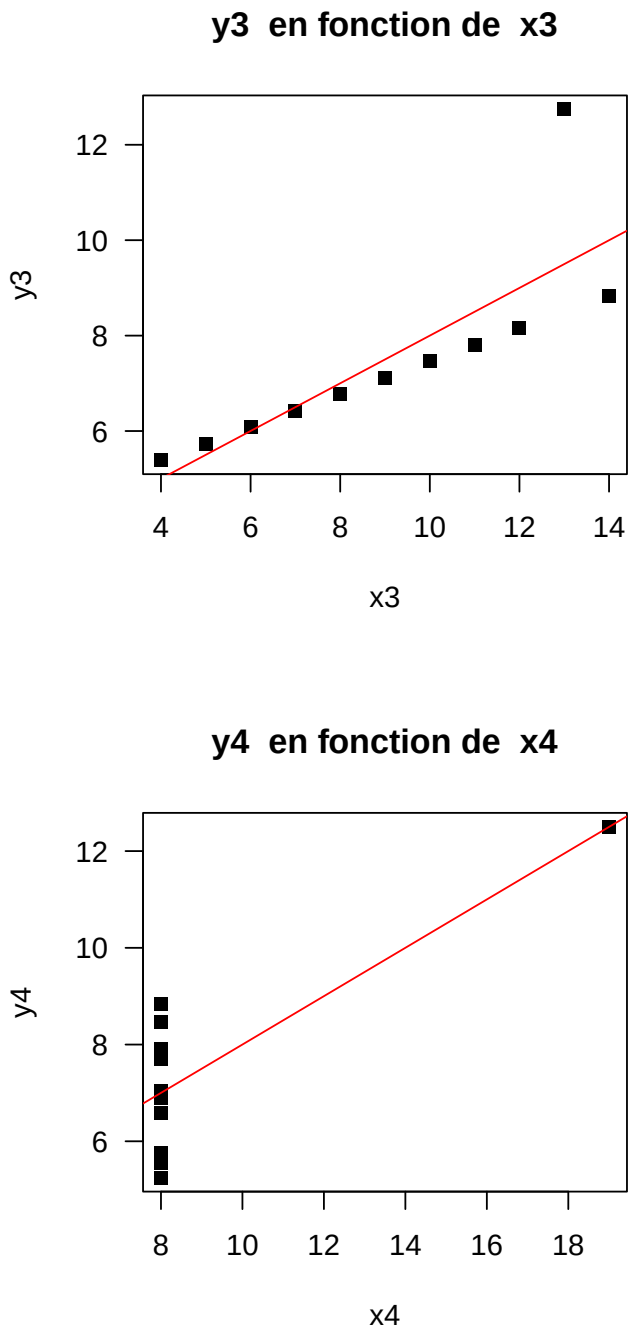


FIGURE L.8. Les deux derniers nuages de points et les droites de régression linéaire pour les données 'anscombe'.

ANNEXE M

Utilisation de fonctions avec R

Ici, la notion de fonction est introduite pour vous simpflifier vos démarches avec R, mais l'usage de ces fonctions n'est nullement imposé (sauf éventuellement pour faire tourner des démonstrations) !

M.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 <- function(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 diférentes 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 R, en allant dans "fichier", "Nouveau script", puis une fois le texte collé, faite "sauver".

Remarque M.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
```

M.2. Une fonction à deux valeurs de sortie

Considérons maintenant la fonction `somme_rapport.R`, disponible à l'URL habituelle. Comme indiqué dans la setion M.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
```

M.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 R, [**17**] disponible sur internet.

ANNEXE N

Vérification expérimentale de la loi des grands nombres et statistique inférentielle

N.1. La loi des grands nombres

La loi des grands nombres est un résultat très connu en probabilité et statistique et peut prendre beaucoup de formes différentes, selon le niveau théorique avec lequel on les étudie !

Cette loi des grands nombres a été utilisée en fait, au cours du chapitre 3, pour la définition 3.4 page 6 des probabilités. On pourra par exemple consulter le chapitre 3 de [**7**] ou la page 174 de [**9**].

D'autres définitions peuvent être données : sur Wikipédia, à l'url :
`http://fr.wikipedia.org/wiki/Loi_des_grands_nombres`, on peut lire par exemple : "En statistiques, la loi des grands nombres indique que lorsque l'on fait un tirage aléatoire dans une série de grande taille, plus on augmente la taille de l'échantillon, plus les caractéristiques statistiques de l'échantillon se rapprochent des caractéristiques statistiques de la population."

Les implication de cette loi peuvent s'exprimer de deux façon selon que la variable aléatoire étudiée est discrète ou continue : on suppose que l'on réalise un "grand" échantillon à partir de la variable aléatoire X.

- Si la variable aléatoire X est discrète, chaque probabilité $P(X = x)$ doit se rapprocher de la fréquence (ou de la proportion de fois) où les valeurs de l'échantillon sont égales à x.
- Si la variable aléatoire est continue, l'histogramme en fréquence de cet échantillon variable doit se rapprocher de la densité de probabilité de la loi initiale, pourvu que chacune des classes soit assez petite.

