ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

Année universitaire 2004 – 2005

EXERCICES

du cours de statistique et analyse de données

25 novembre 2005

La rédaction de ce polycopié a été coordonnée par Jean-François Delmas, professeur responsable du module, et réalisée par Jean-Pierre Raoult et l'équipe des enseignants de l'année universitaire 2004-2005 :

- **Jean-Yves Audibert**, École Nationale des Ponts et Chaussées.
- **Ali Chaouche**, École Nationale du Génie Rural et des Eaux et Forêts.
- **Didier Chauveau**, Université de Marne-la-Vallée.
- **Olivier De Cambry**, École Supérieure d'Ingénieurs en Électronique et Électrotechnique.
- **Jean-François Delmas**, École Nationale des Ponts et Chaussées.
- **Christian Derquenne**, Électricité de France (Recherche et Développement).
- **Marie-Pierre Etienne**, École Nationale du Génie Rural et des Eaux et Forêts.
- **Benjamin Jourdain**, École Nationale des Ponts et Chaussées.
- **Vincent Lefieux**, Réseau de Transport d'Électricité.
- **Eric Parent**, École Nationale du Génie Rural et des Eaux et Forêts.
- **Pierre Vandekerkhove**, Université de Marne-la-Vallée.

Table des matières

Chapitre I

Modèles paramétriques

I.1 Énoncés

Exercice I.1.
Reprenant la suite de 0 et de 1 donnée en cours (**Chapitre Modèle paramétrique**) considérez séparément les deux sous-suites de longueur 50 (première ligne, deuxième ligne). On adopte le modèle selon lequel les observations sont toutes indépendantes, suivant une loi de Bernoulli de paramètre p_a pour la première ligne et une loi de Bernoulli de paramètre p_b pour la deuxième ligne.

1. Calculez des estimations de p_a et p_b ; proposez des estimations des écart-types de chacun de ces estimateurs et donnez ces propriétés.
2. En admettant que 50 est un effectif assez élevé pour utiliser l'approximation normale, calculez, à différents niveaux de confiance $1-\alpha$ choisis par vous, des intervalles de confiance asymptotiques pour p_a et p_b ; on constate que, pour $1-\alpha$ assez faible, ces intervalles sont d'intersection vide ; à partir de quelle valeur de $1-\alpha$ cela se produit-il ?
3. Notons $\overline{X}_a = \frac{1}{n}\sum_{i=1}^{50} X_i$ et $\overline{X}_b = \frac{1}{n}\sum_{i=51}^{100} X_i$. Quelle est approximativement (toujours à l'aide des lois normales) la loi de $\overline{X}_a - \overline{X}_b$? En déduire un test de l'hypothèse $p_a = p_b$ (en remplaçant les variances de $\overline{X}_a$ et $\overline{X}_b$ par des approximations). Pour quelles valeurs de α rejette-t-on l'hypothèse ?
4. Pouvez-vous faire un lien entre les questions 2 et 3 ci-dessus ?
5. Reprenez le test élaboré à la question 3 sur les données figurant dans le texte suivant, extrait d'un article du journal *Le Monde* du 9 février 2003, intitulé *Faut-il traiter la ménopause ?* et relatif aux THS (*Traitements Hormonaux Substitutifs de la ménopause*). Faites un commentaire critique de ce passage.

La dernière étude américaine estime notamment que, chez 10 000 femmes âgées de 50 à 70 ans et non traitées, 450 souffriront d'un cancer du sein, alors que dans la situation inverse, dans un groupe de 10 000 femmes traitées pendant 5 ans, on observera 8 cas supplémentaires par an. Ces chiffres peuvent paraître très faibles et expliquent pourquoi ils sont restés si longtemps indétectables par des médecins isolés. Cependant, appliqués aux 10 millions d'utilisatrices américaines, ils sont évidement inadmissibles (8000 cas supplémentaires par an).

△

Exercice I.2.
On observe n variables aléatoires X_i indépendantes toutes régies par une loi exponentielle de paramètre $\theta \in]0, +\infty[$.

1. Ecrivez le modèle statistique correspondant. Démontrez que $\sum_{i=1}^n X_i$ est une statistique exhaustive. Rappelez quelle est sa loi.

2. On veut estimer la valeur commune des $\mathbb{E}_\theta(X_i)$; rappelez l'expression de cette espérance mathématique en fonction de θ. Donnez toutes les justifications auxquelles vous pouvez penser à l'aide du cours pour utiliser ici l'estimateur empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$; calculez son risque quadratique.

