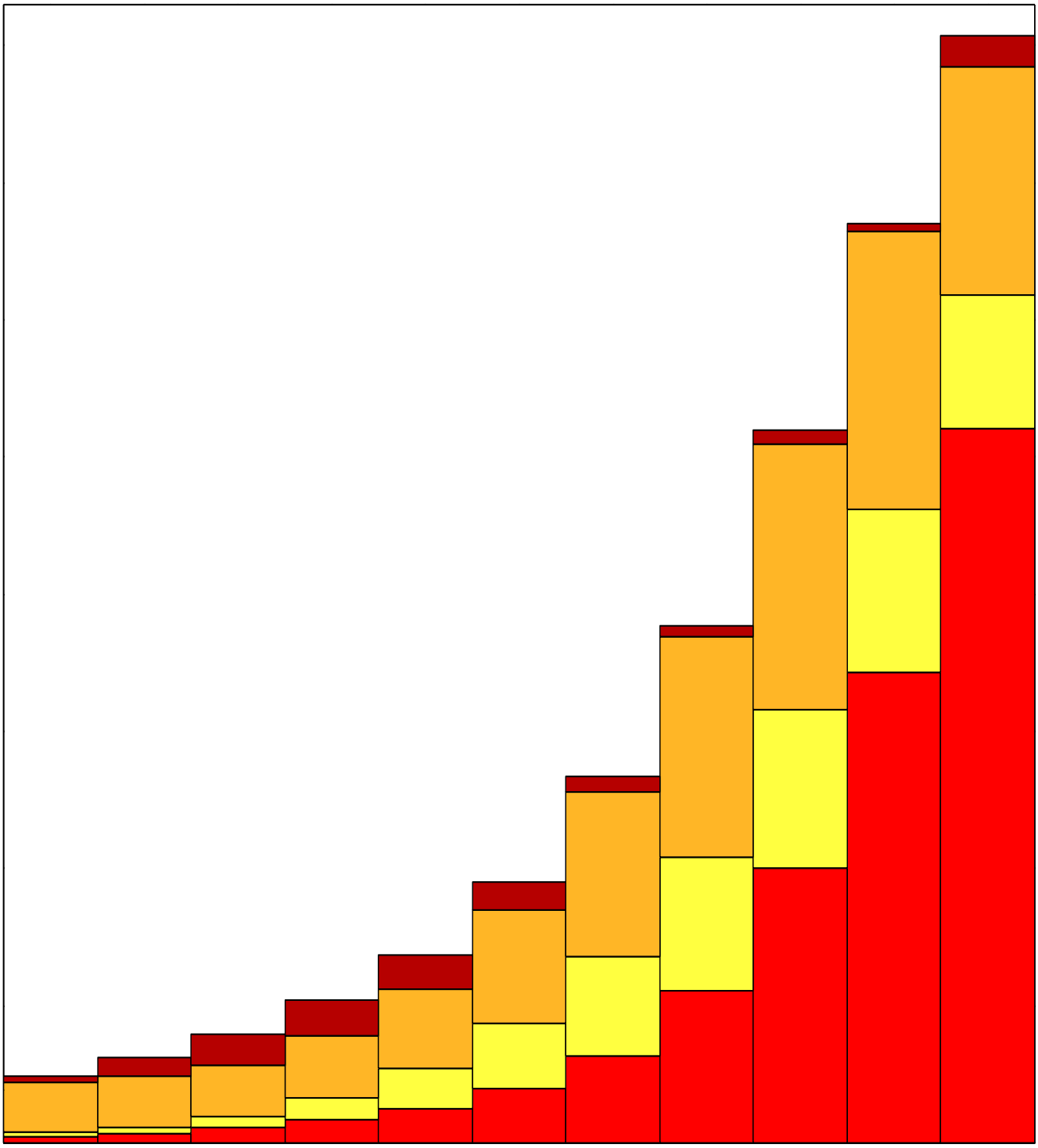
Introduction à la statistique

Corrigés des exercices

JEAN-PAUL BONVIN, ANNE-CATHERINE FAVRE, REINHARD FURRER, THOMAS GSPONER
ESTELLE MARTIN, OLIVIER RENAUD, EVA RESTLE, ANDREI ZENIDE



Mars, 2001

Chaire de statistique appliquée
http ://statwww.epfl.ch/morgenthaler/
DMA/EPFL
1015 Lausanne

Table des matières

Avant propos

Ce manuscrit présente les solutions des exercices du livre Introduction à la statistique. Il a été écrit par Eva Restle, Reinhard Furrer, Thomas Gsponer et Andrei Zenide, qui ont utilisé les esquisses de solutions préparées auparavant par Olivier Renaud, Anne-Catherine Favre, Estelle Martin et Jean-Paul Bonvin. Je suis sûr que les étudiants qui participent au cours *probabilité et statistique* à l'EPFL seront très contents d'avoir accès à cette collection d'exercices avec solutions. Quelques fautes dans les exercices de l'ancienne édition du livre ont été corrigées et les solutions présentées ici se réfèrent à la nouvelle édition. Je remercie mes assistants pour leur excellent travail, qui j'espère sera apprécié par tous les utilisateurs du livre.

Lausanne, mars 2001 Stephan Morgenthaler

Chapitre 1

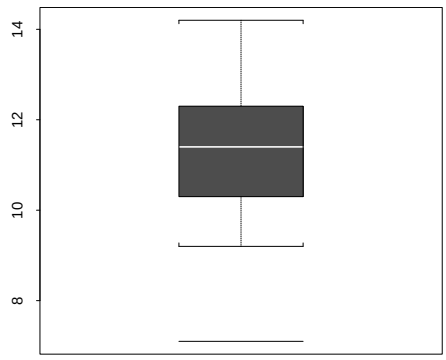
Représentations graphiques de données

1. (a) Le diagramme branche-et-feuille nous donne :

7	1
8	
9	23
10	1158
11	123568
12	2246
13	37
14	2

   (b) 10.3 .

   (c) Dans le boxplot ci-dessous on remarque que la valeur 7.1 est une valeur aberrante. La valeur trouvée au point (b) est le quartile inférieur qui détermine le niveau inférieur de la boîte (le rectangle plein).

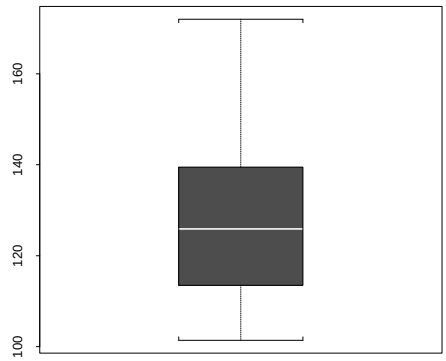


2. (a) Le diagramme branche-et-feuille est le suivant :

10	14599
11	023479
12	023688
13	05589
14	299
15	0
16	1
17	22

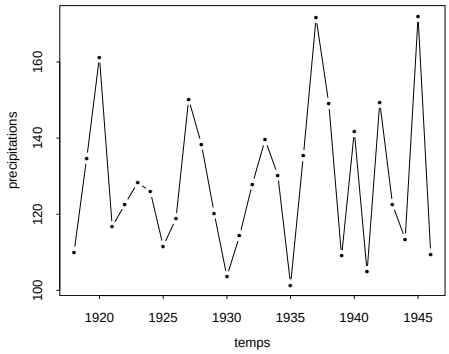
   (b) $\hat{q}(50\%) = 125.9$, $\hat{q}(25\%) = 113.5$, $\hat{q}(75\%) = 139.5$.

(c) Dans le boxplot ci-dessous ainsi que dans le diagramme branche-et-feuille ci-dessus on remarque une asymétrie : la queue droite est plus lourde.



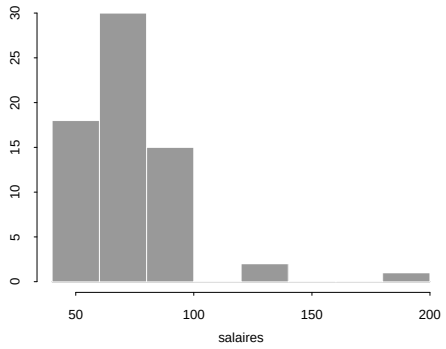
(d) Après avoir tracé le boxplot on peut affirmer qu'il n'y a pas de valeur aberrante.

(e) Un graphique mettant en évidence l'évolution temporelle des précipitations :



Les points correspondent au niveau de précipitations alors que les droites les reliant n'ont qu'un rôle graphique pour mieux mettre en évidence l'évolution temporelle des précipitations.

3. (a) Voici l'histogramme des salaires journaliers :



(b) Le salaire médian est de 70.

(c) $\bar{x} = 77.27$; la médiane calculée au point (b) représente mieux le centre de la distribution.

(d) $\hat{q}(75\%) = 90$.

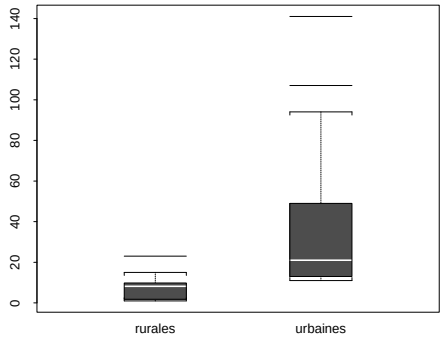
4. (a) Le diagramme branche-et-feuille est le suivant :

0	112235889
1	00112235688
2	12349
3	
4	9
5	
6	
7	
8	
9	4
10	7
11	
12	
13	
14	1

$\hat{q}(50\%) = 12$, $\hat{q}(25\%) = 8.2$, $\hat{q}(75\%) = 22$.

(b) On remarque que les sols des régions urbaines ont un taux de contamination plus élevé.

(c) Les boxplots suivants confirment le point (b) :



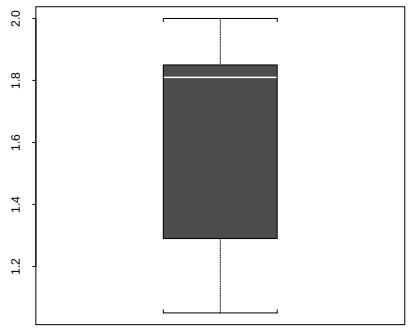
(d) Le boxplot des données urbaines nous renseigne qu'il y a deux valeurs aberrantes, le 107 et le 141 alors que pour les données rurales le 23 est une valeur aberrante.

5. (a) Voici le diagramme branche-et-feuille avec un pas de 0.2 :

1	1
1	223333
1	445
1	67
1	88888899999
2	00

(b) $\hat{q}(50\%) = 1.81$, $\hat{q}(25\%) = 1.29$, $\hat{q}(75\%) = 1.85$.

(c) Dans le boxplot ci-dessous, on remarque que la médiane est beaucoup plus proche du quartile supérieur que du quartile inférieur et qu'il n'y a pas de valeur aberrante.



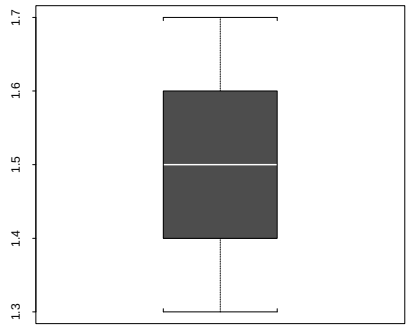
(d) On remarque la bimodalité des données, c'est-à-dire une forte concentration d'observations autour de 1.25 et de 1.85 .

(e) La qualité est médiocre car on est loin de la valeur idéale de 1.5 .

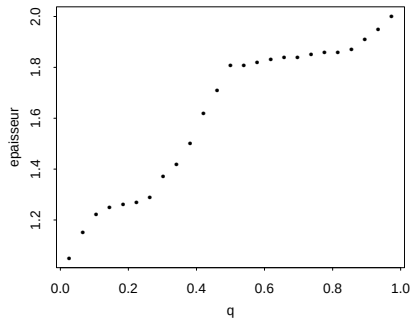
(f) Le diagramme branche-et-feuille avec un pas de 0.1 est le suivant :

1	3
1	444
1	5555
1	666
1	7

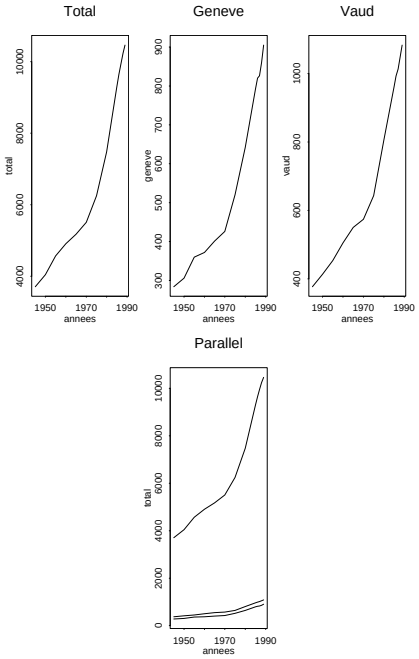
Un boxplot qui représente une production raisonnable est de la forme :



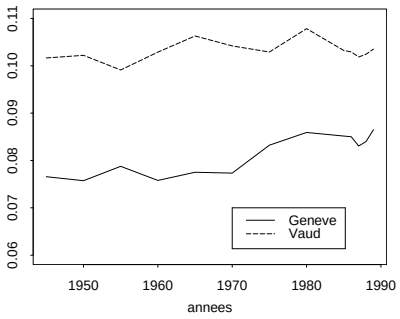
(g) Dans le diagramme des quantiles ci-dessous on remarque de nouveau la concentration autour de 1.25 et 1.85 . Cela était visible sur le graphique en branche-et-feuille également.



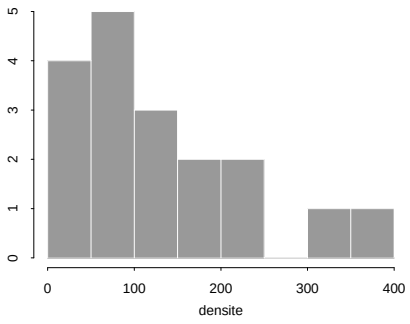
6. (a) Les graphiques suivants mettent en évidence l'évolution temporelle des séries individuellement et en parallèle :



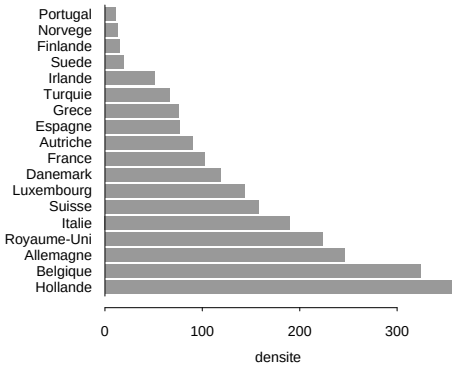
(b) On considère à chaque période le rapport entre le nombre de médecins à Genève et dans le canton de Vaud divisé par le nombre total en Suisse. Dans le graphique ci-dessous on peut voir que globalement la pente de la courbe qui concerne Genève est plus forte.



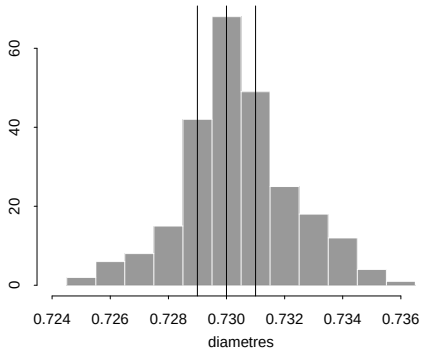
7. Dans l'histogramme ci-dessous on voit que la distribution de la densité de population est asymétrique avec un pic vers la gauche :



Un autre graphique approprié est le diagramme en barres :



8. (a) Voici l'histogramme des diamètres de têtes de vis :

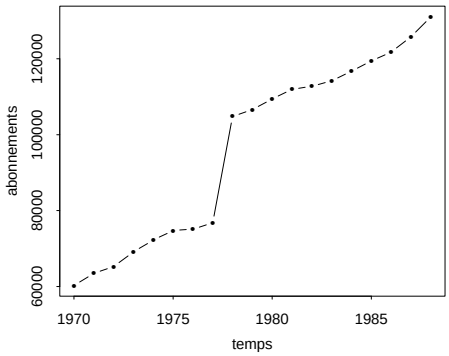


(b) $\hat{q}(50\%) = 0.730$, $\hat{q}(25\%) = 0.729$, $\hat{q}(75\%) = 0.731$, *cf.* les lignes verticales dans l'histogramme sous (a). On remarque que la distribution des diamètres est assez centrée.

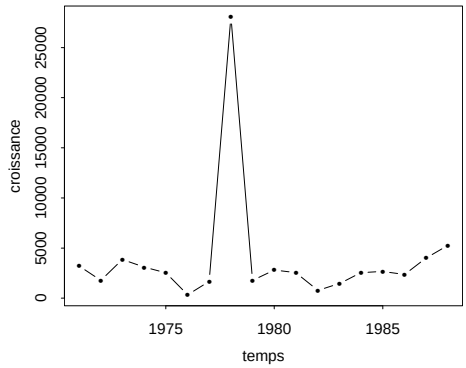
(c) Si on note le diamètre moyen par $\bar{x}$ on trouve la formule :

$$\bar{x} = \sum_{i=1}^{12} x_i \cdot p_i.$$

9. (a) Dans le graphique suivant on voit l'évolution du nombre d'abonnements :

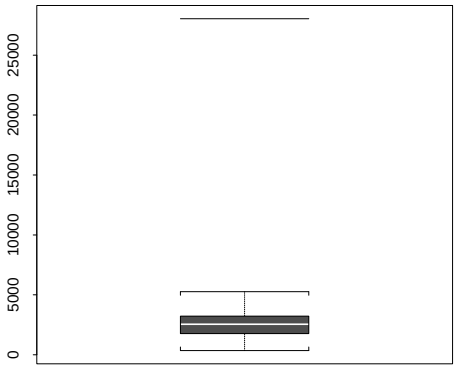


(b) Voici le graphique de la croissance des nouveaux abonnements :

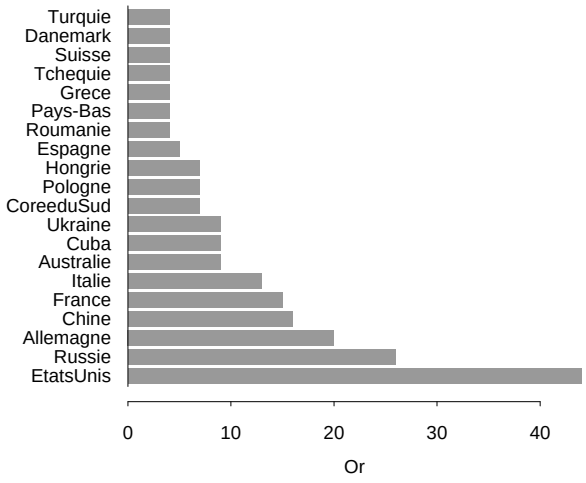


La médiane est égale à 2546.5 et la moyenne à 3929.7 . La grosse différence est due au saut enregistré en 1978 quand le nombre de nouveaux abonnements a été de 28045, une valeur nettement supérieure à la moyenne des autres années.

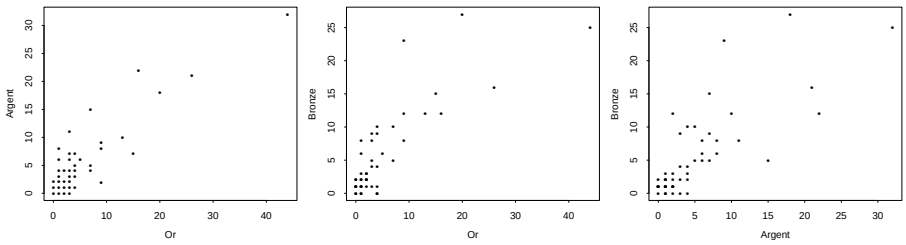
(c) Dans le boxplot ci-dessous, on remarque de nouveau la valeur aberrante :



11. (a) Un graphique approprié est le diagramme en barres. On a considéré seulement les vingt premiers pays du tableau :



(b) On a considéré les couples de médailles : or-argent, or-bronze et argent-bronze pour faire trois graphiques avec des axes perpendiculaires. A chaque point correspond un pays parmi les vingt premiers médaillés d'or :



Il est évident, d'après ces graphiques, qu'il y a un lien entre le nombre des différentes médailles.

Chapitre 2

Moyenne, écart-type et loi normale

1. Le salaire moyen $\bar{x}$ vaut 77.27 ; dans l'intervalle $[\,\bar{x}-s, \bar{x}+s\,] = [\,54.17, 100.37\,]$ nous comptons 55 observations, ce qui fait une proportion empirique de $5/6 = 0.83$. Avec $X \sim \mathcal{N}(77.27, 23.1^2)$, nous trouvons la proportion théorique $P\{\bar{x}-s \leq X \leq \bar{x}+s\} = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.681$. Ainsi, la loi normale n'est pas une bonne approximation.

2. (a) On cherche a et b tel que $P\{a \leq X \leq b\} = 0.95$ et $b - \bar{x} = \bar{x} - a$:

$$P\{a \leq X \leq b\} = P\Big\{\frac{a-\bar{x}}{s} \leq \frac{X-\bar{x}}{s} \leq \frac{b-\bar{x}}{s}\Big\} = 2\Phi(a_{\text{réd}}) - 1 = 0.95\,.$$

   Comme $a_{\text{réd}} = 1.96$, on a $a = \bar{x} - s a_{\text{réd}} = 6.56$ et $b = \bar{x} + s a_{\text{réd}} = 12.44$.

   (b) On calcule

$$P\{6 \leq X \leq 7\} = P\Big\{-2.33 \leq \frac{X-9.5}{1.5} \leq -1.67\Big\} = 0.038.$$

3. La proportion d'automobiles ayant commis un excès de vitesse vaut $P\{X > 80\} = 1 - P\big\{(X-\bar{x})/s \leq (80-72)/8\big\} = 1 - \Phi(1) = 0.159$, où X représente la vitesse.

4. (a) Pour le graphique *cf.* (h).

   (b) La moyenne vaut $\bar{y} = 74.006$ et la médiane $m = 74.002$. La médiane représente mieux le centre de la distribution car la moyenne est très influencée par des valeurs aberrantes.

   (c) Après l'élimination de la valeur aberrante la moyenne vaut $\bar{y} = 74.003$, *i.e.* elle a diminué, tandis que la médiane ne change pas.

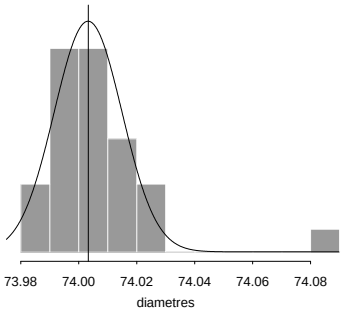
   (d) $s = 0.0118$.

   (e) Dans l'intervalle $[\,\bar{y} - s, \bar{y} + s\,] = [\,73.991, 74.015\,]$ nous comptons 20 observations, ce qui fait une proportion empirique de 0.689. Dans l'intervalle $[\,\bar{y} - 2s, \bar{y} + 2s\,] = [\,73.979, 74.026\,]$ nous comptons 28 observations, ce qui fait une proportion empirique de 0.965.

   (f) Avec $Y \sim \mathcal{N}(\bar{y}, s^2)$, nous trouvons les proportions théoriques $P\{\bar{x}-s \leq X \leq \bar{x}+s\} = 0.681$ et $P\{\bar{y}-2s \leq Y \leq \bar{y}+2s\} = 2\Phi(2) - 1 = 0.954$.

   (g) Comme la distribution empirique est unimodale et quasiment symétrique et en comparant (e) et (f) on peut affirmer que le modèle normal est une bonne approximation.