Pseudo-preuve. On peut montrer, sans rigueur mathématique, le second point.
On renvoie à la figure N.1 page suivante. On suppose que les classes de l'histogramme sont :

$$[x_1, x_2[$$
$$[x_2, x_3[$$
$$\dots$$
$$[x_i, x_{i+1}[$$
$$\dots$$
$$[x_n, x_{n+1}]$$

et qu'elles sont toutes de largeur dl, "petite".

D'autres part, pour la classe $[x_i, x_{i+1}]$, l'ordonnée de l'histogramme est $q(x_i)$. On suppose que $p(x_i)$ est la valeur de la densité de probabilité au point x_i. Il s'agit de montrer que

$$p(x_i) \approx q(x_i). \tag{N.1}$$

Par définition, la probabilité que X appartiennent à l'intervalle $[x_i, x_{i+1}]$ est :

$$P(x_i \leq X \leq x_{i+1}) = \int_{x_i}^{x_{i+1}} p(x)dx.$$

Puisque dl est "petit", on peut supposer que p ne varie pas sur cet intervalle et qu'il y vaut $p(x_i)$; on a donc

$$P(x_i \leq X \leq x_{i+1}) \approx p(x_i)(x_{i+1} - x_i) = p(x_i)dl. \tag{N.2}$$

Par ailleurs, notons $N(x_i)$ le nombre d'éléments de l'échantillon appartenant à l'intervalle $[x_i, x_{i+1}[$ et N le nombre total d'éléments de l'échantillon. Par définition, la hauteur de l'histogramme sur la classe $[x_i, x_{i+1}[$ est proportionnelle à $N(x_i)$. Il existe donc une constante α telle que

$$q(x_i) = \alpha N(x_i) \tag{N.3}$$

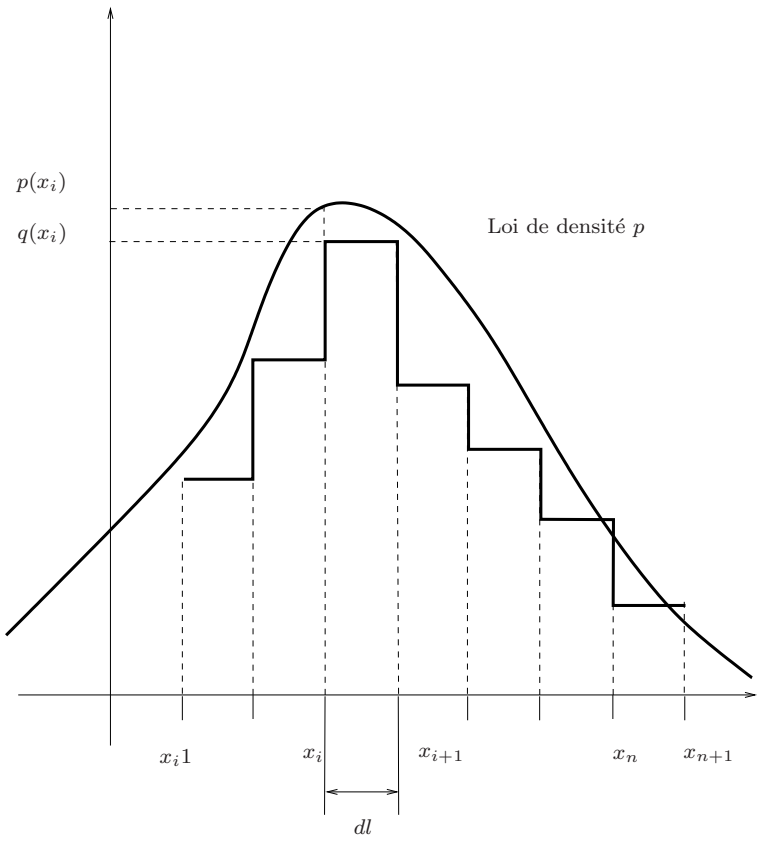


FIGURE N.1. Approximation de la loi de probabilité par l'histogramme

On multiplie cette égalité par dl et on somme sur i :

$$\sum_{i=1}^{n} q(x_i)dl = \sum_{i=1}^{n} \alpha N(x_i)dl = \alpha dl \sum_{i=1}^{n} N(x_i).$$

Le terme de gauche est égale à la somme totale de l'aire des colonnes de l'histogramme (égale à 1 par définition). Le terme de droite vaut $\alpha N dl$ et on a donc

$$1 = \alpha N dl$$

Si on repporte cette valeur de α dans (N.3), on a donc

$$q(x_i) = \alpha N(x_i) = \frac{N(x_i)}{N dl} \tag{N.4}$$

Enfin, la probabilité $P(x_i \leq X \leq x_{i+1})$ est approximativement égale à la proportion de fois où la valeur de l'échantillon est dans l'intervalle $[x_i, x_{i+1}[$:

$$P(x_i \leq X \leq x_{i+1}) = \frac{N(x_i)}{N}$$

et donc (N.4) fournit

$$q(x_i) = \frac{1}{dl} P(x_i \leq X \leq x_{i+1}).$$

Enfin, on déduit alors de (N.2) :

$$p(x_i)dl \approx P(x_i \leq X \leq x_{i+1}) = q(x_i)dl.$$

et donc (N.1) est vrai. □

♢

N.2. "Simulation"

Reprenons les trois lois vues au cours du chapitre 3
- loi uniforme discrète, définition 3.7 page 7,
- loi binomiale, définition 3.15 page 13 ;
- loi normale, définition 3.34 page 17.