3. On veut estimer θ. Pensez-vous que l'estimateur auquel conduit naturellement l'étude faite en **2** soit sans biais ? Pour préciser ce point, et calculer le biais éventuel, utilisez la propriété : *si une v.a. Y suit la loi Gamma de paramètre (a, θ) (notée dans le cours $\mathcal{G}_{a,\theta}$), $\mathbb{E}(Y^r)$ est défini pour $a + r > 0$ et vaut $\frac{\Gamma(a+r)}{\theta^r \Gamma(a)}$; en particulier, si $a > 1$, pour $r = -1$, on obtient $\mathbb{E}(\frac{1}{Y}) = \frac{\theta}{a-1}$.*

4. Donnez le principe de fabrication des tests, au niveau α, des hypothèses $[\theta = 1]$ et $[\theta \leq 1]$. Précisez leur mise en œuvre pour $\alpha = 0,05$ et $n = 15$ (puis $n = 30$) : utilisez pour cela la table des fractiles de la loi du χ^2 donnée en appendice, en vous servant des deux propriétés suivantes :

• si Y suit la loi Gamma de paramètre $(n, 1)$, $2Y$ suit la loi du χ^2 à $2n$ degrés de liberté,

• pour $k > 30$, si Z suit la loi du χ^2 à k degrés de liberté, $\sqrt{2Z} - \sqrt{2k-1}$ suit approximativement la loi normale centrée réduite.

5. Donnez le principe de fabrication d'un intervalle de confiance, au niveau de confiance $1 - \alpha$, pour θ (utilisez pour cela le fait que, quel que soit θ, si Y suit la loi Gamma de paramètre (a, θ), θY suit la loi Gamma de paramètre $(a, 1)$). En réutilisant les lectures de tables faites en **4**, précisez la mise en œuvre de l'intervalle de confiance pour $1 - \alpha = 0,95$ et $n = 15$ (puis $n = 30$).

6. On utilise un modèle bayésien, en prenant pour probabilité a priori la loi Gamma $\mathcal{G}_{b,\eta}$. Démontrez que la probabilité a posteriori est $\mathcal{G}_{b+n,\eta+nm}$. En déduire l'estimateur bayésien de θ. Que retrouve-t-on approximativement pour de "grandes" tailles d'échantillon ?

7. En quoi les études précédentes auraient-elles été modifiées si on avait observé n v.a. de loi Gamma de paramètre (a, θ), avec a connu ?

△

Exercice I.3.
On observe n variables aléatoires X_i indépendantes toutes régies par la loi uniforme sur l'intervalle $[0, \theta]$, $\mathcal{U}_{0,\theta}$ (voir **R.5.7**). θ (> 0) est inconnu.

1. Ecrivez le modèle statistique correspondant. Démontrez que $Y = \sup(X_1, \ldots, X_n)$ définit une statistique exhaustive. Précisez sa loi en en fournissant la fonction de répartition, la densité (par rapport à la mesure de Lebsgue sur $\mathbb{R}_+$), l'espérance mathématique et la variance.

2. On veut estimer θ.
a. Trouvez l'estimateur par maximum de vraisemblance ; est-il sans biais ? sinon, comment le "corriger" pour le "débiaiser" (en conservant le fait qu'il s'agit d'un estimateur fondé sur la statistique exhaustive mise en évidence en **1**) ? Calculez les risques quadratiques de ces deux estimateurs et comparez les.
b. Trouvez un estimateur sans biais de θ fondé non sur $Y = \sup(X_1, \ldots, X_n)$ mais sur $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Comparez son risque quadratique avec ceux des estimateurs étudiés en **a**.

3. Donnez le principe de fabrication de tests, au niveau α, des hypohèses $[\theta = 1]$ et $[\theta \leq 1]$. Précisez leur mise en œuvre pour $\alpha = 0,05$ et $n = 15$.

4. Donnez le principe de fabrication d'un intervalle de confiance, au niveau de confiance $1 - \alpha$, pour θ (utilisez pour cela le fait que, quel que soit θ, $\theta^n Y$ suit une loi ne dépendant plus du paramètre θ) . En réutilisant les calculs faits en **3**, précisez la mise en œuvre pour $\alpha = 0,05$ et $n = 15$.

△

Exercice I.4.
(Cet exercice est extrait de la session d'examen de juin 2002 du module de Statistique et Analyse des données de l'ENPC. Il était acompagné d'un extrait de tables de la fonction de répartition de lois de Poisson, fourni ici en Annexe)

On effectue n observations $(x_1, \ldots, x_n)$ indépendantes à valeurs entières positives ou nulles, supposées toutes suivre une même loi de Poisson de paramètre inconnu θ, loi notée $\mathcal{P}_\theta$.

1. La statistique $(x_1, \ldots, x_n) \mapsto x_1 + \cdots + x_n$ est-elle exhaustive ? Si oui, pourquoi ? Quelle est sa loi ?

2. Proposez un estimateur de θ, en en indiquant (sans démonstration) des propriétés. Calculez son risque quadratique.

3. Pour un échantillon de taille 15, on a observé $x_1 + \cdots + x_n = 26$. Testez, au seuil 0,01, l'hypothèse nulle $\theta \leq 1$; vous justifierez votre méthode en vous appuyant sur la propriété suivante (admise), où on note F_θ la fonction de répartition de $\mathcal{P}_\theta$: pour tout $x > 0$, l'application $\theta \mapsto F_\theta(x)$ est strictement décroissante.