   (h) Voici l'histogramme des données avec la densité de la loi normale superposée :

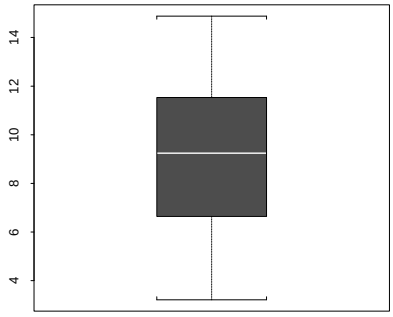


5. (a) Soit Y la quantité concernée, c'est-à-dire $Y \sim \mathcal{N}(100, 5^2)$. Nous trouvons la proportion de bouteilles détruites $P\{Y \leq 90.2\} = P\big\{(Y-100)/5 \leq -1.96\big\} = 1 - \Phi(1.96) = 0.025$.

   (b) Gain moyen $= 0.975 \cdot 80$ cts $- 100$ cl $\cdot\, 0.2$ct/cl $= 58$ ct.

   (c) $\bar{y} = 8.91$, $s = 3.259$.

   (d) Voici le boxplot des ventes journalières :



   (e) La symétrie du boxplot n'indique pas une aberration grossière de la loi normale, de plus dans l'intervalle $[\,\bar{y}-s, \bar{y}+s\,] = [\,5.65, 12.169\,]$ nous comptons 11 observations, ce qui fait une proportion empirique de 0.6875, le modèle normal semble être adéquat. Il faudrait encore examiner si les données sont unimodales.

   (f) On cherche y tel que $P\big\{(Y-\bar{y})/s \leq (y-\bar{y})/s\big\} = 0.90$, donc $y = \bar{y} + s \cdot z = 8.91 + 3.259 \cdot 1.28 = 13.08$ milliers de litres.

6. (a) $\bar{y} = 39.07$, $s = 40.949$.

   (b) Dans l'intervalle $[\,\bar{y}-s, \bar{y}+s\,] = [\,-1.88, 80.02\,]$ on trouve 12 observations, donc une proportion de 0.8, dans $[\,\bar{y}-2s, \bar{y}+2s\,] = [\,-42.83, 120.97\,]$ on trouve 14 observations, donc une proportion de 0.93. Pour le modèle normal $P\{\bar{y} - s \leq Y \leq \bar{y} + s\} = 2\Phi(1) - 1 = 0.682$. L'approximation normale n'est pas bonne. De plus le diagramme branche-et-feuille est fortement asymétrique.

7. Soit X le diamètre, ainsi $X \sim \mathcal{N}(10, 0.2^2)$.

   (a)
$$P\{X \leq 9.95\} = P\Big\{\frac{X-10}{0.2} \leq 0.25\Big\} = 0.401\,,$$
$$P\{9.95 < X \leq 9.99\} = P\Big\{-0.25 < \frac{X-10}{0.2} \leq -0.05\Big\} = 0.079\,,$$
$$P\{X > 9.99\} = 1 - \big(P\{X \leq 9.95\} + P\{9.95 < X \leq 9.99\}\big) = 0.52\,.$$

   (b) Profit $= 5 \cdot 0.079 + 15 \cdot 0.52 - 7 =$ Fr. 1.195.

8. (a) $\bar{y} = 1.6144$, $s = 0.2992$.

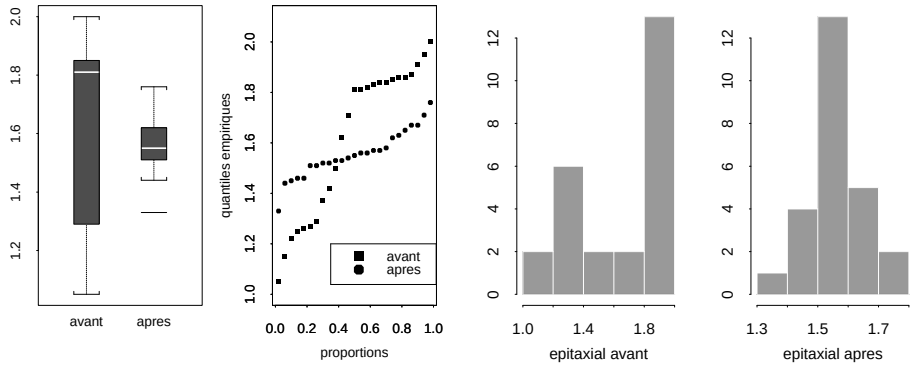
   (b) On trouve 16 observations dans l'intervalle $[\bar{y} - s, \bar{y} + s] = [1.3152, 1.9136]$ et 24 dans $[\bar{y} - 2s, \bar{y} + 2s] = [1.0160, 2.2128]$, ce qui correspond à des proportions empiriques de respectivement 0.64 et 0.96 .

   (c) Pour les proportions de 0.682 et 0.954 calculées sur la base de la loi normale *cf.* exercice 4.

   (d) La distribution n'est pas normale même si les proportions calculées sous (b) et (c) ne sont pas très différentes. Il y a beaucoup d'observations dans la queue de l'échantillon (grandes observations). Un histogramme permet de repérer cette caractéristique.

9. (a) Voici les diagrammes branche-et-feuille :

10	5		10	
11	5		11	
12	25679		12	
13	7		13	3
14	2		14	4566
15	0		15	1122334566778
16	2		16	23577
17	1		17	16
18	1123445667		18	
19	15		19	
20	0		20	
	×0.1			×0.1



   L'objectif de qualité est atteint. Les données sont centrées sur la valeur de référence et l'écart-type est faible.

   (b) Il y a 11 observations de 25 entre 1.45 et 1.55, ce qui fait une proportion de 0.44 . Avec $\bar{x} = 1.556$, $s = 0.094$ et $X \sim \mathcal{N}(1.556, 0.094^2)$ nous trouvons

$$P\{1.45 \leq X \leq 1.55\} = P\Big\{ -1.1277 \leq \frac{X - 1.556}{0.094} \leq -0.0639 \Big\} = 0.345.$$

Chapitre 3

Probabilités d'événements

1. On a par indépendance des événements A et B

$$\begin{aligned} P\{A^c \cap B^c\} &= P\{(A \cup B)^c\} = 1 - P\{A \cup B\} = 1 - P\{A\} - P\{B\} + P\{A \cap B\} \\ &= 1 - P\{A\} - P\{B\} + P\{A\}\,P\{B\} = \big(1 - P\{A\}\big)\,\big(1 - P\{B\}\big) \\ &= P\{A^c\}\,P\{B^c\}, \end{aligned}$$

   donc les événements A^c et B^c sont indépendants.

2. Le vendeur gagne en moyenne $300 \cdot 0.3 = 90$ Francs s'il pleut et il perd en moyenne $60\,(1-0.3) = 42$ Francs s'il fait beau temps. Le vendeur peut donc espérer gagner 48 Francs par jour.

3. (a) $P\{111\} = 1/6^3 = 1/216$.
   (b) $1 - P\{111\} = 215/216$.
   (c) $P\{\text{Somme } \leq 10\} = P\{\text{Somme } = 3\} + \cdots + P\{\text{Somme } = 10\}$. Le nombre de cas favorables vaut $1+3+6+10+15+21+25+27 = 108$ et la probabilité cherchée est donc
$$\frac{108}{6^3} = \frac{108}{216} = 0.5\,.$$

4. Soient les événements A : "le biréacteur termine son vol sans difficultés" et B : "le quadriréacteur termine son vol sans difficultés".

$$P\{A\} = \binom{2}{1}(1-p)\,p + \binom{2}{0}p^2 = 2p - p^2,$$

$$P\{B\} = \binom{4}{2}(1-p)^2\,p^2 + \binom{4}{1}(1-p)\,p^3 + \binom{4}{0}p^4 = 3\,p^4 - 8\,p^3 + 6\,p^2,$$

   $P\{B\} - P\{A\} = 3\,p^4 - 8\,p^3 + 7\,p^2 - 2\,p = p(p-1)^2(3p-2).$

   Comme p et $(p-1)^2$ sont positive $P\{B\} \geq P\{A\} \iff 3\,p - 2 \geq 0 \iff 2/3$, donc les quadriréacteurs sont donc préférables aux biréacteurs si $p > 2/3$.

5. $P\{\,\text{jeter la pièce } n \text{ fois}\,\} = (1-p)^{n-1}p$.

6. Soit p la probabilité de n'obtenir aucun as en un lancer simple. Alors $p = 5/6$ et $P\{A\} = 1 - P\{\text{aucun as en 4 lancers simples}\} = 1 - \big(5/6\big)^4 = 0.518$. Soit q la probabilité de ne pas obtenir un double as en un lancer double. Alors $q = 1 - P\{\text{double as en un lancer double}\} = 1 - 1/36 = 35/36$ et $P\{B\} = 1 - P\{\text{aucun double as en 24 lancers doubles}\} = 1 - \big(35/36\big)^{24} = 0.491$. Donc A est l'événement le plus probable.

7. (i) Le livre 1 restant à sa place, il y a $(n-1)!$ permutations $p = (n-1)!/n! = 1/n$.

   (ii) Les livres 3 et 4 sont à considérer comme un seul livre ; il y a à nouveau $(n-1)!$ permutations et $p = 1/n$.

8. (i) $4! = 24$ arrangements.

   (ii) $2! \cdot 2! = 4$ arrangements.

9. (a) $P\{4 \text{ as}\} = \binom{4}{4}\Big/\binom{36}{4} = 1.698 \cdot 10^{-5}$.

   (b) $P\{\text{exactement 2 as}\} = \binom{4}{2}\binom{32}{2}\Big/\binom{36}{4} = 5.052 \cdot 10^{-2}$,

   $P\{\text{exactement 3 as}\} = \binom{4}{3}\binom{32}{1}\Big/\binom{36}{4} = 2.173 \cdot 10^{-3}$,

   $P\{\text{au moins 2 as}\} = P\{\text{ex. 2 as}\} + P\{\text{ex. 3 as}\} + P\{4 \text{ as}\} = 5.271 \cdot 10^{-2}$.

   (c) $P\{\text{couleur}\} = \binom{18}{4}\Big/\binom{36}{4} = 0.0519$ si on considère les couleurs rouge/noir.

   $P\{\text{couleur}\} = \binom{9}{4}\Big/\binom{36}{4} = 2.139 \cdot 10^{-3}$ si on considère 4 couleurs (pique, coeur, carreau, trèfle).

10. (a) $P\{666\} = 1/6^3 = 0.463 \cdot 10^{-3}$.

    (b) $P\{\text{au moins un 6}\} = 1 - P\{\text{aucun 6}\} = 1 - \big(5/6\big)^3 = 0.421$.

    (c) $P\{\text{aucun dé avec 5 ou 6}\} = \big(4/6\big)^3 = 0.296$.

    (d) Il y a $1 + 3 + 6 + 10 = 20$ combinaisons possibles, donc la probabilité cherchée vaut $20/216 = 0.093$.

11. (a) $\binom{4}{3}\binom{4}{2}/\binom{52}{5} = 1/108290 = 9.234 \cdot 10^{-6}$.

    (b) $\binom{13}{2}\binom{39}{2}/\binom{52}{5} = 741/33320 = 2.224 \cdot 10^{-2}$.

    (c) $\binom{13}{1}\binom{13}{1}\binom{13}{1}\binom{13}{2}\binom{4}{1}/\binom{52}{5} = 2197/8330 = 0.264$.

12. (a) Chaque personne peut répéter l'information à $N-1$ autres personnes (elle ne peut pas se le dire à elle-même). Etant donné qu'il ne faut pas répéter l'information à la première personne, il y a $N-2$ cas favorables. Ainsi la probabilité cherchée est $\big((N-2)/(N-1)\big)^r$.

    (b) A nouveau, chaque personne ne peut indiquer l'information qu'à $N-1$ personnes. Mais la i^{e} personne dans la chaîne ne peut dire l'information qu'aux $N-i$ personnes restantes. Ainsi la probabilité cherchée est

    $$\frac{N-1}{N-1} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-1} \cdot \ldots \cdot \frac{N-r-1}{N-1} = \frac{(N-2)!}{(N-1)^r (N-r-2)!},$$

    si $r \leq N-2$ et 0 si $r > N-2$.

13. (a) Le résultat de l'expérience qui consiste à répartir les n boules indiscernables dans les r urnes est décrit par un vecteur $(x_1, \ldots, x_r)$ où x_i représente le nombre des boules contenues dans la i^{e} urne. Le problème revient alors à trouver le nombre de vecteurs $(x_1, \ldots, x_r)$ à composantes entières non négatives tels que $x_1 + \cdots + x_r = n$. Pour le cas, où on mets des boules dans chaque urne (*i.e.* $x_i > 0$ pour tout i), on doit séparer n objets avec $r-1$ séparateurs et pour cela il y a $\binom{n-1}{r-1}$ possibilités. Pour le cas, où il peut y avoir

aucune boule dans une urne, on pose $y_i = x_i + 1$ et on cherche le nombre de vecteurs $(y_1, \ldots, y_r)$ avec $y_i > 0$ tels que $y_1 + \cdots + y_r = n + r$. Donc il y a $\binom{n+r-1}{r}$ vecteurs distincts à composantes entières et non négatives tels que $x_1 + \cdots + x_r = n$.
Pour notre exemple $r = 5$ (les gens quittent l'ascenseur aux étages 1, 2, 3, 4 ou 5) et $n = 8$; le nombre vaut donc $\binom{8+5-1}{8} = \binom{12}{8} = 495$.

(b) D'après le même principe que sous (a), il y a $\binom{5+5-1}{5}\binom{3+5-1}{3} = 126 \cdot 35 = 4410$ possibilités.

14. (a) $10 \cdot 9 \cdot 8 = \binom{10}{3}\,3\,! = 720$.

(b) Comités avec A, sans B : $\binom{8}{2}\,3\,! = 168$. Comités sans A ni B : $\binom{8}{3}\,3\,! = 336$. Au total : $2 \cdot 168 + 336 = 672$.

(c) Comités avec C et D : $8 \cdot 3\,! = 48$. Au total : $48 + 336 = 384$.

(d) $\binom{9}{2}\,3\,! = 216$.

(e) $\binom{9}{2}\,2\,! + \binom{9}{3}\,3\,! = 72 + 504 = 576$.

15. (a) $\binom{6}{3\,2\,1} = \frac{6\,!}{3\,!\,2\,!\,1\,!} = 60$.

(b) $3\,! = 6$.

(c) $\binom{4}{1\,2\,1} = \frac{4\,!}{1\,!\,2\,!\,1\,!} = 12$.

16. (a) $E^c \cap F^c \cap G^c$.

(b) $(E \cap F^c \cap G^c) \cup (E^c \cap F \cap G^c) \cup (E^c \cap F^c \cap G) \cup (E^c \cap F^c \cap G^c)$.

(c) $(E \cap F \cap G^c) \cup (E^c \cap F \cap G) \cup (E \cap F \cap G^c)$.

(d) $E \cap F^c \cap G$.

17. $P\{\text{ex. } k \text{ boules blanches}\} = P\{k \text{ boules blanches et } r-k \text{ noires}\} = \frac{\binom{B}{k}\binom{N}{r-k}}{\binom{B+N}{r}}$. Pour le cas $B = k = 1$ on trouve $\frac{\binom{N}{r-1}}{\binom{N+1}{r}} = \frac{r}{(N+1)}$.

18.
$$\begin{aligned}
E &= \{(1,2),\ (1,4),\ (1,6),\ (2,1),\ (2,3),\ (2,5),\ (3,2),\ (3,4),\ (3,6),\\
&\qquad (4,1),\ (4,3),\ (4,5),\ (5,2),\ (5,4),\ (5,6),\ (6,1),\ (6,3),\ (6,5)\},\\
F &= \{(1,1),\ (1,2),\ (1,3),\ (1,4),\ (1,5),\ (1,6),\ (2,1),\ (3,1),\ (4,1),\\
&\qquad (5,1),\ (6,1)\},\\
G &= \{(4,1),\ (3,2),\ (2,3),\ (1,4)\},\\
E \cap F &= \{(1,2),\ (1,4),\ (1,6),\ (2,1),\ (4,1),\ (6,1)\},\\
E \cup F &= \{(1,i) \mid i = 1,\ldots,6\} \cup \{(i,1) \mid 1 = 2,\ldots 6\} \cup \{(2,3),\ (2,5),\ (3,2),\\
&\qquad (3,4),\ (3,6),\ (4,3),\ (4,5),\ (5,2),\ (5,4),\ (5,6),\ (6,3),\ (6,5)\}\\
F \cap G &= \{(4,1),\ (1,4)\},\\
E \cap F^c &= \{(2,3),\ (2,5),\ (3,2),\ (3,4),\ (3,6),\ (4,3),\ (4,5),\ (5,2),\ (5,4),\\
&\qquad (5,6),\ (6,3),\ (6,5)\},\\
E \cap F \cap G &= \{(4,1),\ (1,4)\}.
\end{aligned}$$

19. On calcule $P\{14^{\text{e}}\text{ carte est un as}\} = \frac{4}{52} = \frac{1}{13}$; $P\{14^{\text{e}}\text{ carte est le premier as}\} = \frac{48}{52} \cdot \frac{47}{51} \cdots \frac{36}{40} \cdot \frac{4}{39} = 4 \cdot \binom{48}{13} / \left(14 \cdot \binom{52}{14}\right) = 0.031$.

20. (a) $M = \{(F,G),\ (G,F)\}$ et $N = \{(G,G),\ (F,G),\ (G,F)\}$, alors $P\{M\} = 1/2$, $P\{N\} = 3/4$ et $P\{M \cap N\} = P\{M\} = 1/2$.

    (b) $M = \{(F,G,F),\ (F,G,G),\ (G,F,F),\ (G,F,G),\ (F,F,G),\ (G,G,F)\}$
    $N = \{(G,G,G),\ (F,G,G),\ (G,F,G),\ (G,G,F)\}$, alors $P\{M\} = 3/4$, $P\{N\} = 1/2$ et $P\{M \cap N\} = 3/8$.

21. Notons $an =$ "auteur est anglais", $am =$ "auteur est américain" et $v =$ "lettre est une voyelle" : $P\{an \,|\, v\} = \dfrac{P\{v \,|\, an\} \cdot P\{an\}}{P\{v \,|\, an\}P\{an\} + P\{v \,|\, am\}P\{am\}} = \frac{\frac{3}{6}\cdot\frac{2}{5}}{\frac{3}{6}\cdot\frac{2}{5}+\frac{2}{5}\cdot\frac{3}{5}} = \dfrac{5}{11}$.

22. (a) La probabilité cherchée vaut

$$\begin{aligned} &P\{\text{pièce est mauvaise}\} \\ &\quad = P\{\text{pièce est mauvaise} \,|\, \text{machine } A\} \cdot P\{\text{machine } A\} \\ &\qquad + P\{\text{pièce est mauvaise} \,|\, \text{machine } B\} \cdot P\{\text{machine } B\} \\ &\qquad + P\{\text{pièce est mauvaise} \,|\, \text{machine } C\} \cdot P\{\text{machine } C\} \\ &\quad = 0.03 \cdot 0.4 + 0.05 \cdot 0.35 + 0.04 \cdot 0.25 \\ &\quad = 0.0395\,. \end{aligned}$$

    (b) On trouve la probabilité suivante :

$$\begin{aligned} &P\{\text{machine } A \,|\, \text{pièce est mauvaise}\} \\ &\quad = \frac{P\{\text{pièce est mauvaise} \,|\, \text{machine } A\} \cdot P\{\text{machine } A\}}{P\{\text{pièce est mauvaise}\}} \\ &\quad = \frac{0.03 \cdot 0.4}{0.0395} = 0.3038\,. \end{aligned}$$

23. Notons $p1 =$ "1^{re} carte est pique", $p23 =$ "2^{e} et 3^{e} cartes sont piques" :

$$\begin{aligned} P\{p1 \,|\, p23\} &= \frac{P\{p1 \cap p23\}}{P\{p23 \,|\, p1\}P\{p1\} + P\{p23 \,|\, p1^c\}P\{p1^c\}} \\ &= \frac{\frac{13}{52}\,\frac{12}{51}\,\frac{11}{50}}{\left(\frac{12}{51}\cdot\frac{11}{50}\right)\cdot\frac{13}{52} + \left(\frac{13}{51}\cdot\frac{12}{50}\right)\frac{39}{52}} = \frac{11}{50} = 0.22\,. \end{aligned}$$

24. (a) $P\{1^{\text{re}}\text{ pièce est intacte}\} = 7/10$.

    (b) $P\{\text{au moins 1 pièce défectueuse}\} = 1 - P\{\text{aucune pièce défectueuse}\}$
    $= 1 - \frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \approx 0.7$.

    (c) $P\{\text{au moins 1 pièce défectueuse}\} = 1 - P\{\text{aucune pièce défectueuse}\}$
    $= 1 - \frac{7}{10} \cdot \frac{7}{10} \cdot \frac{7}{10} \approx 0.657$,
    $P\{\text{exactement 2 pièces défectueuses}\} = \binom{3}{2}\frac{3}{10} \cdot \frac{3}{10} \cdot \frac{7}{10} \approx 0.189$.

25. (a) Le calcul est le suivant :

$$\begin{aligned} P\{7 \le L_A \le 9\} &= P\Big\{\frac{7-8}{2} \le \frac{L_A-8}{2} \le \frac{9-8}{2}\Big\} \\ &= P\Big\{-\frac{1}{2} \le \frac{L_A-8}{2} \le \frac{1}{2}\Big\} \\ &= \Phi\Big(\frac{1}{2}\Big) - \Phi\Big(-\frac{1}{2}\Big) = 2\,\Phi\Big(\frac{1}{2}\Big) - 1 \approx 0.382, \\ P\{7 \le L_B \le 9\} &= P\Big\{-\frac{1}{2} \le \frac{L_B-7.5}{1} \le \frac{3}{2}\Big\} \\ &= \Phi\Big(\frac{3}{2}\Big) - \Phi\Big(-\frac{1}{2}\Big) = \Phi\Big(\frac{3}{2}\Big) + \Phi\Big(\frac{1}{2}\Big) - 1 = 0.624\,. \end{aligned}$$

On choisit la machine B.