Nous proposons de faire des simulations de ces trois lois qui corroborent expérimentalement les conséquences de la loi des grands nombres.

En fait, ce n'est pas exactement cela que l'on teste ; on vérifie tout simplement que les échantillons produits aléatoirement par ® sont bien conformes aux lois théoriques, autrement dit que l'éléat est bien programmé ! ♢

N.2.1. Loi uniforme discrète

Le dès correspond à $n = 6$.

Nous avons déjà vu, sur la figure 3.1 page 7 une simulation faite avec ®. Sur cette figure, on peut voir six courbes, correspondant aux proportions expérimentales d'apparition de chacun des numéros, qui se rapprochent de 1/6.

Comme dans l'annexe M, vous pouvez télécharger la fonction `graphe.desimple` et taper par exemple

```
graphe.desimple(10000)
```

ce qui produira une figure identique à la figure 3.1 page 7.

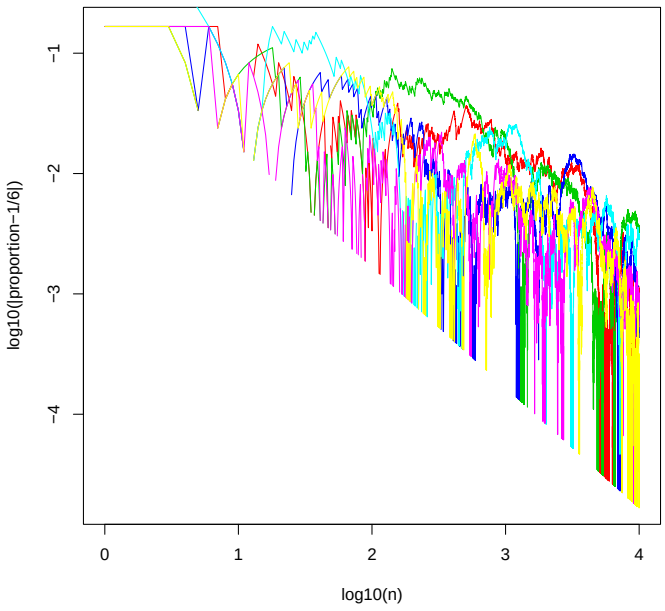


FIGURE N.2. Une simulation de la fonction `graphe.desimple` pour $n = 10000$ en logarithme.

Pour ceux qui sont habitués aux logarithmes, consulter la fonction `graphe.desimple.R` et taper par exemple

```
graphe.desimple(10000,translog=TRUE)
```

Vous obtiendrez la figure du type de la figure N.2.

N.2.2. Loi binomiale

Comme dans l'annexe M, récupérer et sourcer la fonction `verifie.binom.R`.

Cette fonction simule le modèle binomial, pour un nombre de tirage égal à q, avec des paramètres (n, p). Elle affiche le graphe de la loi de probabilité théorique ainsi que celui des proportions observées.

Constater en reprenant les paramètres suivants $p = 0.3$ et $n = 5$ et q de plus en plus grand que les deux graphes sont similaires en tapant par exemple :

```
verifie.binom(5,0.3,q=1000)
```

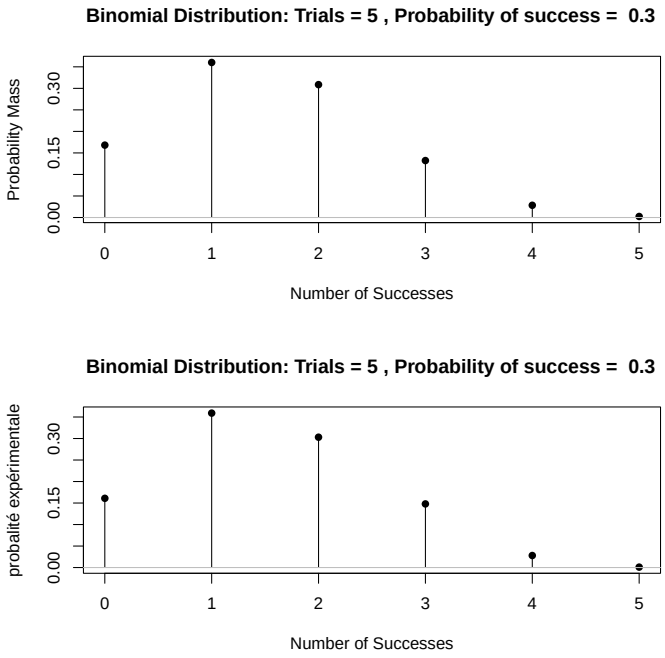


FIGURE N.3. Le résultat de la fonction `verifie.binom` pour $n = 5$, $p = 0.3$ et $q = 1000$.