△

Exercice I.5.
(Cet exercice est extrait de la session d'examen de rappel (septembre 2002) du module de Statistique et Analyse des données de l'ENPC. Son corrigé n'est pas fourni dans ce fascicule.)

La loi de Pareto de paramètre de forme α (> 1) et de paramètre d'échelle β (> 0) est

donnée par sa densité, définie sur $\mathbb{R}_+^*$ ($=]0, \infty[$) par :

$$f_{\alpha,\beta}(x) = \frac{(\alpha-1)\beta^{\alpha-1}}{x^\alpha}\mathbf{1}_{[\beta,\infty[}(x) \ ,$$

où $\mathbf{1}_{[\beta,\infty[}$ désigne la fonction indicatrice de l'intervalle $[\beta, \infty[$.

On effectue n observations indépendantes, $(x_1, \ldots, x_n)$, selon une telle loi, les paramètres étant inconnus.

1.a. Donnez une application de $(\mathbb{R}_+^*)^n$ dans $(\mathbb{R}_+^*)^2$ qui soit un résumé exhaustif des observations.

b. Donnez l'estimation du couple (α, β) par maximum de vraisemblance (m.v.).

N.B. : On pourra commencer par chercher séparément l'estimation m.v. de α si β est connu, puis l'estimation m.v. de β si α est connu.

2. Le paramètre d'échelle β est ici supposé connu ; sans perte de généralité on le prendra égal à 1 (ceci revient à remplacer chaque x_i par $\frac{x_i}{\beta}$).

a. Démontrez (ou admettez) que, si la variable aléatoire (v.a.) X suit la loi de Pareto de paramètres α et 1, la loi de $\ln X$ (logarithme népérien de X) est la loi exponentielle de paramètre $\alpha - 1$.

b. Rappelez quelle est la loi de $\sum_{i=1}^n Y_i$ quand les Y_i sont des v.a. indépendantes et toutes de loi exponentielle de paramètre $\alpha - 1$.

c. Déduisez-en une technique de test de l'hypothèse nulle $[\alpha \geq 2]$.

△

Exercice I.6.
(Cet exercice est extrait de la session d'examen de juin 2003 du module de Statistique et Analyse des données de l'ENPC)

On considère $(X_1, \cdots, X_n)$, n variables aléatoires indépendantes et identiquement distribuées (ou n-échantillon), de loi exponentielle décalée, c'est-à-dire de paramètre $\theta = (\alpha, \beta)$ avec pour densité : $f_{\alpha,\beta}(x) = \mathbf{1}_{[\alpha,+\infty[}(x)\beta e^{-\beta(x-\alpha)}$ (où $\alpha \in \mathbb{R}$, $\beta > 0$ et $\mathbf{1}_{[\alpha,+\infty[}$ désigne la fonction indicatrice de l'intervalle $[\alpha, +\infty[$).

a. Proposez une statistique exhaustive bidimensionnelle pour ce modèle.

b. Fournissez un estimateur par maximum de vraisemblance du couple (α, β).

c. Proposez des estimateurs de l'espérance mathématique et de la variance fondés sur les résultats de **b** .

d. Critiquez et complétez librement l'étude qui vient d'être faite de ce modèle.

△

Exercice I.7.

1. Ecrire la densité de la loi de la variable alatoire $X = \theta Y$, où $\theta > 0$ et Y suit la loi du χ^2 à k degrés de liberté.

2. On effectue n observations $(x_1, \dots x_n)$ selon la loi de paramètre θ considérée à la question précédente.

a. Donner une statistique exhaustive à valeurs dans $\mathbb{R}_+$.

b. Donner un estimateur sans biais de θ fondé sur cette statistique exhaustive.

c. Retrouvez les résulats de a et b *sans utiliser l'expression de la densité de la loi commune des* X_i mais en utilisant la définition de la loi $\chi^2(k)$.

3. Dire, en quelques lignes, comment se présente dans ce modèle un test de l'hypothèse nulle $\theta \leq 0$ contre l'hypothèse alternative $\theta > 0$.

△

I.2 Corrections

Exercice I.1.