(b) $P\Big\{-\frac{1}{\sigma} \le \frac{L_A-8}{\sigma} \le \frac{1}{\sigma}\Big\} = 2\,\Phi\Big(\frac{1}{\sigma}\Big) - 1 = 0.624 \implies \sigma \approx 1.136\,.$

26. (a) $\binom{4}{1}\,0.2^1 \cdot 0.8^3 = 4 \cdot 0.2 \cdot 0.8^3 = 0.41\,.$

(b) $\binom{4}{i}\,0.2^i \cdot 0.8^{4-i}\,.$

(c) $P\{\text{au plus 2 pièces défectueuses}\}$
$= P\{\text{aucune pièce défectueuse}\} + P\{\text{exactement 1 pièce défectueuse}\}$
$\quad + P\{\text{exactement 2 pièces défectueuses}\}$
$= \binom{4}{0}\,0.2^0 \cdot 0.8^4 + \binom{4}{1}\,0.2^1 \cdot 0.8^3 + \binom{4}{2}\,0.2^2 \cdot 0.8^2 = 0.97\,.$

27. (a) $P\{\text{plein d'essence sans plomb}\}$
$= P\{\text{plein} \,|\, \text{essence sans plomb}\} \cdot P\{\text{plein d'essence sans plomb}\}$
$= 0.3 \cdot 0.4 = 0.12\,.$

(b) $P\{\text{plein}\} = \sum_{\text{type}} P\{\text{plein} \,|\, \text{type d'essence}\}\, P\{\text{type d'essence}\}$
$= 0.6 \cdot 0.35 + 0.4 \cdot 0.3 + 0.25 \cdot 0.5 = 0.455\,.$

(c) $P\{\text{normale} \,|\, \text{plein}\} = \dfrac{P\{\text{plein} \,|\, \text{normale}\}\, P\{\text{normale}\}}{P\{\text{plein}\}} = 0.462,$

$P\{\text{sans plomb} \,|\, \text{plein}\} = 0.264,\ P\{\text{super} \,|\, \text{plein}\} = 0.274\,.$

28. (a) $P\{\text{blanche}\} = \sum_{\text{urne}} P\{\text{blanche} \,|\, \text{urne}\}\, P\{\text{urne}\} = \frac{5}{12} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{2} \approx 0.308\,.$

(b) $P\{\text{urne } A \,|\, \text{blanche}\} = P\{\text{blanche} \,|\, \text{urne } A\}\, \dfrac{P\{\text{urne } A\}}{P\{\text{blanche}\}} = \frac{5}{12}\ \frac{\frac{1}{2}}{0.308} \approx 0.676\,.$

29. (a) $P\{\text{au moins un accident}\} = 1 - P\{\text{aucun accident}\}$
$= 1 - \sum_{\text{groupe}} P\{\text{aucun accident} \,|\, \text{groupe}\} \cdot P\{\text{groupe}\}$
$= 1 - (0.95 \cdot 0.2 + 0.85 \cdot 0.5 + 0.7 \cdot 0.3) = 1 - 0.825 = 0.175\,.$

(b) $P\{\text{groupe “bas”} \,|\, \text{pas d'accident}\}$

$$= P\{\text{pas d'accident} \,|\, \text{groupe “bas”}\} \cdot \frac{P\{\text{groupe “bas”}\}}{P\{\text{ pas d'accident}\}}$$

$$= 0.95 \cdot \frac{0.2}{0.825} = 0.23\,.$$

$$P\{\text{groupe“moyen” } |\, \text{pas d'accident}\} = 0.85 \cdot \frac{0.5}{0.825} = 0.515\,.$$

30. (a) $P\{2^{\text{e}} \text{ boule est blanche}\}$

$$= P\{2^{\text{e}} \text{ boule est blanche} \,|\, 1^{\text{re}} \text{ boule est blanche}\}\, P\{1^{\text{re}} \text{ boule est blanche}\}$$

$$+P\{2^{\text{e}} \text{ boule est blanche} \,|\, 1^{\text{re}} \text{ boule est rouge}\}\, P\{1^{\text{re}} \text{ boule est rouge}\}$$

$$= \frac{2}{3} \cdot \frac{2}{6} + \frac{1}{3} \cdot \frac{4}{6} = \frac{8}{18} \approx 0.444\,.$$

(b) $P\{1^{\text{re}} \text{ boule est blanche} \,|\, 2^{\text{e}} \text{ boule est blanche}\}$

$$= P\{2^{\text{e}} \text{ boule est blanche} \,|\, 1^{\text{re}} \text{ boule est blanche}\} \frac{P\{1^{\text{re}} \text{ boule est blanche}\}}{P\{2^{\text{e}} \text{ boule est blanche}\}}$$

$$= \frac{2}{3} \cdot \frac{2}{6} \cdot \frac{18}{6} = \frac{1}{2}.$$

Chapitre 4

Variables aléatoires

1. (a) Comme $\int_{-1}^{1}(1-x^2)\ \mathsf{d}x = 4/3$, $c = 3/4$.

   (b) $F_X(t) = \int_{-1}^{t}\frac{3}{4}(1-x^2)\,\mathsf{d}x = \frac{1}{4}(3t - t^3 + 2)$.

2. Comme $F_Y(t) = P\{\log(X) \le t\} = P\{X \le \mathrm{e}^t\} = 1 - \mathrm{e}^{-\mathrm{e}^t}$, on a $f_Y(t) = \mathrm{e}^{t-\mathrm{e}^t}$.

3. (a) Comme $\int_{10}^{20}\frac{1}{x^2}\ \mathsf{d}x = \frac{1}{20}$, $c = 20$.

   (b) $\mathrm{E}(X) = \int_{10}^{20} 20x\frac{1}{x^2}\ \mathsf{d}x = 20\big(\log(20) - \log(10)\big) = 20\log(2) \approx 13.86$.

   (c) $F_X(t) = \int_{10}^{t} 20\frac{1}{x^2}\ \mathsf{d}x = 2 - \frac{20}{t}$.

   (d) Hypothèse : les composants fonctionnent indépendemment les uns des autres. On a

   $$P\{\text{une composante fonctionne plus de 15 h}\} = 1 - F_X(15) = 1/3.$$

   Donc la probabilité que, parmi six composants, au moins trois fonctionnent plus de 15 h vaut $1 - P\{0,1,2 \text{ fonctionnent plus de 15 h}\} = 1 - \binom{6}{0}\frac{1^0 2^6}{3^6} - \binom{6}{1}\frac{1^1 2^5}{3^6} - \binom{6}{2}\frac{1^2 2^4}{3^6} \approx 0.320$.

4. (a) X suit une variable aléatoire binomiale de paramètres $n = 5$ et $p = 1/4$.

   (b) On trouve

   $$\begin{aligned} P\{\text{un client se plaigne}\} &= P\{X \ge 2\} = 1 - P\{X = 0\} - P\{X = 1\} \\ &= 1 - \binom{5}{0}\frac{1^0 3^5}{4^5} - \binom{5}{1}\frac{1^1 3^4}{4^5} \approx 0.367. \end{aligned}$$

   (c) Il y aura environ 37 plaintes.

5. (a) Soit X le nombre de tirages effectués, on a $P\{X = 0\} = P\{X = 1\} = P\{X = 6\} = 0$. On suppose que la couleur des boules restant dans l'urne après k tirages, soit noire (si $k = 4, 5$ on peut aussi supposer la couleur blanche, de manière analogue). Donc la couleur du tirage k était blanche. Pour les tirages 1 à $k-1$ on trouve l'autre boule blanche et $k-2$ boules blanches avec toutes les permutations possibles. Donc

k	0	1	2	3	4	5	6
$p_k = P\{X = k\}$	0	0	1/15	2/15	4/15	8/15	0

(b) A l'aide du tableau précédent, on a $\mathrm{E}(X) = \sum_{k=2}^{5} k p_k = 64/15$.

6. (a) X est une variable aléatoire exponentielle de paramètre $\lambda = 0.01$. Donc $\mathrm{E}(X) = 1/\lambda = 100$.

(b) Comme $F_X(t) = 1 - \exp(-0.01t)$, $F_X(200) = 1 - \exp(-2)$ et

$$\begin{aligned} &P\{\text{au moins deux composantes fonctionnent}\} \\ &\quad = P\{\text{2 composantes fonctionnent}\} + P\{\text{3 composantes fonctionnent}\} \\ &\quad = \binom{3}{2}\big(1 - F_X(200)\big)^2 F_X(200) + \binom{3}{3}\big(1 - F_X(200)\big)^3 \\ &\quad = 3(\mathrm{e}^{-2})^2(1 - \mathrm{e}^{-2}) - (\mathrm{e}^{-2})^3 = 3\,\mathrm{e}^{-4} - 2\,\mathrm{e}^{-6} \approx 5\%. \end{aligned}$$

7. (a) Le nombre d'homme ayant récupéré leur propre chapeau est donné par $X = X_1 + \cdots + X_N$, où $X_i = 1$ si le i^{e} homme récupère son chapeau et 0 sinon. Comme chaque homme a autant de chances de ramasser n'importe lequel des N chapeaux, $\mathrm{E}(X_i) = P\{X_i = 1\} = 1/N$ pour tout $i = 1, \ldots, N$. Par conséquent $\mathrm{E}(X) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_N) = 1$.

(b) Nous avons la décomposition

$$\mathrm{Var}(X) = \sum_{i=1}^{N} \mathrm{Var}(X_i) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} \mathrm{Cov}(X_i, X_j)$$

et les variances valent $\mathrm{Var}(X_i) = 1/N(1 - 1/N)$ (variables de Bernoulli). La covariance est $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j)$, à noter que le produit $X_i X_j$ vaut 1 si est seulement si les deux hommes concernés ont pu ramasser leur chapeau, *i.e.* $\mathrm{E}(X_i X_j) = P\{X_i = 1, X_j = 1\} = 1/\big(N(N-1)\big)$. Alors nous trouvons

$$\mathrm{Var}(X) = N\frac{1}{N}\Big(1 - \frac{1}{N}\Big) + 2\frac{N(N-1)}{2}\Big(\frac{1}{N}\frac{1}{(N-1)} - \frac{1}{N^2}\Big) = \frac{N-1}{N} + \frac{1}{N} = 1.$$

8. (a) $\sum_{k=0}^{\infty} P\{X = k\} = \mathrm{e}^{-k}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \mathrm{e}^{-k}\,\mathrm{e}^{k} = 1$.

(b) $\mathrm{E}(X) = \mathrm{e}^{-k}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \mathrm{e}^{-k}\,\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda.$

9. (a) $F_X(t) = \int_0^t \lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x = 1 - \mathrm{e}^{-\lambda t}.$

(b) On trouve

$$\begin{aligned} P\{X > s + t \,|\, X > t\} &= \frac{P\{X > s + t, X > t\}}{P\{X > t\}} = \frac{1 - F_X(t+s)}{1 - F_X(t)} \\ &= \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} = \exp(-\lambda s) = P\{X > s\}. \end{aligned}$$

10. Le skieur reprend la même perche si et seulement si X, le nombre de skieurs qui se présentent, est nN, $n = 1, 2, \ldots$. Comme $P\{X = k\} = p(1-p)^{k-1}$ et $(1-p)^N < 1$, on a

$$\begin{aligned} P\{\text{ le skieur reprend la même perche }\} &= \sum_{n=1}^{\infty} P\{X = nN\} \\ &= p\sum_{n=1}^{\infty}(1-p)^{nN-1}\frac{1-p}{1-p} \\ &= \frac{p}{1-p}\sum_{n=0}^{\infty}\Big((1-p)^N\Big)^n - \frac{p}{1-p} \\ &= \frac{p}{1-p}\cdot\frac{1}{1-(1-p)^N} - \frac{p}{1-p} \\ &= \frac{p}{1-p}\Big(\frac{(1-p)^N}{1-(1-p)^N}\Big). \end{aligned}$$

11. On tire sans remise un échantillon de n boules d'une urne en contenant N, dont Np sont défectueuses et $N-Np$ sont de bonne qualité. Soit X le nombre de pièces défectueuses tirées. Alors X est une variable aléatoire hypergéométrique $n = 3$, $N = 20$ et $p = 1/5$. Donc, en général

$$P\{X = k\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}},$$

et en particulier $P\{X = 0\} = 0.49$, $P\{X = 1\} = 0.42$, $P\{X = 2\} = 0.082$, $P\{X = 3\} = 0.0035$. Nous pouvons alors calculer

$$\mathrm{E}(X) = np = \frac{3}{5} \quad \text{et} \quad \mathrm{Var}(X) = \frac{np(1-p)(N-n)}{(N-1)} \approx 0.429.$$

12. Soit X une variable aléatoire géométrique de paramètre p. Donc la fonction de fréquence est $P\{X = k\} = (1-p)^{k-1}p$, $n \geq 1$. Posons $q = 1-p < 1$. Donc

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} = p\sum_{k=1}^{\infty}\frac{d}{dq}q^k = p\frac{d}{dq}\Big(\sum_{k=1}^{\infty} q^k\Big) \\ &= p\frac{d}{dq}\Big(\frac{q}{1-q}\Big) = \frac{p}{(1-q)^2} = \frac{1}{p}, \\ \mathrm{E}(X^2) &= \sum_{k=1}^{\infty} k^2p(1-p)^{k-1} = p\sum_{k=1}^{\infty} k^2q^{k-1} = p\sum_{k=1}^{\infty} k\frac{d}{dq}q^k = p\frac{d}{dq}\Big(\sum_{k=1}^{\infty} kq^k\Big) \\ &= p\frac{d}{dq}\Big(q\sum_{k=1}^{\infty} kq^{k-1}\Big) = p\frac{d}{dq}\Big(\frac{q}{(1-q)^2}\Big) = \frac{p(1+q)}{(1-q)^3} = \frac{2-p}{p^2}, \\ \mathrm{Var}(X) &= \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}. \end{aligned}$$

13. (a) Si $X_0 = 1$, on a $X_1 = 1, 2, 3$. Si $X_1 = 1$, $X_2 = 1, \ldots, 3$, si $X_1 = 2$, $X_2 = 2, \ldots, 6$ et si $X_1 = 3$, $X_2 = 3, \ldots, 9$. A cause de l'indépendance, on trouve le tableau suivant :

k	$P\{X_2 = k\}$
1	p_1^2
2	$p_1p_2 + p_2p_1^2$
3	$p_1p_3 + 2p_2^2p_1 + p_3p_1$
4	$2p_2p_1p_3 + p_2^3 + 3p_3p_1^2p_2$
5	$2p_2^2p_3 + 3p_3p_1^2p_3 + 3p_3p_1p_2^2$
6	$p_2p_3^2 + p_3p_2^3 + 6p_3p_1p_2p_3$
7	$3p_3p_1p_3^2 + 3p_3p_2^2p_3$
8	$3p_3p_2p_3^2$
9	$p_3p_3^3$

(b) $\mathrm{E}(X_2)$ est obtenue directement à l'aide du tableau ci-dessus.

(c) $P\{X_1 = 2 \,|\, X_2 = 2\} = \dfrac{P\{X_1 = 2 \text{ et } X_2 = 2\}}{P\{X_2 = 2\}} = \dfrac{p_2p_1^2}{p_1p_2 + p_2p_1^2} = \dfrac{p_1}{1 + p_1}$.

14. Posons $Z = -\log(X)$, avec $X \sim U(0,1)$. $g(x) = -\log(x)$ est dérivable et décroissante sur $(0, \infty)$. De plus $g^{-1}(x) = \exp(-x)$. Donc

$$F_Z(z) = 1 - F_X\big(g^{-1}(z)\big) = 1 - \exp(-z).$$

15. $F_Y(y) = P\{X/\lambda \le y\} = P\{X \le \lambda y\} = 1 - \exp(-\lambda y)$, $y \ge 0$.
$f_Y(y) = \dfrac{d}{dy}F_Y(y) = \lambda \exp(-\lambda y)$, $y \ge 0$.

16. Sachant l'identité $\displaystyle\int_0^\infty \exp(-ax)x^n \,\mathrm{d}x = n!\, a^{-n-1}$, on trouve les moments $\mathrm{E}(X) = \displaystyle\int_0^\infty \lambda \exp(-\lambda x)x \,\mathrm{d}x = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$ et $\mathrm{E}(X^2) = \displaystyle\int_0^\infty \lambda \exp(-\lambda x)x^2 \,\mathrm{d}x = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2}$, puis $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X) = \dfrac{2}{\lambda^2} - \dfrac{1}{\lambda^2} = \dfrac{1}{\lambda^2}$.

Remarque : si l'identité n'est pas connue, on peut y échapper par intégrations par parties.

17. Sachant l'identité

$$\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big)x^2 \,\mathrm{d}x = \frac{\sigma^2}{2},$$

on a, en posant $b = 1/2\sigma^2$,

$$\int_0^\infty a \exp(-bx^2)x^2 \,\mathrm{d}x = a\frac{\sigma^2}{2}\sigma\sqrt{2\pi} = a\Big(\frac{1}{\sqrt{2b}}\Big)^3\frac{\sqrt{\pi}}{\sqrt{2}} = a\frac{\sqrt{\pi}}{4b^{3/2}},$$

d'où $a = \dfrac{4b^{3/2}}{\sqrt{\pi}}$, avec $b = m/2kT$.

18. Pour toute variable aléatoire continue X de densité $f_X(x)$, on obtient la fonction de répartition de $Y = X^2$ de la manière suivante : pour $y \geq 0$,

$$F_Y(y) = P\{Y = X^2 \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Une dérivation livre la densité $\frac{1}{2\sqrt{y}}\big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\big)$. On sait que la densité d'une variable aléatoire centrée réduite est $f_X(x) = \varphi(x) = \exp(x^2/2)/\sqrt{2\pi}$. Ainsi

$$f_Y(y) = \frac{1}{2\sqrt{y}} 2\Big(\frac{1}{\sqrt{2\pi}} \exp(-y/2)\Big) = \frac{\frac{1}{2}\exp(-y/2)(\frac{1}{2}y)^{-1/2}}{\sqrt{\pi}},$$

qui est bien la densité d'une loi χ^2_1.

19. Soit $g(x) = \log(x)$. Donc
- $\mathrm{E}(X) \approx g\big(\mathrm{E}(X)\big) = \log\big(\mathrm{E}(X)\big) = \log\Big(\frac{1}{\lambda}\Big) = -\log(\lambda)$.
- $\mathrm{Var}(X) \approx \big(g'(\mathrm{E}(X))\big)^2 \mathrm{Var}(X) = \mathrm{E}(X)^{-2}\frac{1}{\lambda^2} = 1$.

20. $$\int_{-\infty}^{\infty} y f_Y(y)\,\mathsf{d}y = \int_{-\infty}^{\infty} y\,\frac{f_X\big(g^{-1}(y)\big)}{g'\big(g^{-1}(y)\big)}\,\mathsf{d}y = \int_{-\infty}^{\infty} y f_X\big(g^{-1}(y)\big)\frac{\mathsf{d}}{\mathsf{d}y}g^{-1}(y)\,\mathsf{d}y$$
$$= \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathsf{d}x,$$

où la dernière égalité est obtenue en faisant le changement de variables $x = g^{-1}(y)$.

21. La volume d'une bille est $V = 4\pi r^3/3$. Donc :
   (a) Comme $F_R(r) = 3r^2 - 2r^3$, on a

$$F_V(v) = P\{V \leq v\} = P\Big\{r \leq \Big(\frac{3}{4\pi}v\Big)^{1/3}\Big\} = 3\Big(\frac{3}{4\pi}v\Big)^{2/3} - \frac{3}{2\pi}v.$$

   D'où $f_V(v) = 2\Big(\frac{3}{4\pi}\Big)^{2/3} v^{-1/3} - \frac{3}{2\pi}$.

   (b) i. L'espérance exacte vaut

$$\mathrm{E}(V) = \int_0^1 g(r) f_R(r)\,\mathsf{d}r = \int_0^1 \frac{4}{3}\pi r^3 6r(1-r)\,\mathsf{d}r = \frac{8}{5}\pi - \frac{4}{3}\pi = \frac{4\pi}{15}.$$

   On obtient le même résultat avec $\displaystyle\int_0^{4\pi/3} v f_V(v)\,\mathsf{d}v$.

   ii. Comme la distribution de R est symétrique autour de $1/2$, $\mathrm{E}(R) = 1/2$. Donc $\mathrm{E}(V) \approx \frac{4}{3}\pi\big(\mathrm{E}(r)\big)^3 = \frac{4}{3}\pi(1/2)^3 = \frac{\pi}{6} \approx 0.523$.

22. (a) Par symétrie, on trouve

s_i	2, 12	3, 11	4, 10	5, 9	6, 8	7
p_i	1/36	2/36	3/36	4/36	5/36	6/36

   (b) $\mathrm{E}(S) = 7$, car la distribution de S est symétrique autour de 7.

   (c) $\mathrm{E}(X_i) = 7/2$, car $(1+2+3+4+5+6)/6 = 7/2$.

23. Sachant $(\lambda_X + \lambda_Y)^n = \sum_{k=0}^{n} \binom{n}{k} \lambda_X^k \lambda_Y^{n-k} = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_X^k \lambda_Y^{n-k}$, on a

$$P\{X+Y=n\} = \sum_{k=0}^{n} P\{X=k\}P\{Y=n-k\} = \sum_{k=0}^{n} \mathrm{e}^{-\lambda_X} \frac{\lambda_X^k}{k!} \mathrm{e}^{-\lambda_Y} \frac{\lambda_Y^{n-k}}{(n-k)!}$$
$$= \mathrm{e}^{-\lambda_X-\lambda_Y} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} \lambda_X^k \lambda_Y^{n-k} = \mathrm{e}^{-(\lambda_X+\lambda_Y)} \frac{(\lambda_X+\lambda_Y)^n}{n!},$$

d'où $Z = X + Y \sim \mathcal{P}(\lambda_X + \lambda_Y)$.

24. (a) $\mathrm{E}(Y) = 2a_1 + 4a_2 = 15$ et $\mathrm{Var}(Y) = a_1^2/4 + a_2^2 = 9.5625$.

(b) $a_1 X_1 \sim \mathcal{N}(2a_1, a_1^2/4)$, $a_2 X \sim \mathcal{N}(4a_2, a_2^2)$ et $Y \sim \mathcal{N}(2a_1 + 4a_2, a_1^2/4 + a_2^2)$.

(c) $P\{Y > 20\} = 1 - P\{Y \le 20\} = 1 - P\left\{ \frac{Y - 2a_1 - 4a_2}{\sqrt{a_1^2/4 + a_2^2}} \le \frac{20 - 2a_1 - 4a_2}{\sqrt{a_1^2/4 + a_2^2}} \right\} =$
$1 - \Phi\left(\frac{20 - 2a_1 - 4a_2}{\sqrt{a_1^2/4 + a_2^2}} \right)$.

(d) Posons $g(x) = 1/(x^2+1)$, donc $g'(x) = -2x/(x^2+1)^2$.

$$\mathrm{E}(W) \approx g\big(\mathrm{E}(Y)\big) = ((2a_1 + 4a_2)^2 + 1)^{-1},$$
$$\mathrm{Var}(W) \approx \mathrm{Var}(Y) g'\big(\mathrm{E}(Y)\big)^2 = (a_1^2/4 + a_2^2) \frac{-2(2a_1 + 4a_2)}{\big((2a_1 + 4a_2)^2 + 1\big)^2}.$$

(e) La variance vaut

$$\begin{aligned} \mathrm{Var}(Y) &= a_1^2/4 + a_2^2 + 2\,\mathrm{Cov}(a_1 X_1, a_2 X_2) \\ &= a_1^2/4 + a_2^2 + 2a_1 a_2 \,\mathrm{Corr}(X_1, X_2) \sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)} \\ &= a_1^2/4 + a_2^2 + a_1 a_2/2\,. \end{aligned}$$

25. Soit la variable aléatoire Z_i qui vaut 1, si le i^{e} tirage est un pique et zéro sinon. Clairement, Z est Bernoulli de paramètre $p = 1/4$.

(a) Comme $X = \sum_{i=1}^{n} Z_i$, X est une variable aléatoire binomiale de paramètres n et p. Y est une variable aléatoire géométrique, car $P\{Y=k\} = (1-p)^{k-1}p$.

(b) $\mathrm{E}(X) = np = n/4$, $\mathrm{Var}(X) = np(1-p) = 3n/16$.
Selon l'exercice 12, $\mathrm{E}(Y) = 1/p = 4$, $\mathrm{Var}(Y) = (1-p)/p^2 = 12$.

26. (a)

y_i	2	4
p_i	1/2	1/2

(b) Par symétrie de la variable aléatoire X, les moments impairs sont nuls, *i.e.* $\mathrm{E}(X) = \mathrm{E}(X^3) = 0$. Donc $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X^3) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0$.

(c) Les variables aléatoires ne sont pas indépendantes. Seulement dans le cas de normalité, corrélation nulle implique indépendance.

27. Soit $X =$ le nombre d'accidents par équipe de travail. Donc $X \sim \mathcal{P}(\lambda)$.
    (a) $P\{X = 2\} = \mathrm{e}^{-1.5}\,\frac{1.5^2}{2!} \approx 0.251$.
    (b) $P\{X < 2\} = P\{X = 0\} + P\{X = 1\} = \mathrm{e}^{-1.5}\,\dfrac{1.5^0}{0!} + \mathrm{e}^{-1.5}\,\dfrac{1.5^1}{1!} \approx 0.558$.
    (c) $\big(P\{X = 0\}\big)^3 = \big(\mathrm{e}^{-1.5}\big)^3 = \mathrm{e}^{-4.5} \approx 0.011$. On peut supposer que les accidents sont indépendants.
    (d) $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) = 1.5 + 1.5 = 3$ et $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) = 1.5 + 1.5 + 2 \cdot 0.9 = 4.8$.