Voir par exemple la figure N.3.

On pourra aussi afficher les différences entre les probabilités théoriques et les proportions observées en tapant

```
binom.complet(5,1000,q=1000,difference=T)
```

N.2.3. Loi normale

On peut reprendre maintenant la simulation donnée page 8 dans le chapitre 3. Elle consistait à tracer l'histogramme d'un grand échantillon issu d'une loi normale et de constater que cet histogramme en densité (voir figure 3.4 page 10) est proche d'une courbe normale.

On peut être un peu plus précis : on se donne

- une moyenne μ ;
- un écart-type σ ;
- la taille de l'échantillon n ;
- un nombre de classe q.

On réalise un échantillon de taille n, issu d'une loi normale de moyenne μ et d'écart-type σ, on trace un histogramme en densité avec q classe et on constate que cet histogramme est proche de la densité de probabilité de la loi normale de moyenne m et d'écart-type σ.

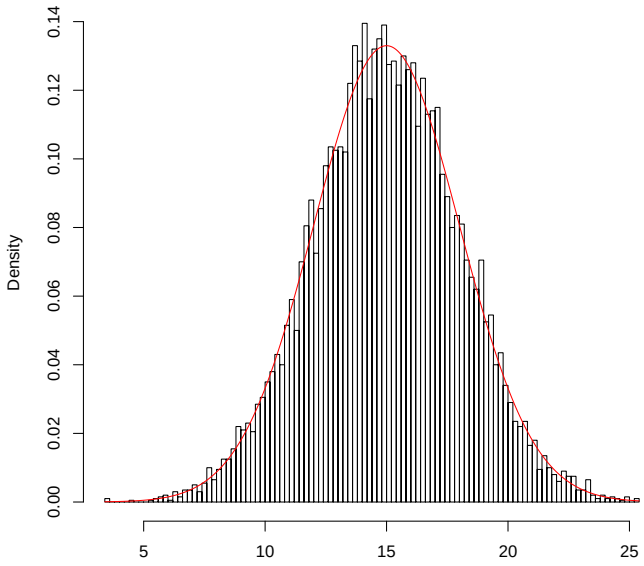


FIGURE N.4. Le résultat de la fonction `verifie.norm` pour $\mu = 15$, $\sigma = 3$, $n = 10000$ et $q = 150$.

Pour cela, il suffit d'utiliser la focntion `verifie.norm`. Taper par exemple

```
verifie.norm(mu=15,sigma=3,10000,150)
```

ce qui produit une figure analogue à la figure N.4.

On peut faire aussi le lien entre la définition 3.10 page 10 et la proposition 3.39 page 21 et calculant la moyenne et l'écart-type de l'échantillon produit et en vérifiant justement qu'ils sont proches de μ et de σ, en tapant par exemple

```
verifie.norm(mu=15,sigma=3,10000,150,echo=T)
```

ce qui donne

```
$mue
[1] 14.93825

$sigmae
[1] 3.007661
```

Attention, la loi des grands nombres est un résultat en moyenne ; il se peut (mais c'est un phénomème très rare) que vous ayez des partie de l'histogramme qui soient éloignées de la densité de probabilité. ♢

N.3. La statistique inférentielle

On suppose qu'une variable aléatoire X provient d'une loi de probabilité (discrète ou continue). Cette loi aléatoire est déterminée par quelques paramètres (par exemple n et p pour la loi Binomiale ou μ et σ pour la loi normale).

Le domaine des probabilités consiste à prévoir les fréquences apparitions des événements à partir de la loi de probabilité. Les statistiques descriptives permettent de vérifier, sur des échantillons aléatoires formés à partir de cette loi (ou des réalisations de la variable aléatoire X), que les statistiques de ces échantillons sont liées à cette loi : par exemple, on a vu que la moyenne et l'écart-type d'un échantillon issu d'une loi normale sont proches de la moyenne et de l'écart-type de la loi normale.

La statistique inférientielle a pour objectif l'inverse : on suppose une ou plusieurs réalisations d'une variable aléatoire connues (sous forme d'échantillons). On en connaît des statistiques (par exemple moyenne et écart-type). On cherche, d'abord, sur le plan théorique, une loi de probabilité dont pourrait venir le ou les échantillons. On cherche ensuite à inférer, c'est-à-dire déduire, les valeurs des paramètres de la loi de probabilité à partir des statistiques connues. Naturellement, l'échantillon est supposé être issu de la loi de probabilité de façon aléatoire et la connaissance des valeurs des paramètres ne sera qu'imparfaite !

Donnons deux exemples du domaine de la statistique inférentielle, que nous developperons lors des chapitres suivants.

(1) On se donne un groupe d'élèves dont on connait la taille. On fait l'hypothèse que ce groupe provient d'une loi normale de moyenne μ et d'écart-type σ inconnus. On montrera qu'il est possible d'approcher μ à partir de m et de s les moyennes et écart-type du groupe d'élèves.