1. Les v.a. X_i $(1 \leq i \leq 50)$ sont indépendantes et de loi de Bernoulli de paramètre p_a. Donc la v.a. moyenne empirique $\bar{X}_a = \frac{1}{50}\sum_{i=1}^{50} X_i$ est l'estimateur par maximum de vraisemblance de p_a. Comme $\forall p_a \in [0,1]$, $\mathbb{E}_{p_a}[\bar{X}_a] = p_a$, l'estimateur est sans biais. Comme (LFGN) $\mathbb{P}_{p_a}$-p.s. $\lim_{n\to\infty} \bar{X}_a = p_a$, l'estimateur est convergent. On peut vérifier grâce au TCL qu'il est asymptotiquement normal, de variance asymptotique $p_a(1-p_a)$.
   L'estimation de p_a vaut ici $\bar{x}_a = 0,08$. De même l'estimation de p_b vaut $\bar{x}_b = 0,14$.
   Chaque v.a. X_i $(1 \leq i \leq 50)$ a pour variance $p_a(1-p_a)$. Donc, les v.a. X_i étant indépendantes, la v.a. $\bar{X}_a$ a pour variance $\frac{1}{50^2}\sum_{i=1}^{50} p_a(1-p_a) = \frac{1}{50}p_a(1-p_a)$, dont l'estimateur du maximum de vraisemblance est $\frac{1}{50}\bar{X}_a(1-\bar{X}_a)$. D'où l'estimation de l'écart-type de $\bar{X}_a$: $[\frac{1}{50}\bar{x}_a(1-\bar{x}_a)]^{1/2}$. (La v.a. $\frac{1}{50}\bar{X}_a(1-\bar{X}_a)$ n'est pas un estimateur sans biais de la variance, ni non plus $\frac{1}{50}\bar{X}_a(1-\bar{X}_a)^{1/2}$ un estimateur sans biais de l'écart-type de $\bar{X}_a$. En revanche, il s'agit d'estimateurs convergents.)
   Numériquement, on a ici : $[\frac{1}{50}\bar{x}_a(1-\bar{x}_a)]^{1/2} = 0.052$. De même l'estimation de l'écart-type de $\bar{X}_b$ est $[\frac{1}{50}\bar{x}_b(1-\bar{x}_b)]^{1/2} = 0.063$. 0,06.

2. $\phi_{1-\frac{\alpha}{2}}$ désignant le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite, on prend pour intervalle de confiance (I.C.) pour p_a, au niveau de confiance $1-\alpha$:

$$[\bar{x}_a \pm \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_a(1-\bar{x}_a)}{50}}].$$

   On rappelle que ceci exprime que, quelle que soit la vraie valeur de p_a, on a :

$$\mathbb{P}_{p_a}(p_a \in [\bar{X}_a \pm \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_a(1-\bar{X}_a)}{50}}]) \simeq 1-\alpha.$$

   L'approximation étant dû au fait que la vraie loi de $\frac{\bar{X}_a - p_a}{\sqrt{\frac{p_a(1-p_a)}{50}}}$ a été remplacée par la loi limite normale centrée réduite.
   De même l'I.C., au niveau de confiance $1-\alpha$, pour p_b est :

$$[\bar{x}_b \pm \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_b(1-\bar{x}_b)}{50}}].$$

   Voici deux exemples de résultats numériques (la précision sur les bornes se limitant à 2 chiffres après la virgule en raison de l'approximation normale) :

Niveau de confiance	α	$\phi_{1-\frac{\alpha}{2}}$	I.C. pour p_a	I.C. pour p_b
0,95	0,05	1,96	[0,06 , 0,26]	[0,16 , 0,40]
0,90	0,10	1,65	[0,07 , 0,25]	[0,18 , 0,38]

Dans les deux cas ci-dessus ($1-\alpha = 0.95$ et $1-\alpha = 0.90$), les I.C. pour p_a et p_b ont une intersection non vide. Les longueurs des intervalles de confiance diminuent quand α augmente (et donc le niveau de confiance diminue) ; elles tendent vers 0 quand α tend vers 1 (situation limite où l'estimation par intervalle se réduit à l'estimation ponctuelle, avec donc une probabilité égale à 1 d'affirmer un résultat faux). Ces I.C. auront donc une intersection vide pour α assez grand (niveau de confiance assez faible), c'est-à-dire, puisqu'ici $\bar{x}_a < \bar{x}_b$, si α est tel que

$$\bar{x}_a + \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_a(1-\bar{x}_a)}{50}} < \bar{x}_b - \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_b(1-\bar{x}_b)}{50}}$$

c'est-à-dire :

$$\phi_{1-\frac{\alpha}{2}} < \frac{\bar{x}_b - \bar{x}_a}{\sqrt{\frac{\bar{x}_a(1-\bar{x}_a)}{50}} + \sqrt{\frac{\bar{x}_b(1-\bar{x}_b)}{50}}} ,$$

soit ici

$$\phi_{1-\frac{\alpha}{2}} < 1.04$$

ou encore, comme par définition $\phi_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ (où Φ^{-1} est la fonction réciproque de le fonction de répartition de la loi normale centrée réduite)

$$1 - \frac{\alpha}{2} = 0.85$$

c'est-à-dire enfin

$$1 - \alpha < 0.70 .$$

C'est donc au niveau de confiance (très mauvais) de 0.70 (ou pire) que les intervalles de confiance pour p_a et p_b sont d'intersection vide.