28. (a) Soit $\mathbf{S} = (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T$. Donc $\mathbf{S}_{ij} = (X_i - \mu_i)(X_j - \mu_j)$ et par conséquent
$$\mathrm{E}(\mathbf{S}_{ij}) = \mathrm{E}\big((X_i - \mu_i)(X_j - \mu_j)\big) = \mathrm{Cov}(X_i, X_j) = \boldsymbol{\Sigma}_{ij}.$$
    Ainsi $\mathrm{E}(\mathbf{S}) = \boldsymbol{\Sigma}$.
    (b) Soit $\boldsymbol{z} \in \mathbb{R}^p$. On a
$$\begin{aligned}\boldsymbol{z}^T\boldsymbol{\Sigma}\boldsymbol{z} &= \boldsymbol{z}^T\,\mathrm{E}\big((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\big)\boldsymbol{z} \\ &= \mathrm{E}\big(\boldsymbol{z}^T(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{z}\big) = \mathrm{E}\big((\boldsymbol{z}^T(\mathbf{X} - \boldsymbol{\mu}))^2\big).\end{aligned}$$
    Puisque $(\boldsymbol{z}^T(\mathbf{X} - \boldsymbol{\mu}))^2 \geq 0$, on en déduit que $\boldsymbol{z}^T\boldsymbol{\Sigma}\boldsymbol{z} \geq 0$, d'où, $\boldsymbol{\Sigma}$ est semi-définie positive.
    (c) On pose $\mathbf{A} = (a_{ij}) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)^T$. De $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$, on a $Y_i = \boldsymbol{a}_i^T\mathbf{X}$, $i = 1, \ldots, p$.
        i. Pour l'espérance on trouve
$$\mathrm{E}(Y_i) = \mathrm{E}\Big(\sum_{j=1}^{p} a_{ji}X_j\Big) = \sum_{j=1}^{p} a_{ji}\,\mathrm{E}(X_j) = \boldsymbol{a}_i^T\,\mathrm{E}(\mathbf{X}), \quad i = 1, \ldots, p.$$
        Donc $\mathrm{E}(\mathbf{Y}) = \mathbf{A}^T\,\mathrm{E}(\mathbf{X})$.
        ii. La variance vaut
$$\begin{aligned}\mathrm{Var}(\mathbf{Y}) &= \mathrm{E}\big((\mathbf{Y} - \mathrm{E}(\mathbf{Y}))(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))^T\big) \\ &= \mathrm{E}\big((\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\boldsymbol{\mu})(\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\boldsymbol{\mu})^T\big) \\ &= \mathrm{E}\big(\mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{A}\big) \\ &= \mathbf{A}^T\,\mathrm{E}\big((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\big)\mathbf{A} = \mathbf{A}^T\,\mathrm{Var}(\mathbf{X})\mathbf{A}.\end{aligned}$$

29. On sait que la densité de X_1 et X_2 est $\lambda\exp(-\lambda x)$. En utilisant la formule pour la convolution (page 80(1997) ou 81(2001)), on trouve :
    (a) $f_{X_1+X_2}(z) = \lambda^2 \displaystyle\int_0^z \exp\big(-\lambda(z - x)\big)\exp(-\lambda x)\,\mathrm{d}x$
$$= \lambda^2\exp(-\lambda z)\int_0^z \mathrm{d}x = \lambda^2\exp(-\lambda z)z.$$
    (b) $f_{X_1+X_2+X_3}(z) = \lambda^3 \displaystyle\int_0^z \exp\big(-\lambda(z - x)\big)x\exp(-\lambda x)\,\mathrm{d}x$
$$= \lambda^3\exp(-\lambda z)\int_0^z z\,\mathrm{d}x = \lambda^3\exp(-\lambda z)\frac{z^2}{2}.$$

(c) L'intégrant de la convolution pour l'étape $n+1$ est x^n, qui augmente, au cours de l'intégration, le dénominateur et la puissance de x, la densité de la loi Gamma de paramètres $p = n$ et λ est

$$\frac{\lambda^n}{(n-1)!}\, x^{n-1} \exp(-\lambda x).$$

Chapitre 6

Modèles statistiques et estimation de paramètres

1. Le biais de $\hat{\theta} = \max(Y_1, \ldots, Y_n)$ est $-\theta/(n+1)$, c'est-à-dire que $\mathrm{E}_\theta(\hat{\theta}) = \theta/(n+1)$, l'estimateur $\theta^* = (n+1)/n \cdot \hat{\theta}$ est alors non biaisé.

2. (a) $\mathrm{E}(T) = (\alpha+\beta)\theta$, c'est-à-dire T est non biaisé pour toutes les valeurs de α et β qui satisfont $\alpha + \beta = 1$.
   (b) $\mathrm{Var}(T) = \alpha^2\sigma^2 + \beta^2\sigma^2 = (2\alpha^2 - 2\alpha + 1)\sigma^2$, en dérivant par rapport à α et en annulant on trouve $\alpha = \beta = 1/2$.

3. L'espérance et la variance d'une loi de Poisson sont égales au paramètre λ. Le biais de $\hat{\lambda}$ est alors $b_{\hat{\lambda}}(\lambda) = \lambda - \lambda = 0$ et la variance vaut $\mathrm{Var}(\hat{\lambda}) = \lambda/n$.

4. (a) Nous avons

$$\begin{aligned}(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 &= Y_1^2 + Y_2^2 - 2\bar{Y}(Y_1 + Y_2) + 2\bar{Y}^2 \\ &= Y_1^2 + Y_2^2 - 2\bar{Y}2\bar{Y} + 2\bar{Y}^2 = Y_1^2 + Y_2^2 - 2\bar{Y}^2 \\ &= Y_1^2 + Y_2^2 - 2\frac{(Y_1+Y_2)^2}{4} = \frac{1}{2}Y_1^2 + \frac{1}{2}Y_2^2 - Y_1Y_2 \\ &= \frac{1}{2}(Y_1 - Y_2)^2.\end{aligned}$$

   (b) $\mathrm{Var}(Y_1 - Y_2) = \mathrm{E}\big((Y_1 - Y_2)^2\big)$, car $\mathrm{E}(Y_1 - Y_2) = 0$. D'autre part, l'indépendance de Y_1 et Y_2 entraine que $\mathrm{Var}(Y_1 - Y_2) = 2\sigma^2$.
   (c) $\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\big(1/2 \cdot 1/2(Y_1 - Y_2)^2\big) = 1/4 \cdot 2/\sigma^2 = \sigma^2/2$ et $\mathrm{E}(s^2) = \mathrm{E}(1/2(Y_1 - Y_2)^2) = 1/2 \cdot 2/\sigma^2 = \sigma^2$.

5. (a) Non, la moyenne est non biaisée comme estimateur de l'espérance.
   (b) $\mathrm{CME}_g(\hat{\mu}_2) = \mathrm{Var}_g(\hat{\mu}_2) + b_{\hat{\mu}_2}(g)^2 = 1/2 \cdot \sigma^2$ et $\mathrm{CME}_g(\hat{\mu}_3) = 1/3 \cdot \sigma^2$.
   (c) Il devrait choisir μ_3, car tous les deux sont non biaisés, mais μ_3 à un CME plus petit.

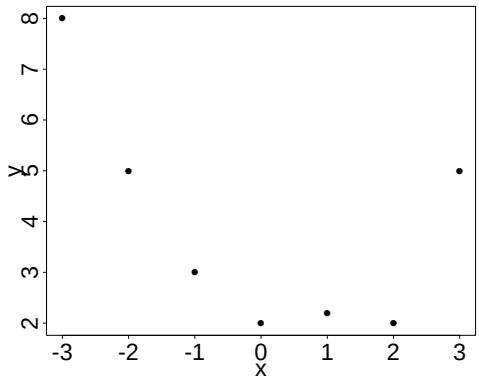
6. (a) Comme $\mathrm{E}(\hat{\mu}) = \mu \sum a_i$, la condition sur les a_i est $\sum a_i = 1$.
   (b) Nous avons

$$\mathrm{CME}_\mu(\hat{\mu}) = \mathrm{Var}_\mu(\hat{\mu}) + b_{\hat{\mu}}(\mu)^2 = \sigma^2 \sum a_i^2 \mu^2 \Big(\sum a_i - 1 \Big)^2.$$

   (c) La fonction à minimiser est $F(a_1, \ldots, a_n, \lambda) = \sigma^2 \sum a_i^2 - \lambda\mu\big(\sum a_i - 1\big)$. En dérivant par rapport à a_i et en annulant nous trouvons $a_i = \lambda\mu/(2\sigma^2)$. En remplaçant ceci dans l'équation qui annule la dérivée de F par rapport à λ nous obtenons $\mu \sum \lambda\mu/(2\sigma^2) = \mu$, donc $\lambda = 2\sigma^2/(n\mu)$ et $a_i = 1/n$.

Chapitre 7
Méthodes d'estimation

1. (a) Selon le graphique, on propose un modèle quadratique.



   (b) $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j)$.

   (c) On ajuste $y_i = \alpha + \beta x_i^2 + \varepsilon_i$, $i = 1, \ldots, n$, où $n = 7$, avec la méthode de la régression linéaire.

   (d) $(\alpha, \beta)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = (1.91, 0.49)^T$.

   (e) Les résidus sont $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (0.09, 1.65, 1.11, 0.59, -0.21, -1.89, -1.35)^T$ et leur somme vaut 0.

   (f) $\hat{y}_i = 1.91 + 0.49 \cdot 4^2 = 9.75$.

   Remarque : En ajustant le modèle $y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $i = 1, \ldots, n$, où $n = 7$, avec la méthode de la régression linéaire multiple ($p = 2$), on obtiendrait

   (d) $(\alpha, \beta_1, \beta_2)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = (1.91, -0.56, 0.49)^T$.

   (e) Les résidus sont $\boldsymbol{r} = (0.09, -0.04, -0.01, 0.03, 0.36, -0.76, 0.34)^T$ et leur somme vaut 0.

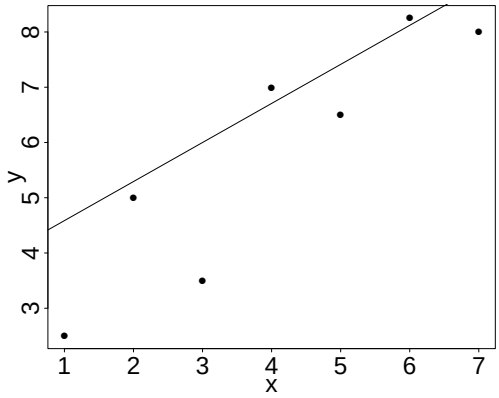
   (f) $\hat{y}_i = 1.91 - 0.56 \cdot 4 + 0.49 \cdot 4^2 = 7.54$.

2. (a) La droite cherchée est d'équation : $\hat{y}_i - y_0 = \hat{\beta}(x_i - x_0)$, $i = 1, \ldots, n$, et on veut minimiser $C(\beta) = \sum_{i=1}^n \big(y_i - y_0 - \beta(x_i - x_0)\big)^2$. Donc

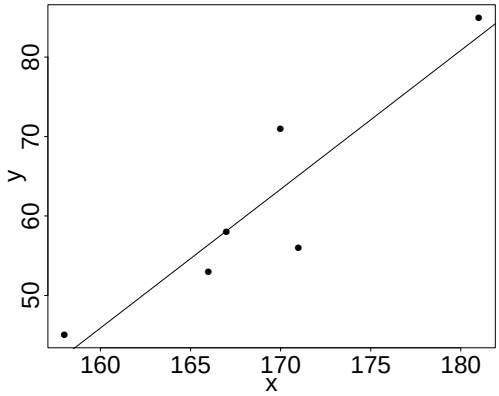
$$\frac{\partial C(\beta)}{\partial \beta}\Big|_{\hat{\beta}} = 0 \Longleftrightarrow \hat{\beta} = \frac{\sum_{i=1}^n (x_i - x_0)(y_i - y_0)}{\sum_{i=1}^n (x_i - x_0)^2}.$$

   (b) $(x_0, y_0) = (0, 0)$ et donc $\hat{\beta} = \sum_{i=1}^n x_i y_i \Big/ \sum_{i=1}^n x_i^2$. Comme la régression non-forcée passe toujours par le centre de gravité $P(\bar{x}, \bar{y})$, les équations ne changent pas.

(c) $\hat{\beta} = (99/4)/35 = 0.7071$ et la droite ajustée est d'équation $y = 3.9 + 0.71x$.



3. (a) Avec la méthode des moindres carrés, on trouve l'équation de la droite $y = -233.75 + 1.75x$.



(b) Selon le schéma sur la page 115(1997) ou 118(2001) en bas, on trouve :

source	SC	d.l.	CM
modèle	$\mathrm{SC_m} = 23434.64$	2	$\mathrm{SC_m}/2$
erreurs	$\mathrm{SC_e} = 165.34$	4	$\mathrm{SC_e}/4$
total	$\mathrm{SC_t} = 23600.00$	6	

La version alternative est (schéma sur la page 116(1997) ou 118(2001) en bas) :

source	SC	d.l.	CM
régression	$\mathrm{SC_r} = 863.99$	1	$\mathrm{SC_r}$
erreurs	$\mathrm{SC_e} = 165.34$	4	$\mathrm{SC_e}/4$
total (corr)	$\mathrm{SC_t} = 1029.33$	5	

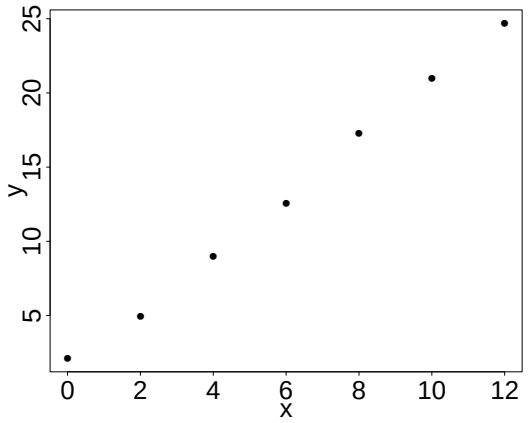
(c) $\hat{\sigma}^2 = \mathrm{CM_e} = 165.34/4 = 41.34$.

(d) $R^2 = \mathrm{SC_r} / \mathrm{SC_t} = 0.84$.

4. (a) On a le modèle $y_i = \tilde{\alpha} + \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, où $\tilde{\alpha} = \alpha - \beta\bar{x}$. Donc l'estimateur des moindres carrés de $\hat{\beta}$ est le même, de plus, $\hat{\alpha} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} = \bar{y}$.

(b) $\hat{\beta} = S_{XY}/S_{XX} = 0.270602$ et $\hat{\alpha} = 41.25$.

5. (a) Selon le graphique suivant, on propose un modèle linéaire.



(b) Avec la méthode des moindres carrés, on trouve l'équation de la droite $y = 1.518 + 1.930x$.

(c) Selon le schéma sur la page 116(1997) en bas ou 118(2001) en bas, on trouve :

source	SC	d.l.	CM
régression	$\mathrm{SC_r} = 417.3432$	1	$\mathrm{SC_r}$
erreurs	$\mathrm{SC_e} = 0.9368$	5	$\mathrm{SC_e}/5$
total (corr.)	$\mathrm{SC_t} = 418.28$	6	

$R^2 = \mathrm{SC_r} \,/\, \mathrm{SC_t} = 0.998$.

6. Pour la loi de Poisson $P\{Y = y\} = \exp(-\lambda)\lambda^y/y!$ on trouve

$$V(\lambda) = \exp(-\lambda 5) \prod_{i=1}^{5} \frac{\lambda^{y_i}}{y_i!},$$

$$L(\lambda) = -\lambda 5 + \sum_{i=1}^{5} y_i \log(\lambda) - \sum_{i=1}^{5} \log(y_i!).$$

En dérivant et annulant, on trouve $\hat{\lambda}_{MV} = \bar{y} = 25/6$.

7. Y est une variable aléatoire binomiale négative de paramètre r et p. Donc

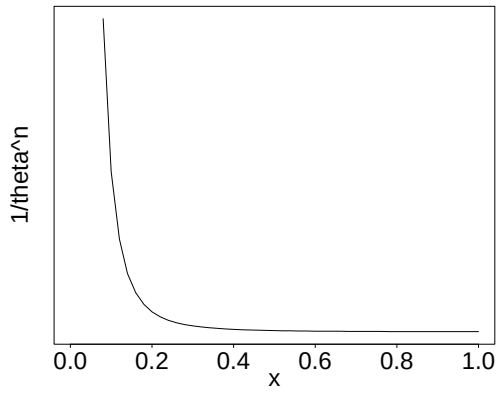
$$V(p) = \prod_{i=1}^{n} P\{Y = k\} = \prod_{i=1}^{n} \binom{k_i - 1}{r - 1} p^r (1-p)^{k_i - r},$$

$$L(p) = \sum_{i=1}^{n} \log\left(\binom{k_i - 1}{r - 1}\right) + nr \log(p) + \sum_{i=1}^{n} (k_i - r) \log(1-p).$$

En dérivant et annulant, on trouve $\hat{p}_{MV} = nr / \sum_{i=1}^{n} k_i = r/\bar{k}$.

8. (a) $V(\theta) = \prod_{i=1}^{n} \theta^{-1} \mathbf{1}_{(0,\theta]} = \theta^{-n}$.

(b) La fonction $1/\theta^n$ n'a pas de maximum dans $[0,1]$. Mais clairement $V(0) > V(\theta)$, $\theta > 0$.



(c) Selon le raisonnement précédent, on devrait choisir $\hat{\theta}_{MV} = 0$, qui est totalement "non sense".

9. (a) Pour $f(y) = \theta/2 \exp(-\theta y)(\theta y)^2$ on trouve

$$V(\theta) = \frac{\theta^n}{2^n} \prod_{i=1}^{n} (\theta y_i)^2 \exp(-\theta y_i),$$

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

En dérivant et annulant, on trouve $\hat{\theta}_{MV} = 3/\bar{y}$.
Pour la loi de Poisson $P\{X = z\} = \exp(-2\theta)(2\theta)^z/z!$ on trouve

$$V(\theta) = \exp(-2\theta m) \prod_{i=1}^{m} \frac{(2\theta)^{z_i}}{z_i!},$$

$$L(\theta) = -2\theta m + \sum_{i=1}^{m} z_i \log(2\theta) - \sum_{i=1}^{m} \log(z_i!).$$

En dérivant et annulant, on trouve $\hat{\theta}_{MV} = \bar{z}/2$.

(b) On a égalité si $\bar{z} = 6/\bar{y}$ est satisfait.

10. (a) Pour une loi exponentielle $\mathcal{E}(\lambda)$, on trouve

$$V(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big),$$

$$L(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i.$$

En dérivant et annulant, on trouve $\hat{\lambda}_{MV} = \hat{\lambda} = 1/\bar{x}$.

(b) $\bar{x} = 92.96$, donc $\hat{\lambda} = 0.01$.

Chapitre 8

Tests statistiques

1. Comme σ est inconnu, on applique un test t de Student avec comme hypothèse nulle $H_0 : \mu = 8.5$ et comme alternative $H_1 : \mu \neq 8.5$. Sous H_0, $t = \sqrt{n}(\bar{Y} - 8.5)/s \sim t_{n-1}$ et nous observons $t_{\text{obs}} = \sqrt{100}\,(9.25 - 8.5)/\sqrt{1.35} = 6.455 > qt_{99}(97.5\%)$, c'est-à-dire qu'on peut rejeter l'hypothèse H_0.

2. (a) $\mathrm{E}(y_i) = \mathrm{E}(\mu + \varepsilon_i) = \mathrm{E}(\mu) + \mathrm{E}(\varepsilon_i) = \mu$, $\mathrm{Var}(y_i) = \mathrm{Var}(\mu + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$.

   (b) $C(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2$ est à minimiser, la dérivée de C par rapport à μ vaut $-2\sum_{i=1}^{n} y_i + 2\,n\,\mu$, en annulant nous trouvons $\hat{\mu} = 1/n \sum_{i=1}^{n} y_i = \bar{y}$.

   (c) Comme σ est inconnu, on applique un test t de Student avec comme hypothèse nulle $H_0 : \mu = 0$ et comme alternative $H_1 : \mu > 0$. Sous H_0, $t = \sqrt{n}(\bar{Y} - 0)/s \sim t_{n-1}$ avec $s^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$. A partir des données, on calcule la valeur observée t_{obs} de la statistique t et ensuite on compare cette valeur au quantile 95% de la loi t avec $n-1$ degrés de liberté. Si t_{obs} est supérieure à ce quantile, on rejette l'hypothèse nulle H_0.

   (d) Dans le cas de $H_1 : \mu \neq 0$, on compare $|t_{\text{obs}}|$ au quantile 97.5% de la loi t avec $n-1$ degrés de liberté. Si $|t_{\text{obs}}|$ est supérieure à ce quantile, on rejette l'hypothèse nulle H_0.

   (e) Comme valeurs numériques, on trouve $\bar{y} = -0.027$, $s^2 = 0.127$ et $t_{\text{obs}} = -0.201$. Donc, comme $qt_6(95\%) = 1.943$, $qt_6(97.5\%) = 2.447$, l'hypothèse nulle H_0 ne peut pas être rejetée dans les deux cas.

   (f) Pour le cas univarié on trouve la p-valeur $= P\{t > t_{\text{obs}}\} = P\{t > -0.221\} = 0.576$ et pour le cas bivarié : p-valeur $= P\{\,|\,t\,|\, > 0.221\} = 0.848$.

3. Les échantillons sont appariés, donc posons $d_i = y_i - x_i$ et $\Delta = \mathrm{E}(d_i)$. On trouve les valeurs numériques suivantes : $\bar{d} = -0.113$ et $s_d^2 = 0.037$. Nous allons tester l'hypothèse nulle $H_0 : \Delta \leq 0$ contre l'alternative $H_1 : \Delta \leq 0$. Sous l'hypothèse nulle, $t = \sqrt{n}\,\bar{d}/s_d \sim t_{14}$ et nous observons $|t_{\text{obs}}| = \sqrt{5}\,(0.113)/\sqrt{0.037} = 2.266 > qt_{14}(97.5\%) = 1.761$, donc l'hypothèse nulle doit être rejetée.

4. (a) Les hypothèses $H_0 : \mu = 365$, $H_1 : \mu < 365$ sont à tester à l'aide de la statistique de test $t = \sqrt{7}\,(\bar{Y} - 365)/s$, sous l'hypothèse nulle $t \sim t_6$ et on observe $t_{\text{obs}} = \sqrt{7}\,(354.6 - 365)/\sqrt{251.8} = -1.734 > qt_6\,(5\%) = -1.943$, donc l'hypothèse nulle ne peut pas être rejetée.

   (b) "H_0 vraie et rejetée" correspond à l'erreur de première espèce = 5%. La proportion de laboratoires qui observeront une évidence plus forte contre H_0 est égal à la p-valeur $= P\{t < t_{\text{obs}}\} = P\{t < -1.734\} = 6.7\%$.

5. Les échantillons sont non appariés avec $\hat{\Delta} = \bar{y} - \bar{x} = -2.9$ et $s_p^2 = (12\,s_x^2 + 22\,s_y^2)/23 = 69.930$. Les hypothèses $H_0 : \mu_x = \mu_y$, $H_1 : \mu_x \neq \mu_y$ sont à tester à l'aide de la statistique de test $t = \sqrt{13 \cdot 12}\,(\bar{Y} - \bar{X})/\big(s_p\,\sqrt{(13+12)}\big)$ qui est distribuée selon une loi t_{23} sous l'hypothèse nulle. On observe $|t_{\text{obs}}| = |-0.8663| < qt\,(97.5\%) = 2.069$, donc l'hypothèse nulle ne peut pas être rejetée.

6. (a) Les échantillons sont non appariés avec $\hat{\Delta} = \bar{y} - \bar{x} = -73$ et $s_p^2 = (11 s_x^2 + 5 s_y^2)/16 = 4490.25$. Les hypothèses $H_0 : \mu_x = \mu_y$, $H_1 : \mu_x \neq \mu_y$ sont à tester à l'aide de la statistique de test $t = \sqrt{12 \cdot 6}\,(\bar{Y} - \bar{X})/(s_p\,\sqrt{(12+6)})$ qui est distribuée selon une loi t_{16} sous l'hypothèse nulle. On observe $|t_{\text{obs}}| = |-2.179| > qt_{16}(97.5\%) = 2.12$, donc l'hypothèse nulle doit être rejetée.