(2) Un autre exemple, qui sera utilisé, lors de la théorie des sondages (très simplifiée) : on suppose que l'on connaisse, lors d'une élection avec seulement deux issues possibles, le choix d'un petit nombre d'électeurs. Il s'agira d'en déduire le choix global de l'ensemble de la population électrice.

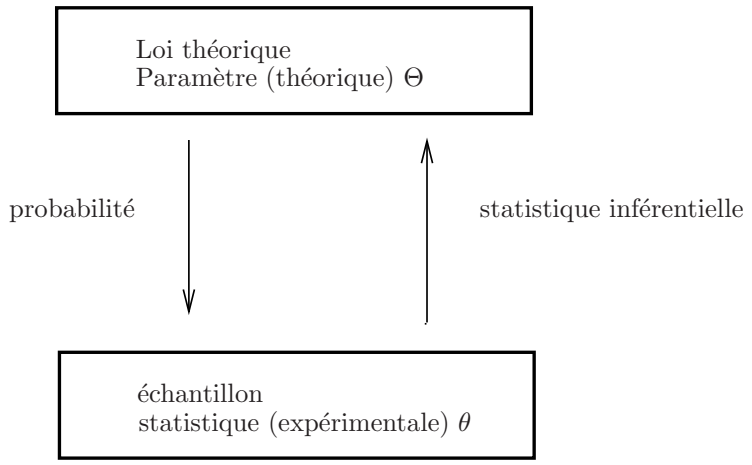


FIGURE N.5. Le principe des probabilités et de l'inférence

Voir la figure N.5 qui résume tout cela.

Pour simplifier, le domaine de la probabilité donne la valeur de la statistique θ en fonction du paramètre Θ, tandis que la statistique inférentielle fournit la valeur du du paramètre Θ en fonction de la statistique θ.

ANNEXE O

Lien entre la moyenne et l'écart-type d'une variable aléatoire et la moyenne et l'écart-type des valeurs prises par cette variable aléatoire au cours expérience

Montrons l'équivalence des définitions 3.10 page 10 et 3.12 page 10. On rappelle que l'on effectue N tirages aléatoires d'une variable aléatoire X et on note $(y_i)_{1\leq i\leq N}$ les valeurs obtenues. Notons $(n_j)_{1\leq j\leq q}$ l'ensemble des valeurs prises par la variable aléatoire X. Notons, pour chaque $j \in \{1, \ldots, q\}$, α_j, le nombre de fois où est apparue la valeur n_j parmi les valeurs $(y_i)_{1\leq i\leq N}$. Ainsi, la moyenne des valeurs $(y_i)_{1\leq i\leq N}$ est égale à

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} y_i &= \frac{1}{N}\left(\alpha_1 n_1 + \alpha_2 n_2 + \ldots + \alpha_q n_q\right),\\ &= \frac{\alpha_1}{N} n_1 + \ldots + \frac{\alpha_q}{N} n_q\end{aligned}$$

Si N est "grand", alors,

$$\begin{aligned}\frac{\alpha_1}{N} &\approx p(X = n_1),\\ \frac{\alpha_2}{N} &\approx p(X = n_2),\\ &\vdots\\ \frac{\alpha_q}{N} &\approx p(X = n_q),\end{aligned}$$

et donc

$$\frac{1}{N}\sum_{i=1}^{N} y_i \approx p(X = n_1)n_1 + \ldots + p(X = n_q)n_q$$

ce qui est exactement la valeur de l'espérance $\mathbb{E}(X)$. On ferait de même pour l'écart-type.

ANNEXE P

Preuve de la proposition 3.23

Calculons tout d'abord l'espérance de la variable aléatoire binomiale de paramètres n et p égale par définition et d'après la proposition 3.17 et la formule du binome

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{n} kP(X=k)\\ &= \sum_{k=0}^{n} kC_n^k p^k (1-p)^{n-k}\end{aligned}$$

Pour calculer cela, on introduit la fonction

$$f(X) = (X+1-p)^n = \sum_{k=0}^{n} C_n^k X^k (1-p)^{n-k}. \tag{P.1}$$

Dérivons ces égalités termes à termes (par rapport à X)

$$f'(X) = ((X+1-p)^n)' = \left(\sum_{k=0}^{n} C_n^k X^k (1-p)^{n-k}\right)',$$

et donc

$$n(X+1-p)^{n-1} = \left(\sum_{k=0}^{n} C_n^k X^k (1-p)^{n-k}\right)'$$

soit encore

$$\begin{aligned}n(X+1-p)^{n-1} &= \left(C_n^0 (1-p)^n\right)' + \left(\sum_{k=1}^{n} C_n^k X^k (1-p)^{n-k}\right)'\\ &= \sum_{k=1}^{n} C_n^k k X^{k-1} (1-p)^{n-k}.\end{aligned}$$

et donc

$$\begin{aligned}nX(X+1-p)^{n-1} &= X\left(\sum_{k=0}^{n} C_n^k k X^{k-1} (1-p)^{n-k}\right)\\ &= \sum_{k=1}^{n} C_n^k k X^k (1-p)^{n-k}\\ &= \sum_{k=0}^{n} C_n^k k X^k (1-p)^{n-k}\end{aligned}$$

En remplaçant X par p, on a donc finalement

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{n} kC_n^k p^k (1-p)^{n-k}\\ &= np(p+1-p)^{n-1}\\ &= np\end{aligned}$$

ce qui est bien la formule annoncée.