3. Test de l'hypothèse $H_0 = \{p_a = p_b\}$ contre $H_1 = \{p_a \neq p_b\}$.

   Les v.a. $\bar{X}_a$ et $\bar{X}_b$ ont respectivement pour lois approchées les lois normales $\mathcal{N}(p_a, \sigma_a^2)$ et $\mathcal{N}(p_b, \sigma_b^2)$, où $\sigma_a^2 = \frac{1}{50}p_a(1-p_a)$ et $\sigma_b^2 = \frac{1}{50}p_b(1-p_b)$; comme $\bar{X}_a$ et $\bar{X}_b$ sont indépendantes, la différence $\bar{X}_a - \bar{X}_b$ a pour loi $\mathcal{N}(p_a - p_b, \sigma_a^2 + \sigma_b^2)$.

   Sous l'hypothèse nulle H_0 (notons $p = p_a = p_b$ la valeur commune du paramètre) on a aussi $\sigma_a^2 = \sigma_b^2 = \frac{1}{50}\sum_{i=1}^{50} p(1-p)$ (notons σ^2 cette valeur commune) et donc la loi de $\bar{X}_a - \bar{X}_b$ est $\mathcal{N}(0, 2\sigma^2)$, qu'on approche par $\mathcal{N}(0, 2s^2)$, où $s^2 = \frac{1}{50}\bar{x}(1-\bar{x})$, avec $\bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i = \frac{1}{2}(\bar{x}_a + \bar{x}_b)$; en effet, sous l'hypothèse nulle, toutes les v.a. X_i (où $1 \leq i \leq 100$) sont de même loi de Bernoulli de paramètre p.

   En revanche, sous l'hypothèse alternative, l'espérance de la loi de $\bar{X}_a - \bar{X}_b$ est non nulle ; il est donc naturel de bâtir un test où le rejet de l'hypothèse nulle s'effectue si $|\bar{x}_a - \bar{x}_b|$ est assez élevé, c'est-à-dire si $|\bar{x}_a - \bar{x}_b| > c$, où c est adapté au niveau de signification choisi.

   Sous l'hypothèse nulle, la loi de $\frac{\bar{X}_a - \bar{X}_b}{\sqrt{2}S}$ (où $S^2 = \frac{1}{50}\bar{X}(1-\bar{X})$) est approximativement $\mathcal{N}(0,1)$ et donc on approxime $\frac{c}{\sqrt{2}s}$ par $\phi_{1-\frac{\alpha}{2}}$, quantile d'ordre $1-\frac{\alpha}{2}$ de $\mathcal{N}(0,1)$

   Ici, numériquement, $\bar{x}_a - \bar{x}_b = -0.12$ et $\sqrt{2}s = 0,08$; donc :
   - si $\alpha = 0.05$, $\phi_{1-\frac{\alpha}{2}} = 1.96$ d'où $c = 0.08 \times 1.96 = 0.16$; comme $0.12 < 0.16$, **il n'y a pas de rejet de l'hypothèse nulle** ;

– si $\alpha = 0.10$, $\phi_{1-\frac{\alpha}{2}} = 1.65$ d'où $c = 0.08 \times 1.65 = 0,13$; ici encore $0.10 < 0.13$, donc **il n'y a pas de rejet de l'hypothèse nulle**.

Envisageons d'autres valeurs de α : il y aurait rejet de l'hypothèse nulle, à partir des données observées $\bar{x}_a - \bar{x}_b = -0.12$ et $\sqrt{2}s = 0.08$, si

$$\phi_{1-\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2}) \leq \frac{0.12}{0.08} = 1.5$$

autrement dit

$$1 - \frac{\alpha}{2} \leq \Phi(1.25) = 0.93$$

ou encore $\alpha \geq 0.14$. En d'autres termes la p-valeur associée aux observations est 0.14. C'est donc avec un risque énorme que l'on rejetterait ici l'hypothèse nulle.

4. Il est évident que constater une intersection vide entre les I.C. pour p_a et p_b "donne envie" de conclure que ces deux paramètres sont différents. Demandons nous donc si, au moins approximativement, conclure ainsi à partir de l'intersection vide des I.C. au niveau de confiance $1 - \alpha$ revient au même que rejeter, par un test au niveau de signification α', l'hypothèse nulle $H_0 = \{p_a = p_b\}$. La première méthode (intersection des I.C.) rejette H_0 si $|\bar{x}_a - \bar{x}_b| > \phi_{1-\frac{\alpha}{2}}(s_a + s_b)$ et la seconde (test) si $|\bar{x}_a - \bar{x}_b| > \phi_{1-\frac{\alpha'}{2}}\sqrt{2}s$ où, rappelons le,

$$s_a^2 = \frac{1}{50}\bar{x}_a(1 - \bar{x}_a) \quad , \quad s_b^2 = \frac{1}{50}\bar{x}_b(1 - \bar{x}_b) \quad \text{et} \quad s^2 = \frac{1}{50}\bar{x}(1 - \bar{x}) \; .$$

Or sous l'hypothèse nulle on a les égalités approximatives $\bar{x}_a \sim \bar{x}_b \sim \bar{x}$, d'où $\bar{x}_a + \bar{x}_b \sim 2\bar{x}$. Donc les deux méthodes conduisent approximativement aux mêmes rejets si $\phi_{1-\frac{\alpha'}{2}} = \sqrt{2}\phi_{1-\frac{\alpha}{2}}$. (La différence sur le calcul des niveaux de rejet provient du fait que pour la méthode des I.C. on s'intéresse à une precision des estimations de p_a et de p_b, alors que pour le test on s'intéresse à une précision sur l'estimation de $p_a - p_b$.)