   (b) $X \sim \mathcal{N}(\mu_x, \sigma^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma^2)$ et $\text{Cov}(X, Y) = 0$.

7. (a) $\overline{Y}_5$ est distribuée selon la loi normale avec espérance $\text{E}(\overline{Y}_5) = \sum_{i=1}^{5} \text{E}(Y_i/5) = \mu = 0$ et variance $\text{Var}(\bar{Y}_5) = \sum_{i=1}^{5} \text{Var}(Y_i)/25 = 1/5$.

   (b) Comme σ est connu, on applique un test z avec comme hypothèse nulle $H_0 : \mu = 0$ et comme alternative $H_1 : \mu \neq 0$. Sous H_0, $z = \sqrt{5}\;\bar{y}_5/\sigma \sim \mathcal{N}(0,1)$ et nous observons $|z_{\text{obs}}| = |\sqrt{5}\,0.649| = 1.450 < 1.96$, c'est-à-dire qu'on peut rejeter l'hypothèse H_0. La p-valeur est $2 - 2\Phi(1.450) = 14.8\%$.

   (c) $\overline{Y}_{10} \sim \mathcal{N}(9, 1/10)$.

   (d) $\bar{y}_{10} = 1.111$, de la même façon que sous (b) on obtient $|z_{\text{obs}}| = |\sqrt{10}\;1.111| = |3.512| > 1.96$, donc l'hypothèse nulle doit être rejetée

   (e) On a commis une erreur de deuxième espèce : on n'a pas rejeté, alors que l'hypothèse nulle est en réalité fausse. La raison est que la détection d'une alternative proche de H_0 est plus facile si n est grand.

8. (a) $\sqrt{10}\,\overline{Y}/s \sim t_9$ sous l'hypothèse H_0.

   (b) $\bar{y} = 1.111$, $s^2 = 0.927$ et $t_{\text{obs}} = \sqrt{10}\,\bar{y}/s = 3.648$, comme la statistique de test $t = \sqrt{10}\,\bar{Y}/s$ est distribuée selon la loi t_9 et $qt_9(97.5\%) = 2.262 < |3.648|$, l'hypothèse nulle doit être rejetée.

9. Les échantillons sont appariés avec $d_i =$ régime $\text{II}_i -$ régime I_i, $\text{E}(d_i) = \Delta$, d'où $\bar{d} = 10.58$ et $s_d^2 = 149.719$. La statistique de test $t = \sqrt{12}\,\bar{d}/s_d \sim t_{11}$, sous l'hypothèse nulle $H_0 : \Delta = 0$. Car $|t_{\text{obs}}| = |2.996| > qt_{11}\,(97.5\%) = 2.201$, l'hypothèse nulle doit être rejetée.

10. L'hypothèse nulle $H_0 : \mu = 3.1$, les athlètes est-allemandes ne sont pas dopées, est testée à l'aide de la statistique de test $t = \sqrt{9}\,(\bar{Y} - 3.1)/s$, qui est distribuée selon une loi t_8 sous l'hypothèse nulle. L'hypothèse alternative est $H_1 : \mu > 3.1$ et nous observons $\bar{y} = 3.257$, $s^2 = 0.055$ et $t_{\text{obs}} = 1.996 > qt_8(95\%) = 1.860$. Donc l'hypothèse nulle doit être rejetée.

Chapitre 9

Intervalle de confiance

1. (a) $\bar{Y} - \mu \sim \mathcal{N}(0, c^2/n)$.
   (b) L'intervalle cherché est $\left[\bar{Y} \pm c \cdot q\mathcal{N}(95\%)/\sqrt{n}\right]$.
   (c) La longueur vaut $2c \cdot q\mathcal{N}(95\%)/\sqrt{n} = 3.3\, c/\sqrt{n}$.
   (d) Plus n augmente, plus l'intervalle de confiance devient petit, c'est-à-dire l'estimation devient plus précise avec plus d'observations. Plus c diminue, plus l'intervalle de confiance devient petit, ce qui est logique car les mesures sont plus précises.

2. (a) L'intervalle de confiance est $\left[\bar{Y} \pm s \cdot qt_{n-1}(95\%)/\sqrt{n}\right]$.
   (b) La longueur vaut $2s \cdot qt_{n-1}(95\%)/\sqrt{n}$.
   (c) $q\mathcal{N}(95\%) < qt_{n-1}(95\%)$, donc l'intervalle de confiance le plus précis est celui de l'exercice précédent. Lorsque n devient grand le quantile de la loi de Student va s'approcher du quantile de la loi normale.

3. (a) $\bar{y} = 1.02$ et $s = 0.264$.
   (b) L'intervalle de confiance cherché vaut $[0.82, 1.23]$.

4. L'intervalle cherché vaut $\left[\bar{Y} \pm c \cdot q_{\text{normale}}(99.5\%)/\sqrt{n}\right] = [99.172, 102.628]$.

5. (a) $Y_1, \ldots, Y_{25}$ i.i.d. selon la loi normale de paramètres μ et σ^2.
   (b) $\hat{\mu} = \bar{Y} = 1.61$ et $\hat{\sigma}^2 = s^2 = 0.090$.
   (c) $[\bar{Y} \pm s \cdot qt_{24}(97.5\%)/\sqrt{n}] = [1.49, 1.74]$.

6. L'intervalle de confiance pour μ_A vaut $\left[\bar{x} \pm s \cdot qt_{n-1}(97.5\%)/\sqrt{n}\right] = [28.78, 31.22]$, avec $\bar{x} = 30$ et $s_x^2 = 2.89$. Pour le deuxième échantillon on trouve $\bar{y} = 28$ et $s_y^2 = 2.86$. La statistique de test t de Student pour deux échantillons vaut alors :

$$\frac{\bar{y} - \bar{x}}{\sqrt{s_p^2(n+m)/(nm)}} = -2.487, \quad \text{où} \quad s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} = 2.87.$$

   Comme $-2.487 < -2.12 = qt_{16}(2.5\%)$ on doit rejeter l'hypothèse nulle de l'égalité des deux moyennes.

7. La pente β est estimée par $\hat{\beta} = \sum_{i=1}^n y_i(x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2 = 1.93$ et l'intervalle de confiance pour β vaut $\left[\hat{\beta} \pm \hat{\sigma} \cdot qt_5(97.5\%)/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = [1.83, 2.04]$, où $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2) = 0.433$.

8. (a) Avec $t_i = x_i^2$ nous trouvons $\hat{\alpha} = \sum_{i=1}^n y_i(t_i - \bar{t})/\sum_{i=1}^n (t_i - \bar{t})^2$ et $\hat{\mu} = \bar{y} - \hat{\alpha}\bar{t}$.
   (b) $\hat{\alpha} = -0.017$, $\hat{\mu} = 1.153$ et $\hat{\sigma}^2 = \mathrm{SC_e}/(n-2) = 22.329$.
   (c) L'intervalle de confiance pour α vaut $\big[\,\hat{\alpha} \pm \hat{\sigma} \cdot qt_5(97.5\%)/\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2}\,\big] = [-1.342, 1.309]$, il se base sur l'hypothèse que les données sont i.i.d. et distribuées selon une loi normale.
   (d) On constate que les points sont à peu près alignés sur une droite, un meilleur modèle est celui de la régression simple $y_i = \mu + \alpha x_i + \varepsilon_i$.

9. (a) Les estimateurs de α et β_2 valent $\hat{\beta}_2 = \sum_{i=1}^n y_i(x_{i2} - \bar{x}_2)/\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 = -5.123$, $\hat{\alpha} = \bar{y} - \hat{\beta}_2 \bar{x} = 321.314$.
   (b) L'intervalle de confiance pour β_2 est $\big[\,\hat{\beta}_2 \pm \hat{\sigma} \cdot qt_{10}(97.5\%)/\sqrt{\sum_{i=1}^n (x_{i2} - \bar{x}_2)^2}\,\big] = [-7.371, -2.875]$, où $\hat{\sigma} = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2) = 25.462$.

10. Les observations $Y_1, \ldots, Y_{12}$ sont supposées i.i.d. selon une loi normale de paramètres μ et σ^2. L'intervalle de confiance pour μ est $[\,\bar{Y} \pm s \cdot qt_{11}(97.5\%)/\sqrt{n}\,] = [12.42, 14.88]$, où $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$.

11. (a) $Y_i \sim \mathcal{E}(\lambda)$ avec $\mathrm{E}(Y_i) = 1/\lambda = \mu$, une estimation de μ est $\hat{\mu} = \bar{Y}$. Les variables λY_i sont distribuées selon la loi $\mathcal{E}(1)$ et la variable $\lambda \sum_{i=1}^n Y_i$ selon la loi Gamma de paramètres n et 1 (*i.e.* $\mathcal{G}(n,1)$). Nous avons alors $90\% = P\big\{c_1 \le \lambda \sum_{i=1}^n Y_i \le c_2\big\} = P\big\{\sum_{i=1}^n Y_i/c2 \le \mu \le \sum_{i=1}^n Y_i/c_2\big\}$, où $c_1 = q\mathcal{G}_{n,1}(5\%)$, $c_2 = q\mathcal{G}_{n,1}(95\%)$ et nous trouvons donc l'intervalle de confiance de niveau 90% pour μ suivant :

$$\left[\, \frac{\sum_{i=1}^n Y_i}{q\mathcal{G}_{n,1}(95\%)}, \frac{\sum_{i=1}^n Y_i}{q\mathcal{G}_{n,1}(5\%)} \,\right].$$

   (b) $\sum_{i=1}^{40} Y_i = 3724.79$, $q\mathcal{G}_{n,1}(5\%) = 39.67$ et $q\mathcal{G}_{n,1}(95\%) = 50.94$, donc nous trouvons l'intervalle de confiance pour μ suivant $[73.12, 93.90]$.

12. (a) Comme $Y \sim \mathcal{B}(1000, p)$ et $\hat{p} = Y/1000$, $\mathrm{E}(\hat{p}) = \mathrm{E}(Y)/1000 = p$ et $\mathrm{Var}(\hat{p}) = \mathrm{Var}(Y)/1000^2 = p(1-p)/1000$.
   (b) Il suffit de montrer que $\mathrm{E}(T) = 0$ et $\mathrm{Var}(T) = 1$, ce qui est juste d'après (a).
   (c) Notons $z^2 = q\mathcal{N}^2(97.5\%)$, $P\big\{T^2 \le z^2\big\} = P\big\{p_1(\hat{p}) \le p \le p_2(\hat{p})\big\}$, où

$$p_{1/2}(\hat{p}) = \frac{2000\hat{p} + z^2 \pm z\sqrt{4000\hat{p} - 4000\hat{p}^2 + z^2}}{2000 + 2z^2}.$$

   (d) Numériquement l'intervalle obtenu sous (c) vaut $[0.490, 0.552]$.
   (e) L'intervalle de confiance cherché est de la forme

$$\left[\, \hat{p} \pm z\sqrt{\frac{1}{1000}\hat{p}(1-\hat{p})} \,\right].$$

(f) Numériquement nous obtenons $[0.49, 0.552]$, ce qui est exactement le même intervalle que sous (d) (au moins si on arrondit à trois décimales).

13. (a) La densité de la loi Gamma de paramètres n et λ, $\mathcal{G}(n,\lambda)$, est $\lambda/\Gamma(n)(\lambda x)^{n-1} \cdot \exp(-\lambda x)$, nous avons alors

$$P\{Z \le x\} = P\Big\{Y \le \frac{x}{k}\Big\} = \int_0^{x/k} \frac{\lambda}{\Gamma(n)}(\lambda u)^{n-1}\exp(-\lambda u)\ \mathsf{d}u$$
$$= \int_0^{x} \frac{\lambda}{k\Gamma(n)}\Big(\frac{\lambda}{k}t\Big)^{n-1}\exp\Big(-\frac{\lambda}{k}t\Big)\ \mathsf{d}t.$$

La dernière égalité a été obtenue avec la substitution $t = ku$, la fonction à l'intérieur de la dernière intégrale est la densité de la loi Gamma de paramètres n et λ/k, alors $Z \sim \mathcal{G}(n, \lambda/k)$.

(b) $\lambda Y \sim \mathcal{G}(n,1)$ car

$$P\{\lambda Y \le x\} = P\Big\{Y \le \frac{x}{\lambda}\Big\} = \int_0^{x/\lambda} \frac{\lambda}{\Gamma(n)}(\lambda u)^{n-1}\exp(-\lambda u)\ \mathsf{d}u$$
$$= \int_0^{x} \frac{1}{\Gamma(n)}t^{n-1}\exp(-t)\ \mathsf{d}t,$$

avec la substitution $u = \lambda x$. D'après (a) nous avons $2\lambda Y \sim \mathcal{G}(n, 1/2) = \chi^2_{2n}$.

(c) Pour $Y \sim \mathcal{G}(n,\lambda)$ un intervalle de confiance de niveau 90% pour λ est

$$\Big[\ \frac{q(5\%)}{2Y}, \frac{q(95\%)}{2Y}\ \Big],$$

où $q(5\%)$ et $q(95\%)$ sont les quantiles de la loi χ^2_{2n}.

14. Comme $\bar{Y} \sim \mathcal{N}(\theta, 1/n)$ nous trouvons

$$P\Big\{\bar{Y} - \frac{2}{\sqrt{n}} \le \theta \le \bar{Y} + \frac{2}{\sqrt{n}}\Big\} = P\big\{-2 \le \sqrt{n}(\bar{Y} - \theta) \le 2\big\} = \Phi(2) - \Phi(-2) = 0.954\,.$$

15. (a) $\mathrm{E}(\hat{\beta}) = \frac{1}{\sum_{i=1}^n x_i^4}\sum_{i=1}^n x_i^2\,\mathrm{E}(Y_i) = \frac{1}{\sum_{i=1}^n x_i^4}\sum_{i=1}^n x_i^2 \beta x_i^2 = \beta$, alors $\hat{\beta}$ est non biaisé.

(b) $\hat{\beta} = 0.780$.

(c) $\mathrm{Var}(\hat{\beta}) = \big(\sum_{i=1}^n x_i^4\big)^{-2}\sum_{i=1}^n x_i^4\,\mathrm{Var}(Y_i) = a^2\big/\sum_{i=1}^n x_i^4$, alors la loi de $\hat{\beta}$ est $\mathcal{N}\big(\beta, a^2\big/\sum_{i=1}^n x_i^4\big)$ donc le pivot cherché est

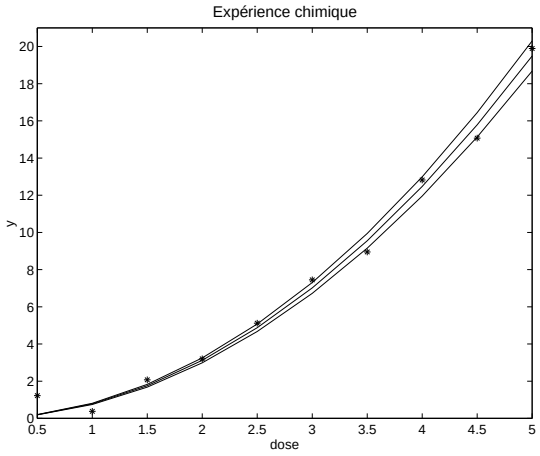
$$\sqrt{\sum\nolimits_{i=1}^n x_i^4}\ \frac{\hat{\beta} - \beta}{a} \sim \mathcal{N}(0,1)$$

et l'intervalle de confiance pour β est

$$[\,B_I, B_S\,] = \Big[\,\hat{\beta} \pm \frac{a}{\sqrt{\sum_{i=1}^n x_i^4}}\ q_{\text{normale}}(99.5\%)\,\Big].$$

(d) $[\,b_I, b_S\,] = [\,0.747, 0.812\,]$.

(e) Voici le graphique représentant les données, la courbe ajustée et ses "courbes de confiance" :



16. (a) $\hat{\beta} = \sum_{i=1}^n y_i(x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2 = 1.307$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 0.432$.

(b) $\hat{y}(-1) = 0.432 - 1.307 = -0.875$.

(c) Avec $t_i = x_i^2$ nous trouvons $\hat{\beta} = \sum_{i=1}^n y_i(t_i - \bar{t}) / \sum_{i=1}^n (t_i - \bar{t})^2 = 0.326$ et $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t} = 1.089$, dans le nouveau modèle la prédiction vaut $\hat{y}(-1) = 1.416$.

(d) La statistique de test pour tester $H_0 : \beta = 1/3$ est

$$t\text{-pente}_{\text{obs}} = \frac{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2}(\hat{\beta} - 1/3)}{\hat{\sigma}} = -0.3396,$$

où $\hat{\sigma} == \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2) = 0.0804$. Comme $|-0.3396| < 3.182 = qt_3(97.5\%)$ on ne peut pas rejeter l'hypothèse nulle.

17. (a) $\hat{\beta} = 20.28$, $\hat{\alpha} = 296.85$, $\hat{\sigma}^2 = 87.8$.

(b) L'intervalle cherché vaut

$$\left[\hat{\beta} \pm \hat{\sigma} \cdot qt_8(97.5\%) / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = [17.86, 22.66].$$

Chapitre 10

Régression multiple

1. (a) $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, où $i = 1, \ldots, 25$ et $\varepsilon_1, \ldots, \varepsilon_{25}$ i.i.d. $\mathcal{N}(0, \sigma^2)$.

   (b) L'estimateur de $\hat{\boldsymbol{\theta}} = (\alpha, \beta_1, \beta_2)^T$ est $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$, où $\mathbf{X} = (\mathbf{1},\ \boldsymbol{x}_1,\ \boldsymbol{x}_2)$, $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ respectivement les vecteurs des quantités de pluie et des températures moyennes relevées.

   (c) $\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, comme σ^2 n'est pas connu, il faut l'estimer par $\sum_{i=1}^{n} r_i^2/(n-p-1) = 243.3/22 = 11.0591$. L'estimation de la variance de $\hat{\boldsymbol{\theta}}$ vaut alors $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = (38.0101, 0.0155, 0.0232)^T$.

   (d) $R^2 = (\mathrm{SC_t} - \mathrm{SC_e})/\mathrm{SC_t} = 184.2/(184.2 + 243.3) = 0.43$.

   (e) $F_{\mathrm{obs}} = \mathrm{CM_r}\,/\,\mathrm{CM_e} = (184.2/2)/(243.3/22) = 8.33$, comme $qF_{2,22}(95\%) = 3.44$ on rejette $H_0 : \beta_1 = \beta_2 = 0$.

   (f) L'intervalle de confiance est $\left[\,\hat{\beta}_2 \pm \hat{\sigma}\sqrt{\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{i+1,i+1}}\, qt_{n-p-1}(97.5\%)\,\right] = [-1.41, 0.30]$, avec $qt_{22}(97.5\%) = 2.074$.

   (g) La statistique de test pour tester H_0 : "ω est le vrai modèle" vaut

$$F_{\mathrm{obs}} = \frac{(\mathrm{SC_e}(\omega) - \mathrm{SC_e}(\Omega))/(p-q)}{\mathrm{SC_e}(\Omega)/(n-p-1)} = \frac{(336.5 - 243)/1}{243/22} = 8.41.$$

   Comme $qF_{1,22}(95\%) = 4.30 < 8.41$, on ne peut pas rejeter H_0.

2. La statistique de test pour comparer deux modèles $\omega \subset \Omega$ est

$$F_{\mathrm{obs}} = \frac{(\mathrm{SC_e}(\omega) - \mathrm{SC_e}(\Omega))/(p-q)}{\mathrm{SC_e}(\Omega)/(n-p-1)},$$

   où q et p sont le nombre de paramètres -1 dans ω et Ω respectivement. Le choix du modèle se fait successivement de la façon suivante :
   - $\omega = \mathrm{cst}$,
     si $\Omega = \{\mathrm{cst}, x_1\}$: $F = 109.65$,
     si $\Omega = \{\mathrm{cst}, x_2\}$: $F = 65.62$,
     si $\Omega = \{\mathrm{cst}, x_3\}$: $F = 3.8$,
     comme $qF_{1,20}(95\%) = 4.351$, on introduit x_1 ;
   - $\omega = \{\mathrm{cst}, x_1\}$,
     si $\Omega = \{\mathrm{cst}, x_1, x_2\}$: $F = 7.35$,
     si $\Omega = \{\mathrm{cst}, x_1, x_3\}$: $F = 0.56$,
     comme $qF_{1,18}(95\%) = 4.414$, on introduit x_2 ;
   - $\omega = \{\mathrm{cst}, x_1, x_2\}$,
     si $\Omega = \{\mathrm{cst}, x_1, x_2, x_3\}$: $F = 0.89$
     comme $qF_{1,16}(95\%) = 4.494$, on n'introduit pas x_3.

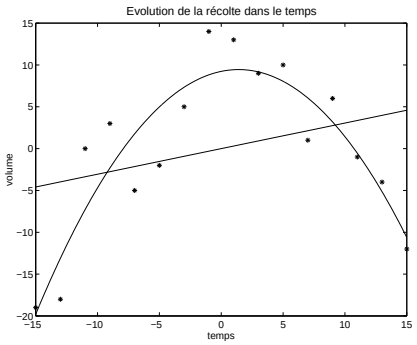
Le modèle sélectionné est donc $\{\text{cst}, x_1, x_2\}$.

3. (a) Pour le graphique *cf.* (d).

   (b) Visiblement le modèle $y_i = \alpha + \beta_1 x_i + \varepsilon_i$ n'est pas satisfaisant. On ne peut pas ajuster une droite à ces données.

   (c) L'estimateur des moindres carrés de $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2)^T$ est donné par $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$, où $\mathbf{X} = (\mathbf{1},\ \boldsymbol{x},\ \boldsymbol{x}^2)$, $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{x} = (x_1, \ldots, \boldsymbol{x}_{16})^T$ et $\boldsymbol{x}^2 = (x_1^2, \ldots, \boldsymbol{x}_{16}^2)^T$. Numériquement nous trouvons les estimations $\hat{\boldsymbol{\theta}} = (9.2336, 0.3059, -0.1086)^T$.

   (d) Le graphique ci-dessous montre le scatterplot des données ainsi que la droite de la régression simple et la parabole ajustée sous (c). Il est évident que le modèle polynômial en x fournit un meilleur ajustement que le modèle linéaire en x.



4. (a) Le premier tableau donne pour Ω les estimations des paramètres ainsi que leurs écart-types estimés et les valeurs de la statistique de test t pour tester l'hypothèse nulle que le paramètre concerné vaut zéro contre l'alternative qu'il est non nul. Dans la dernière colonne nous trouvons les p-valeurs de ces tests, les valeurs numériques indiquent que tous les paramètres sont significativement non nuls à l'exception de β_3.
   Le deuxième tableau contient l'ANOVA pour le modèle Ω. Nous obtenons les sommes des carrés de la régression et des erreurs ainsi que leurs degrés de liberté et les carrés moyen de l'erreur. La valeur de la statistique de test F et la p-valeur correspondent au test F pour tester l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ simultanément, ce qui doit être rejeté d'après la valeur donnée. A la dernière ligne nous trouvons la corrélation multiple R^2, le pourcentage de la variation totale expliquée par le modèle, qui est dans ce cas-ci suffisamment grand.
   Ensuite, nous trouvons les mêmes informations pour le modèle ω, et nous arrivons aux conclusions suivantes : tous les paramètres du modèle sont significativement non nuls, le test F rejette l'hypothèse que $\beta_1 = \beta_2 = 0$ simultanément et le R^2 est suffisamment grand pour garantir un bon ajustement.

   (b) $\hat{\sigma} = \sqrt{\text{CM}_\text{e}} = \sqrt{10.52} = 3.24$.