Pour calculer l'écart-type et la variance, on ferait de même. $\diamondsuit$

ANNEXE Q

Passage d'une loi de probabilité discrète à une loi de probabilité continue

Q.1. Une manipulation sur la loi binomiale

EXERCICE Q.1. Observer comment évolue le graphe de la distribution binomiale évolue lorsque vous conservez $p = 0.3$ et que vous choisissez $n \in \{5, 20, 100, 150, 200, 400\}$.

Voir éléments de correction page 202

Reprenons maintenant le très bon exemple issu de la page 7 de [**18**]. Reprenons loi binomiale de paramètres n et p et modifions-la pour que son espérance ($\mu = np$) soit égale à 0 et son écart-type ($\sigma = \sqrt{np(1-p)}$) égal à 1. On peut montrer qu'il suffit de garder la même loi (définie par (3.8) page 13) et de remplacer chacune des valeurs de succès $k \in \{0, \ldots, n\}$ par $(k - \mu)/\sigma$ pour $k \in \{0, \ldots, n\}$. Autrement dit, on a

$$\forall x \in \{0, \ldots, n\}, \quad P\left(X = \frac{x - \mu}{\sigma}\right) = C_n^x p^x (1-p)^{n-x} \tag{Q.1}$$

On dit que l'on obtient la loi binomiale normalisée (d'espérance nulle et d'écart-type égal à 1).

Sur le graphique Q.1 page suivante, on a indiqué le tracé de la loi binomiale normalisée sur l'intervalle $[-3, 3]$ $p = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) la "loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$", qui est la "courbe en cloche idéale" et qui sera définie dans la section 3.4.3.

Sur la figure suivante Q.2 page 201, on a tracé les mêmes éléments en "normant" les probablilités de telle sorte que la valeur maximale atteinte soit égale à la valeur maximale de la "courbe en cloche" idéale (en 0 : $1/\sqrt{2p}$, ici le "vrai" p).

On constate alors que "la courbe des probabilités discrètes" ont l'air de "se rapprocher" d'une "courbe continue" quand n augmente. Cette courbe est la "courbe en cloche idéale".

Le nombre de valeurs possibles de la loi de probabilité discrète tend vers l'infini, chacune des probabilités tend vers 0, mais ce qui devient constant c'est la probabilité d'être dans un intervalle :

$$P(\alpha \leq X \leq \beta) = \sum_{\alpha \leq x \leq \beta} P(X = x)$$

Il y a dans un intervalle, de plus en plus d'événements mais la somme de leur probabilité tend vers une quantitée fixée

$$P(\alpha \leq X \leq \beta) = \text{Aire "sous la cloche idéale" entre les abscisses } \alpha \text{ et } \beta \tag{Q.2}$$

La forme de la distribution se stabilise. Pour rendre compte de ce phénomène il faut utiliser la fonction de répartition, c'est-à-dire les *probabilités cumulées de la loi binomiale* (voir section 3.3.4 page 15).

En procédant comme dans l'exercice 3.32 page 16, on peut tracer le graphique des probabilités cumulées de la loi binomiale ou ceux de la loi binomiale normalisée. Voir graphique Q.3 page 202. La dernière courbe est celle de la loi normale : elle représente "l'aire sous la cloche idéale"entre $-\infty$ et x.

On appelle ce phénomène la convergence en loi de la loi binomiale normalisée vers la loi normale ("loi de la cloche idéale") quand n tend vers l'infini (avec p constant). La fonction qui est représentée sur le dernier graphe de la figure Q.3 est la fonction de répartition de la loi normale".

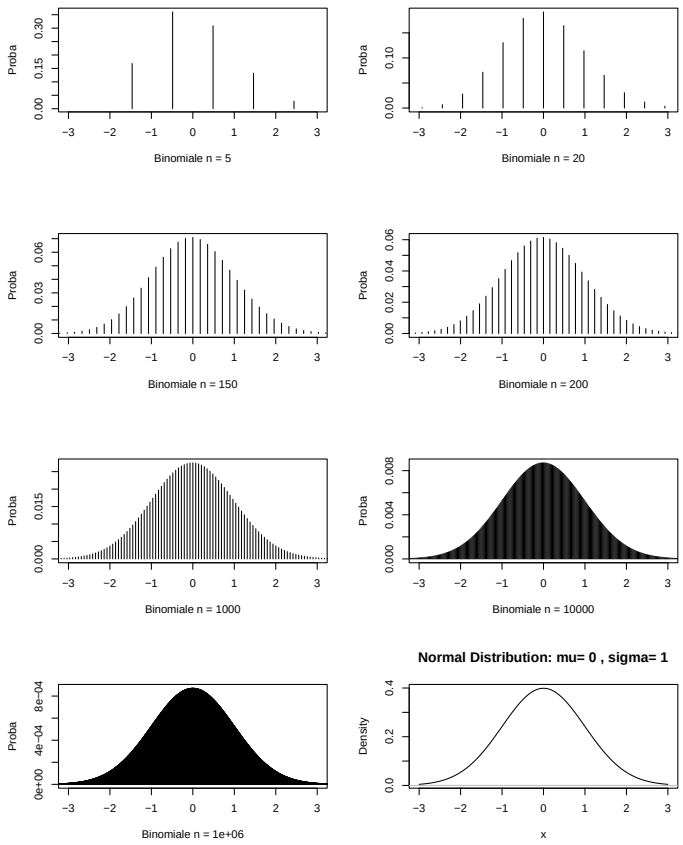


FIGURE Q.1. Les 7 graphes correspondant aux graphes de la loi binomiale normalisée avec $p = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) la loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$.