5. L'article ne précise pas d'où viennent ces "estimations", et on comprend mal l'emploi du futur dans ce texte (*...souffriront ..., ... on observera...*). Admettons que l'étude ait porté sur 2 échantillons de 10 000 femmes chacun (ce qui représente des échantillons très gros, mais accessibles par des enquêtes épidémiologiques à grande échelle) et demandons nous pour quelles valeurs du niveau de signification la différence observée entre ces deux échantillons permettrait de conclure significativement à une différence, induite par la prise du THS, entre les probabilités de développer un cancer du sein.

On se trouve dans la situation étudiée en 3) avec ici : $n = 10000$, $\bar{x}_a = 0.0450$, $\bar{x}_b = 0.0458$, d'où

$|\bar{x}_a - \bar{x}_b| = 0.0008$, $\bar{x} = 0.0454$, $s^2 = 4.33.10^{-6}$, $\sqrt{2}s = 2.94.10^{-3}$

et enfin

$$\frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{2}s} = 0.27.$$

Vu la taille des échantillons, l'approximation, sous l'hypothèse nulle $H_0 = \{p_a = p_b\}$, de la loi de $\frac{|\bar{X}_a - \bar{X}_b|}{\sqrt{2}S}$ par la loi normale centrée réduite est excellente.

La p-valeur est ici la probabilité qu'une réalisation de la loi normale centrée réduite dépasse en valeur absolue 0.27. Elle vaut 0.394. C'est donc avec un risque énorme (presque "4 chances sur 10") que l'on rejetterait ici l'hypothèse nulle. En particulier le

rejet ne serait permis pour aucun des niveaux de signification couramment pratiqués. Ceci met donc gravement en cause la pertinence des conclusions rapportées par cet article, et l'extrapolation aux 10 millions d'américaines suivant ce type de traitement paraît sans fondement.

On peut se demander de quelle taille (supposée commune) n devraient être les deux échantillons (femmes traitées et femmes témoins) pour que des valeurs de $\bar{x}_a$ et $\bar{x}_b$ égales à celles observées ici conduisent à conclure à une différence significative, au seuil de signification usuel de 0.05. Il faudrait que :

$$\frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{2\bar{x}(1-\bar{x})}{n}}} \geq 1.96$$

c'est-à-dire

$$n \geq (1,96)^2 \frac{2\bar{x}(1-\bar{x})}{(\bar{x}_a - \bar{x}_b)^2}$$

soit ici

$$n \geq (1.96)^2 \frac{2 \times 0.0454 \times 0.9546}{(0.0008)^2} \simeq 520300 \ .$$

Sauf information contraire, on peut douter que l'étude ait été menée sur deux échantillons d'effectifs aussi élevés. Précisons d'ailleurs que la suite de l'article du Monde, même si elle ne présente pas l'analyse statistique que nous venons de faire, est assez réservée ; en particulier elle explique pourquoi cette étude américaine ne peut s'appliquer au cas de la France (différence de nature dans la composition des hormones substitutives) et cite à l'appui de cette critique des positions d'autorités médicales françaises.

▲

Exercice I.2.

1. Le modèle ; une statistique exhaustive

La loi exponentielle de paramètre θ admet pour densité, par rapport à la mesure de Lebesgue sur $\mathbb{R}$, l'application $p(x, \theta)$ définie par :

$$p(x, \theta) = \theta \exp(-\theta x) \mathbf{1}_{[0,+\infty[}(x) \ ,$$

où $\mathbf{1}_{[0,+\infty[}$ désigne la fonction indicatrice de la demi-droite $[0, +\infty[$.

On en déduit une densité, par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, de la suite finie $(X_1, \ldots, X_n)$, composée de v.a. indépendantes et de même loi exponentielle de paramètre θ :

$$p_n(x_1, \ldots, x_n, \theta) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \mathbf{1}_{[0,+\infty[^n}(x_1, \ldots, x_n) \ .$$

Cette densité se factorise sous la forme :

$$p_n(x_1, \ldots, x_n, \theta) = \psi\left(\sum_{i=1}^{n} x_i, \theta\right) l(x_1, \ldots, x_n)$$

avec $\psi(y,\theta) = \theta^n \exp(-\theta y)$ et $l(x_1,\ldots,x_n) = \mathbf{1}_{[0,+\infty[}(\min(x_1,\ldots,x_n))$.

On reconnait ainsi (théorème de Halmos-Savage) que la statistique $Y = \sum_{i=1}^n X_i$ est exhaustive dans ce modèle.