   (c) L'intervalle cherché est $\left[\hat{\beta}_2 \pm \left(\widehat{\text{Var}}(\hat{\beta}_2)\right)^{1/2} qt_{17}(97.5\%)\right] = [1.22, 1.38]$, avec $qt_{17}(97.5\%) = 2.11$ et $\left(\widehat{\text{Var}}(\hat{\beta}_2)\right)^{1/2} = 0.04$.

(d) La valeur observée de la statistique de test F (*cf.* exercice 3) vaut $F_{\text{obs}} = (188.8 - 178.8)/(178.8/17) = 0.95$. Comme $qF_{1,17}(95\%) = 4.45 > 0.95$ on ne peut pas rejeter l'hypothèse nulle que ω est le vrai modèle.

5. Nous considérons le modèle $\boldsymbol{y} = \boldsymbol{\theta}\mathbf{X} + \boldsymbol{\varepsilon}$ pour des erreurs $\boldsymbol{\varepsilon}$ i.i.d. selon $\mathcal{N}(0, \sigma^2)$, où $\mathbf{X} = (\mathbf{1},\ \boldsymbol{x},\ \boldsymbol{x}^2)$, $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{x} = (x_1, \ldots, x_{12})^T$ et $\boldsymbol{x}^2 = (x_1^2, \ldots, x_{12}^2)^T$. L'estimateur des moindres carrés vaut $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$. Nous trouvons

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 12 & \sum \boldsymbol{x}_{1i} & \sum \boldsymbol{x}_{2i} \\ \sum \boldsymbol{x}_{1i} & \sum \boldsymbol{x}_{1i}^2 & \sum \boldsymbol{x}_{1i}\boldsymbol{x}_{2i} \\ \sum \boldsymbol{x}_{2i} & \sum \boldsymbol{x}_{1i}\boldsymbol{x}_{2i} & \sum \boldsymbol{x}_{2i}^2 \end{pmatrix}^{-1} = \begin{pmatrix} 7.506 & -0.328 & 0.030 \\ -0.328 & 0.026 & -0.003 \\ -0.030 & -0.003 & 0.002 \end{pmatrix}$$

et $\mathbf{X}^T\boldsymbol{y} = (1033.18, 680, 44169)^T$. Finalement $\hat{\boldsymbol{\theta}} = (323.89, -0.205, -5.096)^T$.

6. (a) $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,\ i = 1, \ldots, 18$, où les erreurs ε_i sont i.i.d. selon une loi $\mathcal{N}(0, \sigma^2)$.

(b) L'estimateur de $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)^T$ est $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ où $\mathbf{X} = (\mathbf{1},\ \boldsymbol{x}_1,\ \boldsymbol{x}_2)$, $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ les vecteurs des puissances et des poids respectivement.

(c) $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

(d) $R^2 = \text{SC}_\text{t} - \text{SC}_\text{e} \,/\, \text{SC}_\text{t} = 0.68$, la valeur maximale théorique de R^2 est 1.

(e) $F_{\text{obs}} = \text{CM}_\text{r} \,/\, \text{CM}_\text{e} = 16.43 > 3.682 = qF_{2,15}(95\%)$ et donc on rejette l'hypothèse H_0.

(f) Les statistiques de test t pour les pentes β_1 et β_2 valent :

$$t_{\text{obs}}^{(1)} = \frac{\hat{\beta}_1}{\sqrt{\widehat{\text{Var}}(\beta_1)}} = 2.40\,, \quad t_{\text{obs}}^{(2)} = \frac{\hat{\beta}_2}{\sqrt{\widehat{\text{Var}}(\beta_2)}} = 1.45\,.$$

Comme $qt_{15}(97.5\%) = 2.131$ l'hypothèse nulle $H_0 : \beta_1 = 0$ doit être rejetée, tandis que $H_0 : \beta_2 = 0$ ne peut pas être rejetée.

(g) $y_i = \alpha + \beta_1 x_{1i} + \varepsilon_i$, où les erreurs ε sont i.i.d. selon une loi $\mathcal{N}(0, \sigma^2)$.

(h) La valeur observée de la statistique du test F (*cf.* exercice 3) vaut $F_{\text{obs}} = (266 - 230)/(230/15) = 2.34 < 4.54 = qF_{1,15}(95\%)$ et donc on ne peut pas rejeter l'hypothèse nulle que le modèle donné sous (h) est suffisant.

Chapitre 11

Plans d'expériences

1. (a) Tableau ANOVA :

source	SC	d.l.	CM	Test F
A	3063.188	3	1021.063	9.277
erreur	1320.750	12	110.062	
total (corr.)	4383.938	15		

On a $qF_{3,12}(95\%) = 3.490 < 9.277$, donc on rejette H_0.

(b) Tableau ANOVA :

source	SC	d.l.	CM	Test F
A	3063.188	3	1021.063	11.337
B	510.188	3	170.063	1.888
erreur	810.562	9	90.062	
total (corr.)	4383.938	15		

(1) On a $qF_{3,9}(95\%) = 3.863 < 11.377$, donc on rejette H_0.

(2) On a $qF_{3,9}(95\%) = 3.863 > 1.888$, donc on ne peut pas rejeter H_0.

(c) On constate que le fait d'inclure un nouveau facteur dans le modèle n'influence que l'erreur. Avec le deuxième tableau, on conclut que le facteur B n'est pas significatif à 95%, donc on conserve le premier modèle.

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande permet d'imposer les contraintes du type $\sum_i \alpha_i = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo1 <- c(44,47,0,36,22,43,9,14,36,41,10,1,34,53,17,34)
> A <- factor(rep(c("A1","A2","A3","A4"),4))
> summary(aov(exo1~A))
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
A          3  3063.188 1021.063 9.277115 0.001887965
Residuals 12  1320.750  110.062
> B <- factor(rep(c("B1","B2","B3","B4"),rep(4,4)))
> summary(aov(exo1~A+B))
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
A          3  3063.188 1021.063 11.33727 0.0020648
B          3   510.188  170.063  1.88827 0.2020355
Residuals  9   810.562   90.062
```

2. (a) $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, 2, 3 \text{ et } j = 1, 2)$.

   (b) $\hat{\mu} = y_{..} = 3$,
   $\hat{\alpha}_1 = y_{1.} - y_{..} = 0$, $\hat{\alpha}_2 = y_{2.} - y_{..} = -1$, $\hat{\alpha}_3 = -\hat{\alpha}_1 - \hat{\alpha}_2 = 1$,
   $\hat{\beta}_1 = y_{.1} - y_{..} = 0$ et $\hat{\beta}_2 = -\hat{\beta}_1 = 0$
   et on obtient la décomposition suivante :

$$\begin{bmatrix} -1 & 1 \\ -2 & 0 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ -2 & 2 \end{bmatrix}$$

   Le tableau ANOVA est :

source	SC	d.l.	CM	Test F
Machine	0	1	0	0
Ouvrier	4	2	2	1/3
erreur	12	2	6	
total (corr.)	18	11		

   On peut donc calculer $\mathrm{SC_e} = (-1)^2 + 1^2 + (-1)^2 + 1^2 + (-2)^2 + 2^2 = 12$, d'où $\hat{\sigma}^2 = \mathrm{CM_e} = 6$.

   (c) On veut tester $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$.
   On a $F_{\text{obs}} = \dfrac{\mathrm{CM_{Ouvrier}}}{\mathrm{CM_e}} = \dfrac{1}{3} < 19.00 = qF_{2,2}(95\%)$, donc on ne peut pas rejeter H_0.

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo2 <- c(2,1,6,4,3,2)
> Machine <- factor(rep(c("A","B"),c(3,3)))
> Ouvrier <- factor(rep(1:3,2))
> summary(aov(exo2~Machine+Ouvrier))
          Df Sum of Sq Mean Sq   F Value Pr(F)
Machine    1         0       0 0.0000000  1.00
Ouvrier    2         4       2 0.3333333  0.75
Residuals  2        12       6
> aov(exo2~Machine+Ouvrier)$coef
 (Intercept)     Machine     Ouvrier1 Ouvrier2
           3 4.07922e-16 4.440892e-16       -1
```

3. (a) $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \ldots, 4 \text{ et } j = 1, \ldots, 5)$.

   (b) Les paramètres ajustés sont les suivants :

$$\begin{aligned}
\hat{\mu} &= y_{..} = 87.65\,, & \hat{\beta}_1 &= y_{.1} - y_{..} = 0.35\,,\\
\hat{\alpha}_1 &= y_{1.} - y_{..} = -2.45\,, & \hat{\beta}_2 &= y_{.2} - y_{..} = -4.90\,,\\
\hat{\alpha}_2 &= y_{2.} - y_{..} = 1.55\,, & \hat{\beta}_3 &= y_{.3} - y_{..} = 4.85\,,\\
\hat{\alpha}_3 &= y_{3.} - y_{..} = -1.25\,, & \hat{\beta}_4 &= y_{.4} - y_{..} = 1.60\,,\\
\hat{\alpha}_4 &= -\hat{\alpha}_1 - \hat{\alpha}_2 - \hat{\alpha}_3 = 2.15\,, & \hat{\beta}_5 &= -\hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3 - \hat{\beta}_4 = -1.90\,.
\end{aligned}$$

   (c) Tableau ANOVA :

source	SC	d.l.	CM	Test F
Méthode	72.95	3	24.317	0.326
Ecole	215.30	4	53.825	0.721
erreur	896.30	12	74.692	
total (corr.)	1184.55	19		

   (d) $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$.
   On a $F_{\text{obs}} = \dfrac{\text{CM}_{\text{Méthode}}}{\text{CM}_{\text{e}}} = \dfrac{24.317}{74.692} = 0.326 < 3.49 = qF_{3,12}(95\%)$, donc on ne peut pas rejeter H_0.

   $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.
   On a $F_{\text{obs}} = \dfrac{\text{CM}_{\text{Ecole}}}{\text{CM}_{\text{e}}} = \dfrac{53.825}{74.692} = 0.721 < 3.26 = qF_{4,12}(95\%)$, donc on ne peut pas rejeter H_0.

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo3 <- c(79,84,95,94,92,96,65,78,86,94,98,92,85,
            91,86,95,84,81,88,90)
> Ecole <- factor(rep(1:5,rep(4,5)))
> Methode <- factor(rep(1:4,5))
> summary(aov(exo3~Ecole+Methode))
          Df Sum of Sq  Mean Sq   F Value     Pr(F)
Ecole      4    215.30 53.82500 0.7206293 0.5941645
Methode    3     72.95 24.31667 0.3255606 0.8069295
Residuals 12    896.30 74.69167
> aov(exo3~Ecole+Methode)$coef
 (Intercept) Ecole1 Ecole2 Ecole3 Ecole4 Methode1 Methode2 Methode3
       87.65   0.35   -4.9   4.85    1.6    -2.45     1.55    -1.25
```

4. On définit un facteur Personne à huit niveaux α_i $(i = 1, \ldots, 8)$, un facteur Agent à deux niveaux β_1 et β_2 et on pose le modèle $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $(i = 1, \ldots, 8$ et $j = 1, 2)$.

   On teste ensuite l'hypothèse $H_0 : \beta_1 = \beta_2 = 0$ pour savoir si les deux agents diffèrent significativement l'un de l'autre.

   Le tableau ANOVA est donné par :

source	SC	d.l.	CM	Test F
Personne	49.454	7	7.065	11.735
Agent	0.681	1	0.681	1.131
erreur	4.214	7	0.602	
total (corr.)	54.349	16		

   On a $F_{\text{obs}} = \dfrac{\text{CM}_{\text{Agent}}}{\text{CM}_{\text{e}}} = \dfrac{0.681}{0.602} = 1.131 < 5.591 = qF_{1,7}(95\%)$, donc on ne peut pas rejeter H_0.

   Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo4 <- c(8.4,12.8,9.6,9.8,8.4,8.6,8.9,7.9,9.4,15.2,9.1,8.8,
            8.2,9.9,9.0,8.1)
> Pers <- factor(rep(1:8,2))
> Agent <- rep(c("A","B"),c(8,8))
> summary(aov(exo4~Pers+Agent))
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Pers       7  49.45438 7.064911 11.73469 0.0021598
Agent      1   0.68062 0.680625  1.13051 0.3229653
Residuals  7   4.21438 0.602054
```

5. On pose le modèle $y_{ijkl} = \mu + f_{1_i} + f_{2_j} + f_{3_k} + f_{4_l} + \varepsilon_{ijkl}$ $(i = 1, 2,\ j = 1, 2,\ k = 1, 2$ et $l = 1, 2)$.

   (a) $\hat{\mu} = y_{....} = 9.5075$,
   $f_{1_1} = y_{1...} - y_{....} = 0.3675$, $f_{1_2} = -0.3675$,
   $f_{2_1} = y_{2...} - y_{....} = 0.0725$, $f_{2_2} = -0.0725$,
   $f_{3_1} = y_{3...} - y_{....} = -0.5575$, $f_{3_2} = 0.5575$,
   $f_{4_1} = y_{4...} - y_{....} = 4.5575$, $f_{4_2} = -4.5575$.

(b) Le tableau ANOVA est donné par :

source	SC	d.l.	CM	Test F
f_1	1.080	1	1.080	6.086
f_2	0.042	1	0.042	0.237
f_3	2.486	1	2.486	14.007
f_4	166.l66	1	166.166	936.061
erreur	0.532	3	0.177	
total (corr.)	170.306	7		

$H_0 : f_{1_1} = f_{1_2} = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{f_1}}{\text{CM}_\text{e}} = \dfrac{1.080}{0.177} = 6.086 < 10.13 = qF_{1,3}(95\%)$, donc on ne peut pas rejeter H_0.

$H_0 : f_{2_1} = f_{2_2} = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{f_2}}{\text{CM}_\text{e}} = \dfrac{0.042}{0.177} = 0.237 < 10.13 = qF_{1,3}(95\%)$, donc on ne peut pas rejeter H_0.

$H_0 : f_{3_1} = f_{3_2} = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{f_3}}{\text{CM}_\text{e}} = \dfrac{2.486}{0.177} = 14.007 > 10.13 = qF_{1,3}(95\%)$, donc on rejette H_0.

$H_0 : f_{4_1} = f_{4_2} = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{f_4}}{\text{CM}_\text{e}} = \dfrac{166.166}{0.177} = 936.061 > 10.13 = qF_{1,3}(95\%)$, donc on rejette H_0.

(c) On a la décomposition des données en valeurs ajustées ($\hat{y}$) et résidus (r), $y = \hat{y} + r$:

essai	y	$\hat{y}$	r
1	14.26	13.9475	0.3125
2	6.14	5.9475	0.1925
3	14.62	14.9175	−0.2975
4	4.48	4.6875	−0.2075
5	14.12	14.3275	−0.2075
6	3.80	4.0975	−0.2975
7	13.26	13.0675	0.1925
8	5.38	5.0675	0.3125

(d) $\hat{\sigma}^2 = \text{CM}_\text{e} = 0.177$.

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $f_{m_1} + f_{m_2} = 0 \quad (m = 1, \ldots, 4)$) :

```
> exo5 <- c(14.26,6.14,14.62,4.48,14.12,3.8,13.26,5.38)
> f1 <- factor(rep(c("j","n"),c(4,4)))
```

```
> f2 <- factor(rep(c("A","A","B","B"),2))
> f3 <- factor(c(1.2,1.3,1.3,1.2,1.3,1.2,1.2,1.3))
> f4 <- factor(rep(c(100,120),4))
> summary(aov(exo5~f1+f2+f3+f4))
            Df Sum of Sq  Mean Sq  F Value     Pr(F)
f1           1    1.0804   1.0804   6.0865 0.0902982
f2           1    0.0420   0.0420   0.2369 0.6598110
f3           1    2.4864   2.4864  14.0069 0.0332856
f4           1  166.1664 166.1664 936.0611 0.0000767
Residuals    3    0.5325   0.1775
> aov(exo5~f1+f2+f3+f4)$coef
 (Intercept)      f1      f2      f3     f4
      9.5075  0.3675  0.0725 -0.5575 4.5575
> sum(summary(aov(exo5~f1+f2+f3+f4))$Sum)
[1] 170.3079
> aov(exo5~f1+f2+f3+f4)$fit
       1      2       3      4       5      6       7      8
 13.9475 5.9475 14.9175 4.6875 14.3275 4.0975 13.0675 5.0675
> aov(exo5~f1+f2+f3+f4)$res
      1      2       3       4       5       6      7      8
 0.3125 0.1925 -0.2975 -0.2075 -0.2075 -0.2975 0.1925 0.3125
```

6. (a) On définit un facteur Jour à trois niveaux α_i $(i = 1, 2, 3)$, un facteur Agents à quatre niveaux β_j $(j = 1, \ldots, 4)$ et on pose le modèle $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $(i = 1, 2, 3$ et $j = 1, \ldots, 4)$.
Le facteur Jour, s'il est significatif, implique que l'expérience n'est pas réalisée chaque jour dans les mêmes conditions et que par conséquent l'efficacité peut varier d'un jour à l'autre. Le facteur Agents quant à lui signifie qu'il y a une différence d'efficacité entre les quatre agents chimiques.

   (b) $\hat{\mu} = y_{..} = 80$,
$\hat{\alpha}_1 = y_{1.} - y_{..} = 1.5$, $\hat{\alpha}_2 = y_{2.} - y_{..} = -1.25$, $\hat{\alpha}_3 = -\hat{\alpha}_1 - \hat{\alpha}_2 = -0.25$,
$\hat{\beta}_1 = y_{.1} - y_{..} = 2$, $\hat{\beta}_2 = y_{.2} - y_{..} = -1.\bar{6}$ et $\hat{\beta}_3 = -\hat{\beta}_1 - \hat{\beta}_2 = -0.\bar{3}$.

   La décomposition des valeurs observées selon les effets estimés et les résidus est la suivante :

$$\begin{array}{|cccc|} \hline 84 & 80 & 83 & 79 \\ 79 & 77 & 80 & 79 \\ 83 & 78 & 80 & 78 \\ \hline \end{array} = \begin{array}{|cccc|} \hline 83.50 & 79.83\bar{3} & 82.50 & 80.16\bar{6} \\ 80.75 & 77.08\bar{3} & 79.75 & 77.41\bar{6} \\ 81.75 & 78.08\bar{3} & 80.75 & 78.41\bar{6} \\ \hline \end{array}$$

$$+ \begin{array}{|cccc|} \hline 0.50 & 0.16\bar{6} & 0.50 & -1.16\bar{6} \\ -1.75 & -0.08\bar{3} & 0.25 & 1.58\bar{3} \\ 1.25 & -0.08\bar{3} & -0.75 & -0.41\bar{6} \\ \hline \end{array}$$

(c) Le tableau d'ANOVA est :

source	SC	d.l.	CM	Test F
Jour	15.500	2	7.750	4.729
Agents	28.667	3	9.556	5.831
erreur	9.833	6	1.639	
total (corr.)	54.000	11		

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo6 <- c(84,79,83,80,77,78,83,80,80,79,79,78)
> Jour <- factor(rep(1:3,4))
> Agents <- factor(rep(1:4,rep(3,4)))
> summary(aov(exo6~Agents+Jour))
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
Agents     3  28.66667 9.555556 5.830508 0.03275623
Jour       2  15.50000 7.750000 4.728814 0.05848241
Residuals  6   9.83333 1.638889
> aov(exo6~Agents+Jour)$coef
 (Intercept) Agents1   Agents2 Agents3 Jour1 Jour2
          80       2 -1.666667       1   1.5 -1.25
> aov(exo6~Agents+Jour)$fit
    1     2     3        4        5        6    7     8     9
 83.5 80.75 81.75 79.83333 77.08333 78.08333 82.5 79.75 80.75
      10       11       12
 80.16667 77.41667 78.41667
> aov(exo6~Agents+Jour)$res
   1     2    3         4           5           6   7    8     9
 0.5 -1.75 1.25 0.1666667 -0.08333333 -0.08333333 0.5 0.25 -0.75
       10       11         12
 -1.166667 1.583333 -0.4166667
```

7. (a) On définit un facteur Engrais à quatre niveaux α_i ($i = 1, \ldots, 4$), un facteur Plant à quatre niveaux β_j ($j = 1, \ldots, 4$) et on pose le modèle $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $(i = 1, \ldots, 4$ et $j = 1, \ldots, 4)$.

(b)

					$y_{i.}$	α_i
	44.0000	22.0000	36.0000	34.0000	34.0000	6.4375
	47.0000	43.0000	41.0000	53.0000	46.0000	18.4375
	0.0000	9.0000	10.0000	17.0000	9.0000	−18.5625
	36.0000	14.0000	1.0000	34.0000	21.2500	−6.3125
$y_{.j}$	31.7500	22.0000	22.0000	34.5000	$y_{..} = 27.5625 = \hat{\mu}$	
β_j	4.1875	−5.5625	−5.5625	6.9375		

(c) Le tableau d'ANOVA est :

source	SC	d.l.	CM	Test F
Engrais	3063.188	3	1021.063	11.337
Plants	510.188	3	170.063	1.888
erreur	810.562	9	90.062	
total (corr.)	4383.938	15		

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{\text{Engrais}}}{\text{CM}_{\text{e}}} = \dfrac{1021.063}{90.062} = 11.337 > 3.863 = qF_{3,9}(95\%)$, donc on rejette H_0.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.
On a $F_{\text{obs}} = \dfrac{\text{CM}_{\text{Plants}}}{\text{CM}_{\text{e}}} = \dfrac{170.063}{90.062} = 1.888 < 3.863 = qF_{3,9}(95\%)$, donc on ne peut pas rejeter H_0.

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo7 <- c(44,47,0,36,22,43,9,14,36,41,10,1,34,53,17,34)
> Engrais <- factor(rep(c("A","B","C","D"),4))
> Plant <- factor(rep(c("I","II","III","IV"),rep(4,4)))
> summary(aov(exo7~Engrais+Plant))
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Engrais    3  3063.188 1021.063 11.33727 0.0020648
Plant      3   510.188  170.063  1.88827 0.2020355
Residuals  9   810.562   90.062
> aov(exo7~Engrais+Plant)$coef
 (Intercept) Engrais1 Engrais2 Engrais3 Plant1  Plant2  Plant3
     27.5625   6.4375  18.4375 -18.5625 4.1875 -5.5625 -5.5625
```

8. (a) On définit un facteur Groupe à trois niveaux α_i ($i = 1, 2, 3$), avec deux observations pour le niveau 1 ($n_1 = 2$) , quatre observations pour le niveau 2 ($n_2 = 4$) et trois observations pour le niveau 3 ($n_3 = 3$) et on pose le modèle $y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (i = 1, 2, 3 \text{ et } j = 1, \ldots, n_i)$.
On définit ainsi un niveau pour chaque méthode et l'analyse de variance permet ensuite de tester $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$, c'est-à-dire s'il y a ou non une différence significative entre le placebo, l'aspirine et le nouveau médicament.

(b) $\hat{\mu} = y_{..} = 2.158\bar{3}$,
$\hat{\alpha}_1 = y_{1.} - y_{..} = 1.041\bar{6}$, $\hat{\alpha}_2 = y_{2.} - y_{..} = 0.61\bar{6}$ et $\hat{\alpha}_3 = -\hat{\alpha}_1 - \hat{\alpha}_2 = -1.658\bar{3}$.