Q.2. Passage du discret au continu

Idées lors du passage du discret au continu
- Nombre de valeurs possible tend vers l'infini.
- La loi de probabilité cumulée discrète :

$$P(X \leq n_i) = \sum_{j:n_j \leq n_i} P(X = x_j)$$

remplacée par cette même somme multipliée par un coefficient "petit" et qui représente une aire approchée (formule des rectangles) qui est l'aire "sous la courbe" pour $x \leq X$, soit encore, on connaît

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy \tag{Q.3}$$

p est appelée la *densité de la loi de probabilité (continue)*. Dans ce cas, on a pour a et b :

$$P(a \leq X \leq b) = \int_{-\infty}^{b} p(y)dy - \int_{-\infty}^{a} p(y)dy = \int_{a}^{b} p(y)dy \tag{Q.4}$$

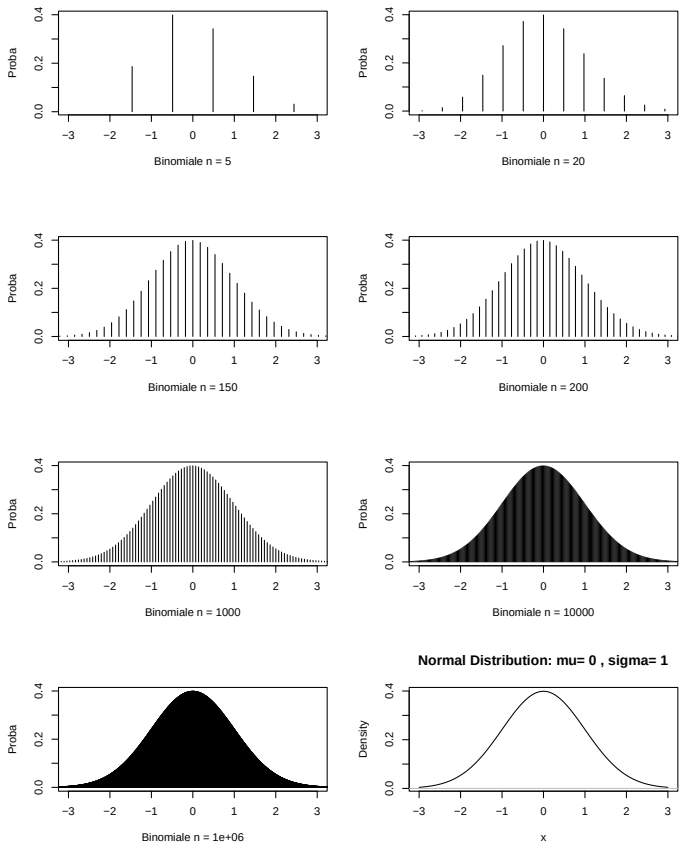


FIGURE Q.2. Les 7 graphes correspondant aux graphes de la loi binomiale normalisée avec $p = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e + 06\}$, ainsi que (pour la dernière courbe) la loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$ avec une valeur maximale imposée.

Si $a = x$ et $b = x + dx$, où dx est une "petite variation",

$$P(x \leq X \leq x + dx) = \int_x^b x + dx p(y)dy \approx p(x)dx \tag{Q.5}$$

Ainsi, la formule (Q.2) s'écrit rigoureusement

$$P(a \leq X \leq b) = \int_a^b p(x)dx \tag{Q.6}$$

où p est "la densité de la cloche idéale". Elle sera définie dans la section 3.4.3.

– Enfin, les notions d'espérance et d'écart-type des variables discrète "passent" à la limite, en remplaçant les sommes par des intégrales :

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx, \tag{Q.7a}$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 p(x)dx}. \tag{Q.7b}$$