La loi de Y est (résultat classique de calcul des probabilités) le loi Gamma de paramètres n et θ, de densité

$$y \mapsto \frac{\theta^n y^{n-1}}{(n-1)!} \exp(-\theta y) \mathbf{1}_{[0,+\infty[}(y)$$

.

2. Estimation de l'espérance mathématique des X_i

Pour tout i, on a $\mathbb{E}_\theta(X_i) = \frac{1}{\theta}$ d'où : $\forall \theta > 0\ \mathbb{E}_\theta(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{\theta}$.

Donc, *comme c'est toujours le cas pour des observations i.i.d. dont la loi commune admet une espérance mathématique finie*, la moyenne empirique des éléments de l'échantillon observé fournit une estimation sans biais de cette espérance mathématique.

On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Cet estimateur $\overline{X}_n$ est de manière évidente fonction de la statistique exhaustive mise en évidence en **1**. Il est fortement convergent d'après la loi forte des grands nombres et asymptotiquement normal d'après le théorème de la limite centrale.

$\overline{X}_n$ est aussi, dans ce modèle, l'estimateur du maximum de vraisemblance de $\frac{1}{\theta}$. Pour l'établir, considérons la log-vraisemblance de ce modèle qui est définie, pour $\theta > 0$ et $x = (x_1,\ldots,x_n) \in]0,+\infty[^n$, par :

$$\ell_n(x,\theta) = \ln p_n(x_1,\ldots,x_n,\theta) = n\ln\theta - \theta\sum_{i=1}^n x_i$$

L'application $\ell_n(x,.)$, de $\mathbb{R}_+^*$ dans $\mathbb{R}$ est dérivable ; sa dérivée est $\theta \mapsto \frac{n}{\theta} - \sum_{i=1}^n x_i$, qui s'annulle pour $\theta = \frac{n}{\sum_{i=1}^n x_i}$, dont on vérifie que c'est bien un maximum de $\ell_n(x,.)$. Donc l'estimateur du maximum de vraisemblance de θ est $\frac{1}{\overline{X}_n}$. On en déduit que l'estimateur du maximum de vraisemblance de $\frac{1}{\theta}$ est $\overline{X}_n$.

N.B. Le fait de se limiter à des observations strictement positives n'est pas gênant car, quel que soit θ, la probabilité de l'évènement $[\forall i\ X_i > 0]$ est égale à 1.

Le risque quadratique de cet estimateur est, comme pour tout estimateur sans biais, sa variance :

$$\mathrm{Var}_\theta\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\theta(X_i) = \frac{1}{n^2}\frac{n}{\theta^2} = \frac{1}{n\theta^2}\ ,$$

qui tend vers 0 quand n tend vers l'infini.

3. Estimation du paramètre θ

L'estimateur du maximum de vraisemblance de θ est, on vient de le voir, $\frac{1}{\overline{X}_n} = \frac{n}{\sum_{i=1}^n X_i}$.

Mais cet estimateur n'a "aucune raison" d'être sans biais. En effet, ce n'est que pour les applications affines qu'on sait que le caractère sans biais "passe bien" : si une v.a. Y estime sans biais une certaine fonction $\phi(\theta)$ du paramètre, alors $aY + b$ estime sans biais $a\phi(\theta) + b$.

De fait ici (voir **1**) $\sum_{i=1}^n X_i$ suit la loi Gamma $\mathcal{G}(n, \theta)$ dont on sait que le moment d'ordre -1 est $\frac{\theta}{n-1}$. Donc $\mathbb{E}_\theta(\frac{1}{\overline{X}_n}) = n\mathbb{E}_\theta(\frac{1}{\sum_{i=1}^n X_i}) = \frac{n}{n-1}\theta$ (et non pas θ qui serait nécessaire pour que cet estimateur soit sans biais).

Le biais de $\frac{1}{\overline{X}_n}$ en tant qu'estimateur de θ est $\mathbb{E}_\theta(\frac{1}{\overline{X}_n}) - \theta = \frac{\theta}{n-1}$; il est strictement positif : on dit que $\frac{1}{\overline{X}_n}$ est biaisé par excès ; il tend vers 0 quand n tend vers l'infini : on dit que $\frac{1}{\overline{X}_n}$ est un estimateur asymptotiquement sans biais de θ.

Enfin on dispose, à l'évidence, d'un estimateur sans biais de θ : c'est $\frac{n-1}{\sum_{i=1}^n X_i}$.

4. Tests

a. Hypothèse nulle $[\theta = 1]$.

Pour tester cette hypothèse nulle, qu'on va plutôt écrire $[\frac{1}{\theta} = 1]$, contre l'hypothèse alternative $[\frac{1}{\theta} \neq 1]$, il est naturel de procéder au rejet si $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$, estimation sans biais de $\frac{1}{\theta}$, est assez loin de 1, autrement dit si $\overline{x}_n < c_1$ ou $\overline{x}_n > c_2$, avec $c_1 < 1 < c_2$, ces valeurs c_1 et c_2 étant à adapter au niveau de signifiation adopté pour le test.