On a la décomposition des valeurs observées en somme de valeurs ajustées ($\hat{y}$) et résidus (r), $y = \hat{y} + r$:

y	$\hat{y}$	r
0.0	0.500	−0.500
1.0	0.500	0.500
2.3	2.775	−0.475
3.5	2.775	0.725
2.8	2.775	0.025
2.5	2.775	−0.275
3.1	3.200	−0.100
2.7	3.200	−0.500
3.8	3.200	0.600

(c) Le tableau d'ANOVA est :

source	SC	d.l.	CM	Test F
Groupes	9.701	2	4.851	14.944
erreur	1.948	6	0.325	
total (corr.)	11.649	9		

Les commandes suivantes permettent d'obtenir les résultats ci-dessus avec `Splus` (la première commande permet d'imposer la contrainte $\sum_i \alpha_i = 0$) :

```
> options(contrasts=c("contr.sum","contr.sum"))
> exo8 <- c(0,1,2.3,3.5,2.8,2.5,3.1,2.7,3.8)
> Groupes <- factor(c(rep("P",2),rep("N",4),rep("A",3)))
> summary(aov(exo8~Groupes))
            Df Sum of Sq  Mean Sq  F Value       Pr(F)
Groupes      2  9.701389 4.850694 14.94437 0.004672818
Residuals    6  1.947500 0.324583
> aov(exo8~Groupes)$coef
 (Intercept) Groupes1  Groupes2
    2.158333 1.041667 0.6166667
> aov(exo8~Groupes)$fit
   1   2     3     4     5     6   7   8   9
 0.5 0.5 2.775 2.775 2.775 2.775 3.2 3.2 3.2
> aov(exo8~Groupes)$res
    1   2      3     4     5      6    7    8   9
 -0.5 0.5 -0.475 0.725 0.025 -0.275 -0.1 -0.5 0.6
```

Chapitre 12

Tests khi-deux

1. Le tableau théorique est

	$R \geq 60000$	$R \geq 30000$ $R < 60000$	$R \geq 12000$ $R < 30000$	$R < 12000$
Républicain	36.7	25.7	25.7	22
Démocrate	53.3	37.3	32	32
Autre	10	7	7	2

Pour la statistique de Pearson, on obtient $\text{Pearson}_{\text{obs}} = 45.12$; le quantile correspondant est $q\chi^2_6(95\%) = 12.59 < 45.12$. Donc on rejette l'hypothèse de l'indépendance entre le salaire et l'affiliation politique.

2. Les répartitions théoriques sont

[0,70)	[70,85)	[85,100)	[100,115)	[115,130)	$[130, \infty)$
22.75	135.91	341.34	341.34	135.91	22.75

Pour la statistique de Pearson, on obtient $\text{Pearson}_{\text{obs}} = 13.21$; le quantile correspondant est $q\chi^2_5(95\%) = 11.07 < 13.21$. Donc on rejette l'hypothèse $Z \sim \mathcal{N}(100, 225)$.

3. Tableau de contingence

	f	nf	total
m	400	200	600
nm	300	100	400
total	700	300	1000

Tableau théorique

	f	nf
m	420	180
nm	280	120

Pour la statistique de Pearson, on obtient $\text{Pearson}_{\text{obs}} = 7.94$; le quantile correspondant est $q\chi^2_1(95\%) = 3.84 < 7.94$. Donc on rejette l'hypothèse de l'indépendance entre la consommation du chocolat et des cigarettes.

4. Dans cet exercice la statistique de Pearson est donnée par

$$\text{Pearson} = \sum \frac{(2000 \cdot \text{Michon} - 2000 \cdot \text{inconnu})^2}{2000 \cdot \text{Michon}}$$

et on observe $\text{Pearson}_{\text{obs}} = 380.08$; le quantile correspondant est $q\chi^2_9(99\%) = 21.67$. Donc on peut rejeter l'hypothèse nulle que Ephrem Michon est l'auteur du nouvel ouvrage au niveau $\alpha = 1\%$.

5. (a) On obtient $\bar{x} = 4.5 <,,$ $\hat{\lambda} = 2/9$ et les fréquences théoriques suivantes

intervalle	0–2	2–4	4–6	6–8	8–10	10–12
fréquence	0.3588	0.2301	0.1475	0.0946	0.0606	0.0389
$76 \cdot$ fréquence	27.27	17.49	11.21	7.19	4.61	2.96

intervalle	12–14	14–16	16–18	18–26	26–28	28–∞
fréquence	0.0249	0.0150	0.0162	0.0152	0.0011	0.002
$76 \cdot$ fréquence	1.89	1.21	0.78	1.16	0.08	0.15

   (b) Les nombres d'observations obtenus par la répartition théorique dans les dernières classes sont trop petites ; pour effectuer le test khi-carré on doit regrouper les classes :

intervalle	0–2	2–4	4–6	6–8	8–12	12–∞
observation o_j	30	15	11	9	5	6
$76 \cdot$ fréquence	27.27	17.49	11.21	7.19	7.57	5.28

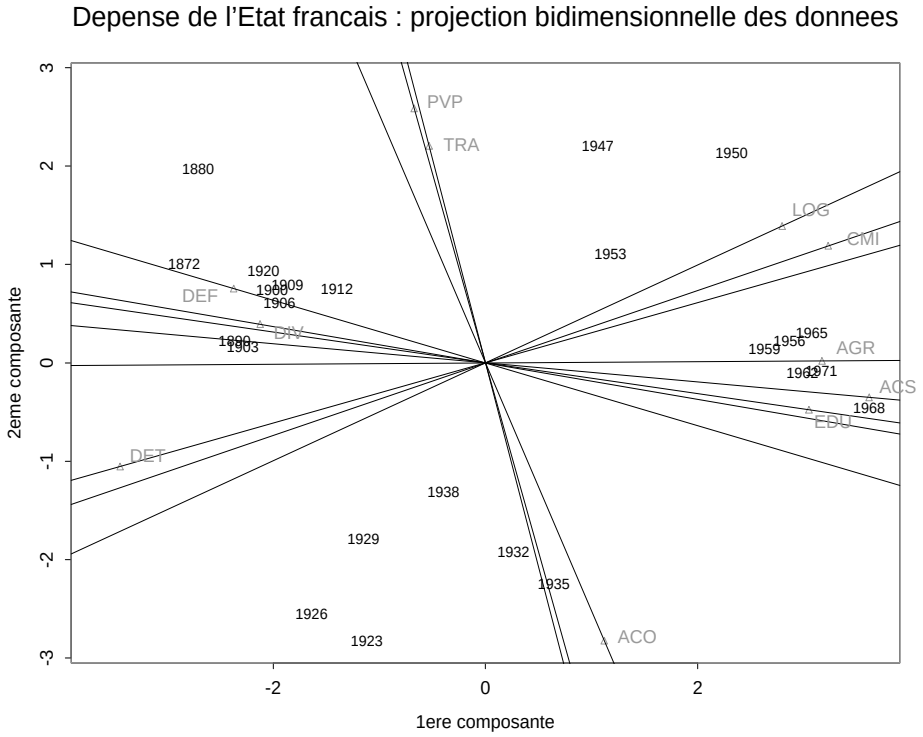
   Pour la statistique de Pearson, on obtient $\text{Pearson}_{\text{obs}} = 2.06$; les degrés de liberté sont $6 - 1 - 1 = 4$; le quantile correspondant est $q\chi^2_4(95\%) = 9.488 > 2.06$. Donc on ne peut pas rejeter l'hypothèse nulle, la qualité de l'ajustement de la loi exponentielle est suffisante.

Chapitre 13

Analyse en composantes principales

1. (a) La variable Y_4 pour le premier vecteur propre et la variable Y_1 pour le deuxième vecteur propre.
   (b) Les variables Y_3, Y_4 et Y_5 (éventuellement Y_2) pour le premier vecteur propre et la variable Y_1 pour le deuxième vecteur propre.
   (c) Lorsque l'ACP est effectuée sur la matrice de covariance, toutes les variables sont supposées avoir la même échelle. Or, dans cet exemple, il est clair que la variable Y_4, ainsi que la variable Y_1, ont des valeurs nettement supérieures aux valeurs des autres variables, Y_2, Y_3 et Y_5. Il n'est donc pas surprenant que ces deux variables soient les plus prépondérantes dans la projection bidimensionnelle trouvée dans i).
   (d) Il faudrait centrer et réduire les variables (les standardiser).

2. (a) On aimerait connaître la structure de ce jeu de données, mais comme onze variables sont observées, il est difficile de les représenter graphiquement. L'ACP permet de résumer ces variables à deux ou trois composantes importantes.
   (b) On explique $\dfrac{5.04 + 1.95}{11} = 63.55\%$ de la variation (11 étant le nombre de variables).
   (c)



On distingue essentiellement trois groupes :

- années 1872–1920, caractérisées par de fortes dépenses dans le ministère DET, DIV et DET mais peu de dépenses pour AGR, CMI, EDU, ACS et LOG ;
- années 1923–1935, caractérisées par de fortes dépenses pour ACO et faibles pour PVP et TRA ;
- années 1956–1971, caractérisées par de grandes dépenses pour AGR, ACS et EDU.

3. Les valeurs des moyennes empiriques des données manquent dans l'énoncé de l'exercice. Soient $\bar{u}$, $\bar{w}$, $\overline{uw}$, $\overline{u^2}$, respectivement $\overline{w^2}$ ces moyennes pour u_i, w_i, $u_i w_i$, u_i^2, respectivement w_i^2 $(i = 1, \ldots, n)$, alors on donne $\bar{u} = 1.98$, $\bar{u} = -3.02$, $\overline{uw} = -5.86$, $\overline{u^2} = 4.09$ et $\overline{w^2} = 9.81$.

   (a) La dimension de $\mathbf{Y}$ est $n \times 5$; les colonnes de cette matrice sont

$$\boldsymbol{y}_1 = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \ \boldsymbol{y}_2 = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \ \boldsymbol{y}_3 = \begin{pmatrix} u_1 w_1 \\ \vdots \\ u_n w_n \end{pmatrix},$$
$$\boldsymbol{y}_4 = \begin{pmatrix} u_1^2 \\ \vdots \\ u_n^2 \end{pmatrix} \text{ et } \boldsymbol{y}_5 = \begin{pmatrix} u_1^2 \\ \vdots \\ w_n^2 \end{pmatrix}.$$

   (b) Posons

$$\mathbf{E} = \begin{pmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{pmatrix}, \ \boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \text{ et } \boldsymbol{z} = \begin{pmatrix} u \\ w \end{pmatrix}.$$

   On a

$$\begin{aligned} 0 = {} & e_{11}\, u^2 + 2\, e_{12}\, uw + e_{22}\, u^2 \\ & - 2\, u\, (e_{11}\, m_1 + e_{12}\, m_2) - 2\, w\, (e_{21}\, m_1 + e_{22}\, m_2) + f \end{aligned}$$
$$\iff 0 = \boldsymbol{z}^T \mathbf{E}\, \boldsymbol{z} - 2\, \boldsymbol{z}^T \mathbf{E}\, \boldsymbol{m} + f = 0.$$

   A partir des données de l'exercice, on définit

$$\hat{\mathbf{E}} = \begin{pmatrix} -0.164 & 0.041 \\ 0.041 & -0.038 \end{pmatrix},$$

   et en résolvant le système

$$-2\, \hat{\mathbf{E}}\, \hat{\boldsymbol{m}} = \begin{pmatrix} 0.902 \\ -0.390 \end{pmatrix}, \text{ on trouve } \hat{\boldsymbol{m}} = \begin{pmatrix} 2.010 \\ -2.969 \end{pmatrix}.$$

   (c) Transformons l'équation de l'ellipse sous la forme $\dfrac{(x - c_1)^2}{a^2} + \dfrac{(y - c_2)^2}{b^2} = 1$, où (c_1, c_2) est son centre et a et b sont les longueurs de ses demi-axes. Soit $\mathbf{P}$, la matrice dont les colonnes sont les vecteurs propres (normés) de $\hat{\mathbf{E}}$ et

λ_1 , λ_2 les valeurs propres associées. Posons $\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. On a alors $\hat{\mathbf{E}} = \mathbf{P}^T \mathbf{D}\, \mathbf{P}$ ($\mathbf{P}^{-1} = \mathbf{P}^T$ car $\mathbf{P}$ orthogonale).
Ainsi, on a

$$\begin{aligned} & 0 = \boldsymbol{z}^T \hat{\mathbf{E}}\, \boldsymbol{z} - 2\, \boldsymbol{z}^T \hat{\mathbf{E}}\, \hat{\boldsymbol{m}} + f \\ \iff \quad & 0 = 2\, \boldsymbol{z}^T \mathbf{P}^T \mathbf{D}\, \mathbf{P}\, \boldsymbol{z} - 2\, \boldsymbol{z}^T \mathbf{P}^T \mathbf{P}\, \hat{\mathbf{E}}\, \hat{\boldsymbol{m}} + f \\ \iff \quad & 0 = \tilde{\boldsymbol{z}}^T \mathbf{D}\, \tilde{\boldsymbol{z}} - 2\, \tilde{\boldsymbol{z}}^T \mathbf{P}\, \hat{\mathbf{E}}\, \hat{\boldsymbol{m}} + f \end{aligned}$$

avec $\tilde{\boldsymbol{z}} = \mathbf{P}\, \boldsymbol{z} = \begin{pmatrix} \tilde{u} \\ \tilde{w} \end{pmatrix}$, ($\boldsymbol{z} = \mathbf{P}^T \tilde{\boldsymbol{z}}$). En posant $\mathbf{P}\, \hat{\mathbf{E}}\, \hat{\boldsymbol{m}} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, on obtient :

$$\begin{aligned} & 0 = \lambda_1\, \tilde{u}^2 + \lambda_2\, \tilde{w}^2 - 2\, \alpha\, \tilde{u} - 2\, \beta\, \tilde{w} + f \\ \iff \quad & \lambda_1 \left(\tilde{u}^2 - 2\, \frac{\alpha}{\lambda_1}\, \tilde{u} + \frac{\alpha^2}{\lambda_1^2} \right) + \lambda_2 \left(\tilde{w} - 2\, \frac{\beta}{\lambda_2}\, \tilde{w} + \frac{\beta^2}{\lambda_2^2} \right) = \frac{\alpha^2}{\lambda_1^2} + \frac{\beta^2}{\lambda_2^2} - f \\ \iff \quad & \frac{(\tilde{u} - \alpha/\lambda_1)^2}{1/\lambda_1} + \frac{(\tilde{w} - \beta/\lambda_2)^2}{1/\lambda^2} = \frac{\alpha^2}{\lambda_1^2} + \frac{\beta^2}{\lambda_2^2} - f\,. \end{aligned}$$

On en déduit :

$$c_1 = \frac{\alpha}{\lambda_1},\ c_2 = \frac{\beta}{\lambda_2},\ a^2 = \frac{\alpha^2/\lambda_1^2 + \beta^2/\lambda_2^2 - f}{\lambda_1} \text{ et } b^2 = \frac{\alpha^2/\lambda_1^2 + \beta^2/\lambda_2^2 - f}{\lambda_2^2}\,.$$

Valeurs numériques : $c_1 = -2.276$, $c_2 = 2.770$, $a = 1.309$ et $b = 0.502$.
L'ellipse cherchée s'obtient en appliquant la transformation $\boldsymbol{z} = \mathbf{P^T} \tilde{\boldsymbol{z}}$ aux points de l'ellipse obtenus ci-dessus.
Il reste encore à déterminer la constante f. Soient m_u, m_w, m_{uw}, m_u^2, m_w^2, l'espérance de respectivement U, W, UW, U^2 et W^2. On a donc

$$a_1(U - m_u) + a_2(W - m_w) + a_3(UW - m_{uw}) + a_4(U^2 - m_{u^2}) + a_5(W^2 - m_{w^2}) = 0$$

pour tout $a_1, \ldots, a_5 \in \mathbb{R}$. Comme on connaît les moyennes empiriques, on peut écrire :

$$\begin{aligned} 0 &= a_{10}(u - \bar{u}) + a_{01}(u - \bar{u}) + a_{11}(uw - \overline{uw}) + a_{20}\,(u^2 - \bar{u}^2) + a_{02}(w^2 - \bar{w}^2) \\ &= a_{10}\, u + a_{01}\, w + a_{11}\, uw + a_{20}\, u^2 + a_{02}\, w^2 \\ &\quad + (-a_{10}\, \bar{u} - a_{01}\, \bar{w} - a_{11}\, \overline{uw} - a_{20}\, \bar{u}^2 - a_{02}\, \bar{w}^2). \end{aligned}$$

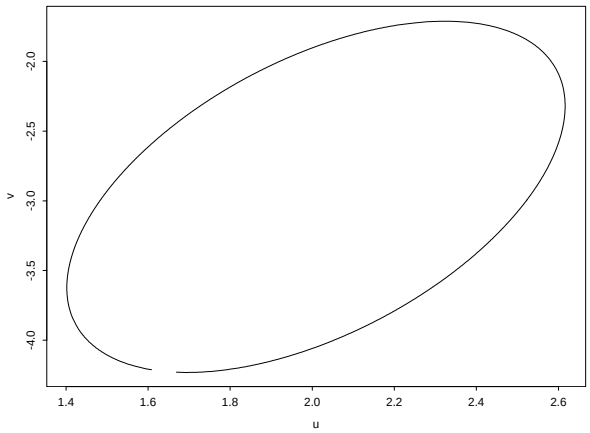
Ainsi

$$f = (a_{10}\, \bar{u} + a_{01}\, \bar{w} + a_{11}\, \overline{uw} + a_{20}\, \bar{u}^2 + a_{02}\, \bar{w}^2) = -1.441.$$

Les commandes SPLUSsuivantes permettent de calculer l'équation de l'ellipse et de la dessiner :

```
E <- matrix(c(-0.164,0.0409,0.0409,-0.038),ncol=2)
d <- eigen(E)$value
m <- solve(-2*E,c(0.902,-0.39))
```

```
P <- eigen(E)$vector
r <- P%*%E%*%m
ce <- r/d
f <- -(0.902*1.98-0.39*(-3.02)+0.0818*(-5.86)-0.164*4.09
         -0.038*9.81)
k <- (r[1]^2)/d[1] + (r[2]^2)/d[2] - f
a <- sqrt(k/d[1])
b <- sqrt(k/d[2])
x <- seq(ce[1]-a,ce[1]+a,length=1000)
z1 <- cbind(x,ce[2]-b*sqrt(1-((x-ce[1])/a)^2))
z2 <- cbind(rew(x),ce[2]+b*sqrt(1-((rew(x)-ce[1])/a)^2))
u <- rbind(z1,z2)%*%P
plot(u,type="l",xlab="u",ylab="w")
```



Chapitre 14

Modèles linéaires

1. Le diagramme branche-et-feuille nous donne :

0	133345678
1	06689
2	02227
3	059
4	5
5	3
6	
7	
8	
9	1
10	
11	2

La forme du diagramme nous suggère une transformation avec $\lambda < 1$ (*cf.* page 219/225(1997) ou page 231/237(2001)). Le diagramme branche-et-feuille avec $\lambda = \frac{1}{2}$, c'est-à-dire les racines carrées des observations originales, nous donne :

1	16789
2	2568
3	1
4	01335777
5	249
6	37
7	3
8	
9	5
10	6

Nous constatons encore une asymétrie sur la gauche. Le diagramme branche-et-feuille avec $\lambda = 0$, c'est-à-dire les logarithmes des observations originales, nous donne :

9	5
10	23458
11	1125
12	001123335689
13	0279

Cette fois-ci les données sont asymétriques vers la droite. Les deux derniers diagrammes nous suggèrent une transformation avec λ entre 0 et 1/2.

Remarque : Il existe des mesures statistiques numériques caractérisant la symétrie. Parmi elles, on peut citer le coefficient d'asymétrie défini par :

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

des données symétriques sont caractérisées par $\gamma = 0$.

2. Soient $\xi_1, \xi_2, \xi_3, \xi_4$ les vraies dénivellations et x_i, $i = 1, \ldots, 4$, les valeurs mesurées, c'est-à-dire $x_1 = 1040\,$, $x_2 = -480\,$, $x_3 = 800$ et $x_4 = -1350\,$. La contrainte est $\xi_1 + \cdots + \xi_4 = 0$. Le modèle linéaire est :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

$$\text{avec } \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} = 0.$$

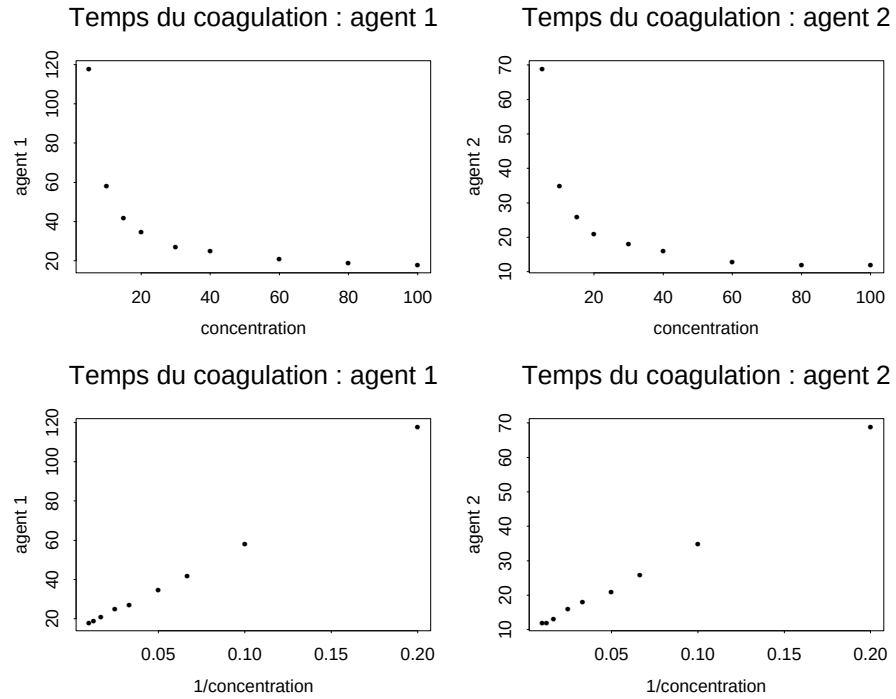
Nous trouvons alors

$$\begin{pmatrix} \hat{\xi}_1 \\ \hat{\xi}_2 \\ \hat{\xi}_3 \\ \hat{\xi}_4 \\ \lambda \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ 0 \end{pmatrix}$$

$$= \frac{1}{4} \begin{pmatrix} 3 & -1 & -1 & -1 & 1 \\ -1 & 3 & -1 & -1 & 1 \\ -1 & -1 & 3 & -1 & 1 \\ -1 & -1 & -1 & 3 & 1 \\ 1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ 0 \end{pmatrix}.$$

D'où l'on tire la solution $\hat{\xi}_1 = 1037.5\,$, $\hat{\xi}_2 = -482.5\,$, $\hat{\xi}_3 = 797.5$ et $\hat{\xi}_4 = -1352.5\,$. (Le surplus $-10\,$m est distribué uniformément sur toutes les dénivellations.)

3. Le temps de coagulation sanguin doit être expliqué par la concentration et le type de l'agent coagulant. Pour commencer nous représentons graphiquement le temps en fonction de la concentration séparément pour les deux agents. Il est évident que la relation entre les variables n'est pas linéaire et nous prenons la réciproque de la concentration comme nouvelle variable explicative, *cf.* la figure ci-dessous.