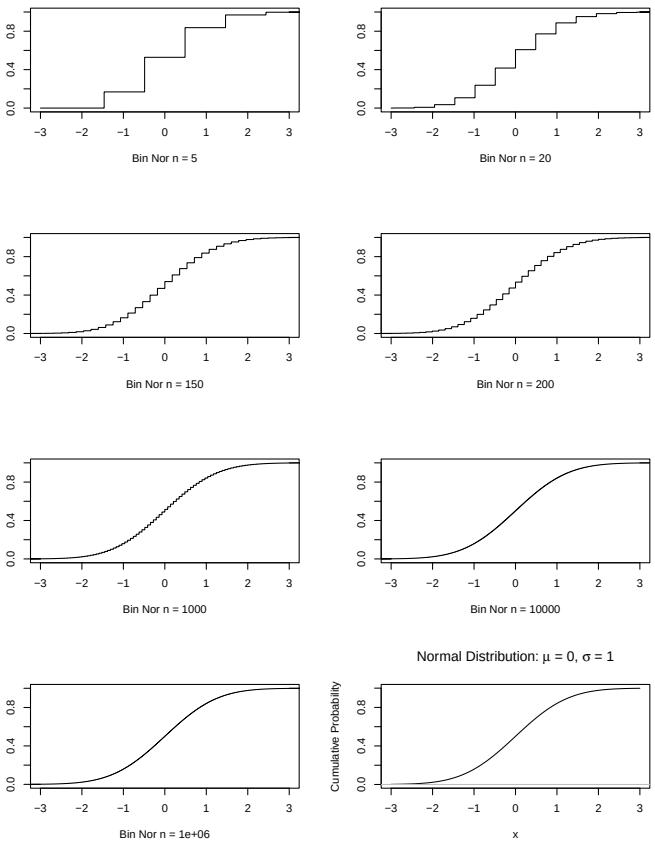


FIGURE Q.3. Les 7 graphes correspondant aux graphes des probabilités cumulées de la loi binomiale normalisée avec $p = 0.3$ et $n \in \{5, 20, 150, 200, 1000, 10000, 1e+06\}$, ainsi que (pour la dernière courbe) l'aire de la "cloche idéale".

Q.3. Éléments de correction

ÉLÉMENTS DE CORRECTION DE L'EXERCICE Q.1

Voir en figure Q.4 les 6 loi de probabilité obtenues, qui ont l'air de "se rapprocher" d'une "courbe continue" quand n augmente.

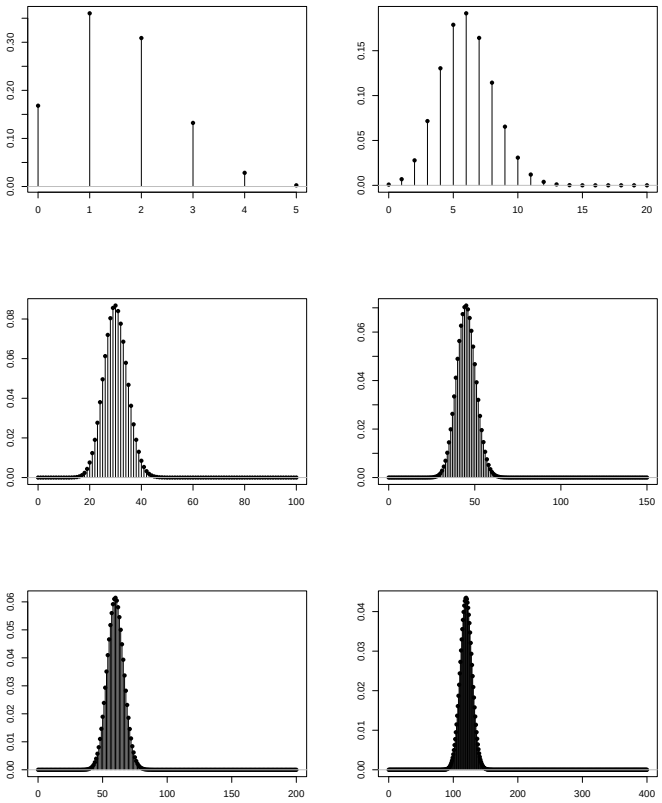


FIGURE Q.4. Les 6 graphes correspondant aux graphes de la loi binomiale avec $p = 0.3$ et $n \in \{5, 20, 100, 150, 200, 400\}$.

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] A.B Dufour and T. Jombart. Fiche de td 27 : Intervalles de confiance. Disponible sur `http://pbil.univ-lyon1.fr/R/enseignement.html`, 2007.

[9] J.P. Lecoutre. *Statistiques et probabilités*. Dunod, Paris, 2006. disponible à la BU de Lyon I sous la cote 519.2 LEC.

[10] J. Lobry. Fiche bs31 : Étude empirique de l'approximation de la loi de student par la loi de laplace-gauss. Disponible sur `http://pbil.univ-lyon1.fr/R/enseignement.html`, 2008.

[11] Alain Rey, editor. *Le robert, dictionnaire historique de la langue française*. Dictionnaires le Robert, Paris, 1998.

[12] 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.

[13] J Cohen. A power primer. *Psychological bulletin*, 112(1) :155–159, 1992.

[14] R.D Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28(9–12), 1974.

[15] Éric Marie. Initiation à la démarche qualité appliquée au champ médico-éducatif. Note de cours, 2007.

[16] F.J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27 :17–21, 1964.

[17] Émmanuel 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`, 2005.

[18] AB Dufour and D. Chessel. Fiche de cours bs1 (cours de biostatistique, illustrations dans r) : Vraisemblance d'une hypothèse. Disponible sur `http://pbil.univ-lyon1.fr/R/enseignement.html` rubrique cours, puis test d'hypothèse, 2008.