Nous commençons en ajustant le modèle le plus simple, où le temps `y` est expliqué par la variable `z`, la réciproque de la concentration et le facteur `f` le type de l'agent coagulant. Voici la sortie du logiciel statistique `Splus` :

```
> x <- c(5,10,15,20,30,40,60,80,100)
> y1 <- c(118,58,42,35,27,25,21,19,18)
> y2 <- c(69,35,26,21,18,16,13,12,12)
> z <- 1/x
> f <- factor(rep("a",length(x)), rep("b",length(x)))
> y <- c(y1,y2)
> z <- c(z,z)
> options( contrasts=c('contr.sum','contr.sum'))
> summary( lm(y~z+f))

Call: lm(formula = y ~ z + f)
Residuals:
    Min      1Q Median    3Q   Max
 -14.06 -3.003 -2.217 4.113 19.28

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   9.1510   2.4415     3.7481   0.0019
          z 408.7016  30.1003    13.5780   0.0000
          f   7.8333   1.7332     4.5195   0.0004

Residual standard error: 7.353 on 15 degrees of freedom
Multiple R-Squared: 0.9318
F-statistic: 102.4 on 2 and 15 degrees of freedom, the p-value is 0
```

L'ajustement semble être satisfaisant, mais le graphique des résidus en fonction des valeurs ajustées (*cf.* la figure ci-dessous) révèle que l'écart-type des résidus augmente linéairement, nous allons par conséquent ajuster une régression pondérée. Voici la sortie de `Splus` :

```
> summary( lm(y~z+f,weights=1/z^2))

Call: lm(formula = y ~ z + f, weights = 1/z^2)
Residuals:
    Min     1Q Median    3Q   Max
 -90.08 -44.35 -12.25 47.46 156.9

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  11.1850   0.5304    21.0880   0.0000
          z 358.9878  26.9582    13.3165   0.0000
          f   3.6306   0.3272    11.0949   0.0000

Residual standard error: 70.56 on 15 degrees of freedom
Multiple R-Squared: 0.9524
F-statistic: 150.2 on 2 and 15 degrees of freedom, the p-value is 0
```

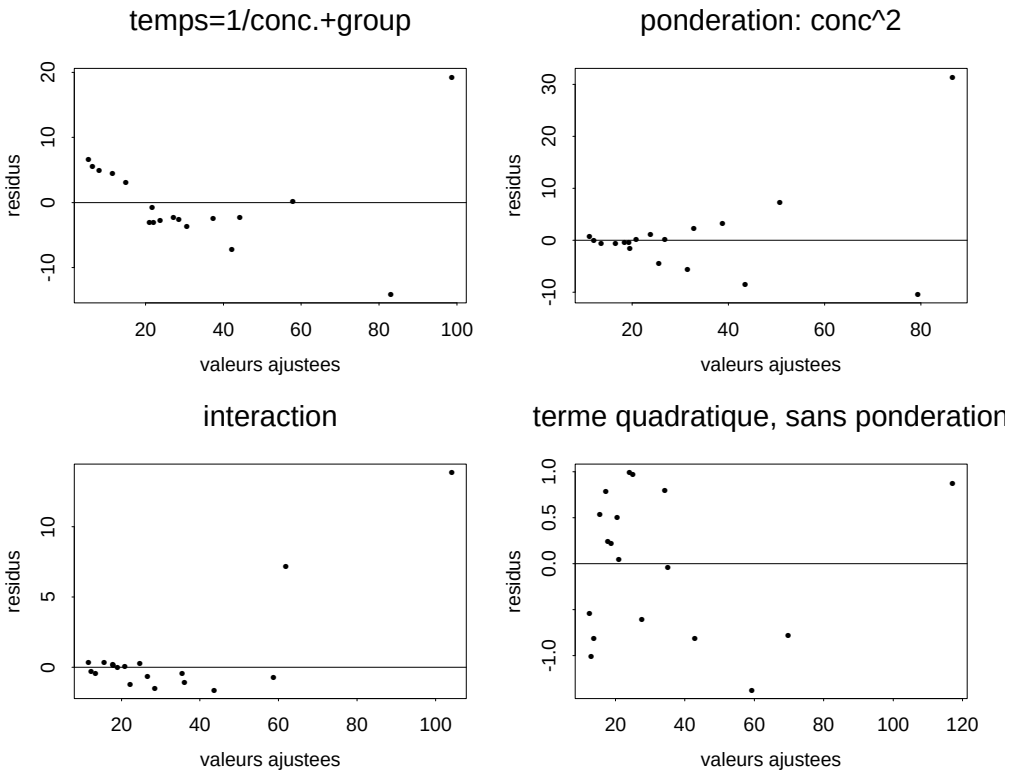
Les résidus, qui étaient grands avant l'introduction de la pondération, ont encore augmenté, mais nous constatons une amélioration dans l'ajustement des autres points (*cf.* figure). L'importance des grands résidus dans ce graphique est surestimé, car leur taille devrait aussi être pondérée. Finalement un terme d'interaction entre `z` et `f` est encore introduit, le sortie de `Splus` indique qu'il est significatif, le graphique des résidus en fonction des valeurs ajustées révèle encore une amélioration (*cf.* figure).

```
> summary( lm(y~z+f+z:f, weights=1/z^2))

Call: lm(formula = y ~ z + f + z:f, weights = 1/z^2)
Residuals:
    Min     1Q Median    3Q   Max
 -44.33 -23.34   -2.9 12.05 69.28

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  11.1850   0.2277    49.1200   0.0000
          z 358.9878  11.5736    31.0178   0.0000
          f   2.1596   0.2277     9.4841   0.0000
        z:f  95.0047  11.5736     8.2087   0.0000

Residual standard error: 30.29 on 14 degrees of freedom
Multiple R-Squared: 0.9918
F-statistic: 565.8 on 3 and 14 degrees of freedom, the p-value is 0
```

La démarche d'une régression pondérée est hasardeuse, car la structure dans les premiers résidus pourrait aussi être parabolique, ce qui nous amène à introduire plutôt un terme `z^2`. Le graphique correspondant nous montre qu'en effet les

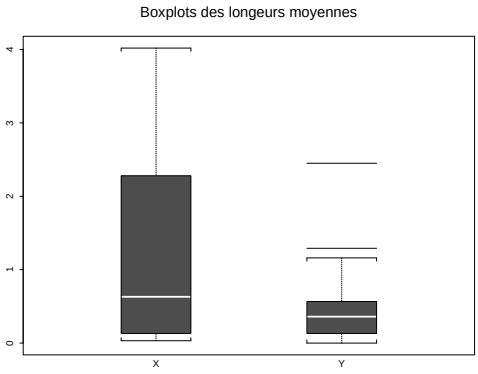
résidus n'augmentent plus.

```
> summary( lm(y~z+z^2+f+z:f))

Call: lm(formula = y ~ z + z^2 + f + z:f)
Residuals:
    Min      1Q Median     3Q    Max
 -1.372 -0.7314  0.133 0.7231 0.9897

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  12.2384   0.4414    27.7235   0.0000
          z 284.8950  13.7531    20.7149   0.0000
     I(z^2) 605.7822  64.9460     9.3275   0.0000
          f   1.4990   0.2921     5.1320   0.0002
        z:f 110.8768   3.6010    30.7908   0.0000

Residual standard error: 0.8797 on 13 degrees of freedom
Multiple R-Squared: 0.9992
F-statistic: 3836 on 4 and 13 degrees of freedom, the p-value is 0
```



4. i. Pour la loi $\chi^2_\nu = \mathcal{G}(\nu/2, 1/2)$ et pour la loi de Poisson, cette relation est vraie.
   ii. Pour la loi exponentielle, cette relation est vraie.

5. Non, le test t de Student ne peut pas s'appliquer directement, car les variances des deux échantillons ne sont pas les mêmes (*cf.* figure). De plus, les deux jeux de données sont fortement asymétriques. Les données doivent être transformées, d'après l'indication à la page 227(1997)/239(2001) du livre nous trouvons $\lambda = -1.92$.



Nous appliquons le test t de Student pour tester l'égalité des moyennes des deux échantillons x et y transformés. Les valeurs numériques sont : $n = m = 16$, $\bar{x} = 113.21$, $\bar{y} = 585.14$, $s_x^2 = 241.67$, $s_y^2 = 1792.08$, $T_{obs} = -0.3691$ et $qt_{30}(95\%) = 1.697$. Donc, les moyennes ne sont pas significativement différentes.
Remarque : la statistique de test appliquée aux données brutes est 0.6761.

6. i. On a

$$\mathbf{X}^T\mathbf{X} = \mathbf{I}_4, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

et en résolvant le système correspondant, on trouve $\hat{\alpha}_1 = \hat{\alpha}_3 = 119.925$ et $\hat{\alpha}_2 = \hat{\alpha}_4 = 60.075$.

ii. On trouve

$$\hat{\sigma}^2 = \frac{1}{4-4+3}\sum_{i=1}^{4}(a_i - \hat{\alpha}_i)^2 = 2.429,$$

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}) = \hat{\sigma}^2\mathbf{P} = 2.429 \begin{pmatrix} \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

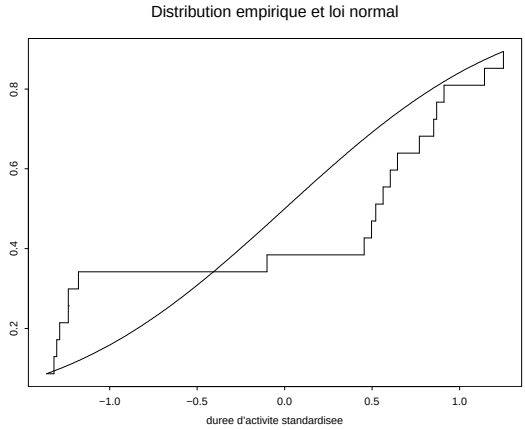
iii. Chaque ligne de la matrice $\mathbf{X}$ apparaît deux fois, mais la matrice des contraintes $\mathbf{B}$ reste la même.

iv. La variance diminue.

Chapitre 15

Inférence non paramétrique

1. La statistique D_n de Kolmogorov-Smirnov est égale à 0.2743. Avec $n = 20$ on obtient $n^{1/2}D_n = 1.227$, la valeur critique du tableau 15.5 du livre (page 260(1997) ou 272(2001)) pour $n = 20$ est égale à $1.315 > 1.227$, on ne peut par conséquent pas rejeter l'hypothèse que les données suivent une loi normale. La figure ci-dessous montre la distribution empirique des données centrées réduites et la distribution de la loi normale centrée réduite.



2. Supposons $F_{\text{paysan}}(t) = F_{\text{autre}}(t - \Delta)$, les hypothèses sont $H_0 : \Delta = 0$ et $H_1 : \Delta \neq 0$. La statistique de Wilcoxon est $W = 41.5 + 22 + 13 + 18 + 20 + 10 + 30.5 + 36 + 39 + 45 + 34 + 6 = 321$. Les échantillons sont assez grands pour justifier l'approximation suivante

$$\frac{W - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \sim \mathcal{N}(0,1).$$

   La valeur observée est égale à $0.6428 < 1.96$, donc on ne peut pas rejeter l'hypothèse qu'il n'y a pas de différence physique.

3. On regarde les différences "après-avant" car les échantillons sont appariés. Les hypothèses pour le centre de symétrie μ sont $H_0 : \mu = 0$ et $H_1 : \mu < 0$, car on s'intéresse à la diminution de la tension du sang. Le score de Wilcoxon pour un échantillon est $W^+ = 8.5 + 4 + 1 + 10 + 6 = 29.5$; la valeur critique du tableau 15.3 du livre (page 258(1997) ou 270(2001)) pour $n = 15$ est $qW_{15}^+(5\%) = 30 > 29.5$, donc on rejette l'hypothèse nulle : "l'effet du traitement est significatif".

4. (a) Les échantillons sont non-appariés donc on utilise le test de Wilcoxon pour deux échantillons. Les hypothèses sont $H_0 : \Delta = 0$ et $H_1 : \Delta < 0$ (diminution). Le score de Wilcoxon est égal à $W = 115$; la valeur critique du tableau 15.1 du livre (page 254(1997) ou 266(2001)) est $qW_{9,9}(5\%) = 66$ et alors $qW_{9,9}(95\%) = 9 \cdot 19 - 66 = 105 < 115$, donc on doit rejeter l'hypothèse nulle : "l'effet du traitement est significatif".

(b) L'autre test que l'on peut utiliser est le test t de Student pour deux échantillons ; il est fiable sous les conditions suivantes :
- les variances des deux échantillons sont égales,
- les observations sont non-corrélées,
- les deux échantillons sont non-corrélés,
- les mesures sont distribuées selon une loi normale.

Pour vérifier ces conditions on utilise le diagramme branche-et-feuille

avant		après
	3	15589
	4	9
75	5	
943	6	05
74	7	3
	8	
21	9	

Le diagramme montre une asymétrie dans les observations, donc la loi normale n'est pas un bon modèle, il faut par conséquent préférer le test de Wilcoxon.

5. On regarde les différences "nouveau-contrôle" car les échantillons sont appariés.
   (a) Les hypothèses sont $H_0 : \mu = 0$ et $H_1 : \mu > 0$ (amélioration). Le score de Wilcoxon pour un échantillon est égal à $W^+ = 8 + 10 + 7 + 9 + 2 + \frac{1}{2} \cdot 1 + 6 = 42.5$; la valeur critique du tableau 15.3 du livre (page 258(1997) ou 270(2001)) est égale à $qW^+_{10}(5\%) = 10$ d'où $qW^+_{10}(95\%) = 55 - 10 = 45$ et on ne rejette pas l'hypothèse nulle.
   (b) L'autre test que l'on peut utiliser est le test t de Student ; il est fiable si les observations sont indépendantes et distribuées selon une loi normale de même variance.

6. (a) H_0 : Les performances des trois types de montres sont identiques.
   H_1 : Les performances des trois types de montres sont différentes.
   On utilise la statistique de Kruskal-Wallis K avec $R_{1+} = 50$, $R_{2+} = 29$, $R_{3+} = 41$, $I = 3$ et $n = 5$. On observe $K_{obs} = \frac{12}{15 \cdot 16} \sum R^2_{i+}/5 - 3 \cdot 16 = 2.22$. La valeur critique de la loi khi-carré avec $3 - 1$ degrés de liberté est égale à $q\chi^2_2(95\%) = 5.991 > 2.22$, donc on ne peut pas rejeter l'hypothèse nulle.
   (b) On considère le modèle $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ avec $i = 1, 2, 3$ et $j = 1, 2, 3, 4, 5$. L'hypothèse nulle est alors $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$, c'est-à-dire les trois échantillons ont la même moyenne. Le tableau d'ANOVA construit avec `Splus` est (la première commande impose la contrainte $\sum_i \alpha_i = 0$) :

```
> options(contrasts=c('contr.sum','contr.sum'))
> montre <- c(0.1,0.08,0.22, ...   ,0.72)
> alpha <- factor(rep(1:3,c(5,5,5)))
> summary(aov(montre~alpha))
          Df Sum of Sq   Mean Sq  F Value     Pr(F)
alpha      2   1.96572 0.9828600 1.084886 0.3689013
Residuals 12  10.87148 0.9059567
```

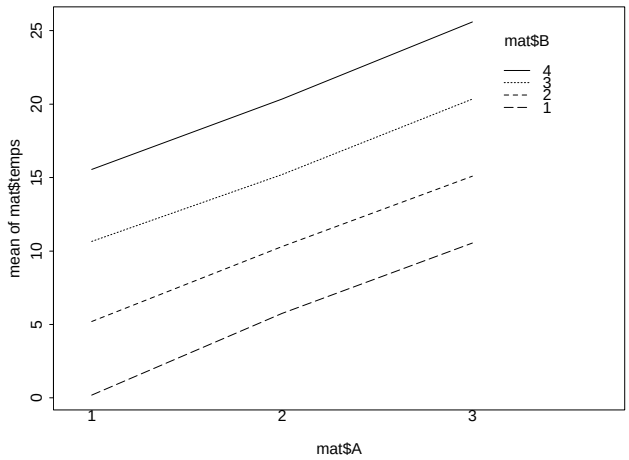
   La p-valeur du test de Fisher est supérieure à 5%, donc on ne peut pas rejeter l'hypothèse nulle.

7. (a) Les commandes suivantes permettent d'obtenir les résultats avec `Splus` (la première commande impose les contraintes $\sum_i \alpha_i = 0$ et $\sum_j \beta_j = 0$) :

```
> options(contrasts=c('contr.sum','contr.sum'))
> temps <- c(0.3,0.1,5.7,5.8, ... 25.8,25.4)
> A <- factor(rep(c(1,1,2,2,3,3),4))
> B <- factor(rep(1:4,c(6,6,6,6)))
> summary(aov(temps~A+B))
          df Sum of Sq  Mean Sq  F Value Pr(F)
A          2    400.00 200.0000 2535.211     0
B          3    756.36 252.1200 3195.887     0
Residuals 18      1.42   0.0789
```

Ainsi le facteur A et le facteur B sont hautement significatifs.

(b) La commande `interaction.plot(A,B,temps)` nous donne la figure suivante :



Comme les droites sont quasiment parallèles, on peut conclure qu'il n'y a pas d'interaction. La p-valeur calculée avec `Splus` pour l'hypothèse nulle : "il n'y a pas d'interaction" est d'ailleurs 23.1%.

(c) Le test de Kruskal-Wallis permet d'effectuer une analyse de variance à une voie. On procède donc en deux étapes : on teste si le facteur A est significatif, puis la même démarche pour le facteur B.

Facteur A : On a $I = 3$, $n = 8$ et $N = 24$. On trouve $R_{1+} = 65.5$, $R_{2+} = 98$ et $R_{3+} = 136.5$. Donc la statistique de test de Kruskal-Wallis est $K_{\text{obs}} = 6.32 > 5.991 = q\chi^2_2(95\%)$ et on peut rejeter l'hypothèse H_0 : "le facteur A n'est pas significatif" avec une p-valeur de 4.2%.

Facteur B : On a $I = 4$, $n = 6$ et $N = 24$. On trouve $R_{1+} = 34$, $R_{2+} = 52$, $R_{3+} = 91$ et $R_{4+} = 123$. Donc la statistique de test de Kruskal-Wallis est $K_{\text{obs}} = 15.9 > 7.815 = q\chi^2_3(95\%)$ et on peut rejeter l'hypothèse H_0 : "le facteur B n'est pas significatif" avec une p-valeur de 0.1%.

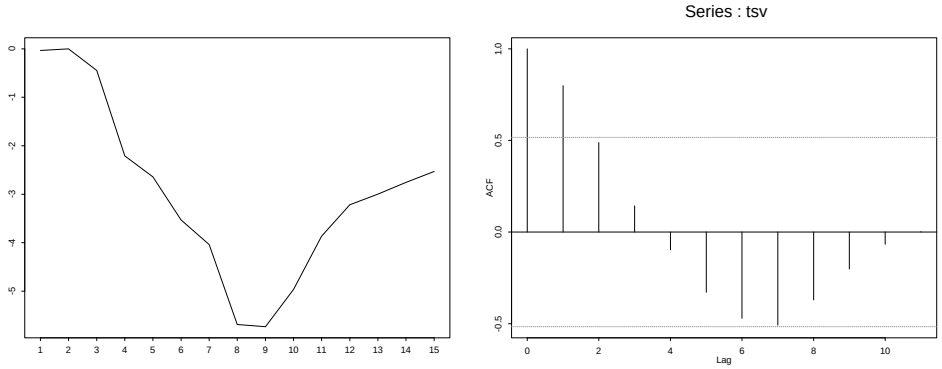
Les commandes `Splus` pour effectuer un test de Kruskal-Wallis sont :

```
> kruskal.test(temps, A)
> kruskal.test(temps, B)
```

Chapitre 16

Séries temporelles

1. Selon les graphiques, on constate une corrélation dans la série, car les autocorrélations ne sont pas toutes contenues dans l'intervalle de confiance $\left[\pm 1.96/\sqrt{n}\right]$.



2. Si $Y(t)$ est un processus de moyenne μ, alors on peut l'écrire comme $Y(t) = \mu + Y^*(t)$, où $Y^*(t)$ est un processus de moyenne 0. Donc $Y(t) - \mu$ est un processus AR(2) et les autocovariances s'écrivent

$$\begin{aligned}
\gamma_0 &= \alpha_1\gamma_1 + \alpha_2\gamma_2 + \sigma^2, \\
\gamma_1 &= \alpha_1\gamma_0 + \alpha_2\gamma_1\,, \\
\gamma_2 &= \alpha_1\gamma_1 + \alpha_2\gamma_0\,, \\
\gamma_3 &= \alpha_1\gamma_2 + \alpha_2\gamma_1\,, \\
&\vdots \\
\gamma_k &= \alpha_1\gamma_{k-1} + \alpha_2\gamma_{k-2}\,.
\end{aligned}$$

Pour les autocorrélations, on a pour $k = 0$

$$1 = \alpha_1\rho_1 + \alpha_2\rho_2 + \sigma^2/\gamma_0\,,$$

et pour $k > 0$, on remplace les γ_k par ρ_k (et on obtient les équations de Yule-Walker).

3. (a) Les équations de Yule-Walker pour un processus AR(2) s'écrivent :

$$\begin{aligned}
\rho_1 &= \alpha_1 + \alpha_2\rho_1\,, \\
\rho_2 &= \alpha_1\rho_1 + \alpha_2\,,
\end{aligned}$$

avec $\rho_1 = 0.354$ et $\rho_2 = 0.26$. En résolvant le système on obtient : $\hat{\alpha}_1 = 0.299$ et $\hat{\alpha}_2 = 0.154$.

(b) Comme $\bar{y} = 0$. $\hat{y}_{101} = \hat{\alpha}_1\hat{y}_{100} + \hat{\alpha}_2\hat{y}_{99} = -0.234$ et $\hat{y}_{102} = \hat{\alpha}_1\hat{y}_{101} + \hat{\alpha}_2\hat{y}_{100} = -0.207$.

4. (a) C'est la série des deuxièmes différences, car les observations ont une tendance quadratique. Dans la série des premières différences on constate encore une tendance linéaire. Par contre, dans les séries des différences supérieures, on ne voit plus de tendances au cours du temps.

(b) i. Pour $k = 0$, on a

$$1 = \alpha_1\rho_1 + \alpha_2\rho_2 + \alpha_3\rho_3 + \sigma^2/\gamma_0\,, \tag{1}$$

pour $k > 0$, on a les équations de Yule-Walker (*cf.* page 277(1997) ou 296(2001)).

ii. De (1), on tire

$$\gamma_0 = \frac{\sigma^2}{1 - \alpha_1\rho_1 - \alpha_2\rho_2 - \alpha_3\rho_3}\,.$$

iii. En se basant sur les équations de Yule-Walker, on ne doit que résoudre le système linéaire

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}. \tag{2}$$

(c) A partir de (2), on trouve $\hat{\alpha}_1 = 0.5282$, $\hat{\alpha}_2 = 0.3253$ et $\hat{\alpha}_3 = -0.0654$.

5. Soient les identités suivantes

$$\sum_{t=1}^{N} \cos\Big(\frac{2\pi pt}{N}\Big)\cos\Big(\frac{2\pi qt}{N}\Big) = \begin{cases} 0 & \text{si } 0 \le p \ne \lfloor N/2 \rfloor, \\ N/2 & \text{si } 0 < p = q < \lfloor N/2 \rfloor, \\ N & \text{sinon.} \end{cases}$$

$$\sum_{t=1}^{N} \sin\Big(\frac{2\pi pt}{N}\Big)\cos\Big(\frac{2\pi qt}{N}\Big) = 0, \tag{3}$$

$$\sum_{t=1}^{N} \sin\Big(\frac{2\pi pt}{N}\Big)\sin\Big(\frac{2\pi qt}{N}\Big) = \begin{cases} 0 & \text{si } 0 \le p \ne \lfloor N/2 \rfloor, \\ N/2 & \text{si } 0 < p = q < \lfloor N/2 \rfloor, \\ N & \text{sinon.} \end{cases}$$

qu'on peut prouver en utilisant la représentation

$$\cos(x) = \frac{1}{2}\big(\mathrm{e}^{ix} + \mathrm{e}^{-ix}\big), \quad \sin(x) = \frac{1}{2}\big(\mathrm{e}^{ix} - \mathrm{e}^{-ix}\big).$$

Afin d'obtenir l'orthogonalité des colonnes de la matrice sous (3), il faut que

$$\sum_{i=1}^{n} \cos\Big(\frac{2\pi pt_i}{2n}\Big)\cos\Big(\frac{2\pi qt_i}{2n}\Big) = 0,$$

$$\sum_{i=1}^{n} \cos\Big(\frac{2\pi pt_i}{2n}\Big)\sin\Big(\frac{2\pi qt_i}{2n}\Big) = 0,$$

$$\sum_{i=1}^{n} \sin\Big(\frac{2\pi pt_i}{2n}\Big)\sin\Big(\frac{2\pi qt_i}{2n}\Big) = 0,$$

avec $p, q = 2k$, $k = 0, 1, 2, \ldots, n-1$. Ainsi, on peut directement appliquer (3) en posant $N = 2n$.