Or, sous l'hypothèse nulle, la v.a. $2n\overline{X}_n = 2\sum_{i=1}^n X_i$ suit la loi du χ^2 à $2n$ degrés de liberté. On doit choisir c_1 et c_2 de sorte que $\mathbb{P}_1([\overline{X}_n \notin [c_1, c_2]]) = \alpha$ ce qui s'écrit aussi $\mathbb{P}_1(2n\overline{X}_n \notin [2nc_1, 2nc_2]]) = \alpha$. Pour des raisons de symétrie, on prend respectivement pour $2nc_1$ et $2nc_2$ le quantile d' ordre $\frac{\alpha}{2}$ et le quantile d' ordre $1 - \frac{\alpha}{2}$ de la loi du χ^2 à $2n$ degrés de liberté.

Exemples : Soit $\alpha = 0,05$.
Pour $n = 15$, cela donne (voir la table de la loi du $\chi^2(30)$) $30c_1 = 16,8$ et $30c_2 = 47,0$, d'où le test : on rejette l' hypothèse $[\theta = 1]$ si $\frac{1}{15}\sum_{i=1}^{15} x_i \notin [0,56\ ,\ 1,57]$
Pour $n = 30$, on utilise l'approximation normale de la loi du χ^2 à nombre de degrés de liberté élevé, de sorte que les quantiles d'ordre 0,025 et 0,975 de $\chi^2(60)$ sont respectivement approchés par $\frac{(-1,96+\sqrt{119})^2}{2}$ et $\frac{(-1,96+\sqrt{119})^2}{2}$, c'est-à-dire 40,04 et 82,8, d'où le test : on rejette l'hypothèse $[\theta = 1]$ si $\frac{1}{30}\sum_{i=1}^{30} x_i \notin [0,67\ ,\ 1,38]$.

On remarque que la zone dans laquelle la valeur de la moyenne empirique conduit au rejet de l' hypothèse nulle grossit quand on passe de $n = 15$ à $n = 30$; c'est normal : mieux renseigné par un échantillon plus gros, on est, à seuil de signification fixé, "plus audacieux"

pour conclure que $\underline{x}_n$ est significativement distant de 1.

b. Hypothèse nulle $[\theta \leq 1]$.

Pour tester cette hypothèse nulle, qu'on va plutôt écrire $[\frac{1}{\theta} \geq 1]$, contre l'hypothèse alternative $[\frac{1}{\theta} < 1]$, il est naturel de procéder au rejet si $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$, estimation sans biais de $\frac{1}{\theta}$, est assez petit, autrement dit si $\overline{x}_n < c$, cette valeur c vérifiant, pour $\theta = 1$ (valeur frontière entre l'hypothèse nulle et l'hypothèse alternative) $\mathbb{P}_1([\overline{X}_n < c]) = \alpha$; alors, a fortiori, pour tout $\theta \leq 1$, on a $\mathbb{P}_\theta([\overline{X}_n < c]) \leq \alpha$; en effet la famille des lois Gamma de paramètre de taille n fixé, c'est-à-dire $(\mathcal{G}(n,\theta))_{\theta\in\mathbb{R}_+^*}$, est stochastiquement décroissante : si on note $G_{n,\theta}$ la fonction de répartition de $\mathcal{G}(n,\theta)$, on a, si $\theta < \theta'$, pour tout $x > 0$, $G_{n,\theta}(x) < G_{n,\theta'}(x)$.

De manière analogue à l'étude faite en **a** ci-dessus, on établit que $2nc$ est le quantile d'ordre α de la loi du χ^2 à $2n$ degrés de liberté.

Exemples : Soit $\alpha = 0,05$.
Pour $n = 15$, cela donne (voir la table de la loi du $\chi^2(30)$) $30c = 18,5$, d'où le test : on rejette l' hypothèse $[\theta \leq 1]$ si $\frac{1}{15}\sum_{i=1}^{15} x_i < 0,62$.
Pour $n = 30$, le quantile d'ordre 0,05 de $\chi^2(60)$ est approché par $\frac{(-1,65+\sqrt{119})^2}{2} = 42,86$, d'où le test : on rejette l' hypothèse $[\theta \leq 1]$ si $\frac{1}{30}\sum_{i=1}^{30} x_i < 0,71$.

5. Intervalle de confiance

Pour tout θ, la loi de la v.a. $\theta\sum_{i=1}^n X_i$ est $\mathcal{G}(n,1)$, qui ne dépend plus de θ et peut donc nous servir de "pivot" pour construire un intervalle de confiance. En effet, si on note $\gamma_{n,\frac{\alpha}{2}}$ et $\gamma_{n,1-\frac{\alpha}{2}}$ les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de $\mathcal{G}(n,1)$, on a :

$$\forall\theta \;\; P_\theta([\gamma_{n,\frac{\alpha}{2}} \leq \theta\sum_{i=1}^n X_i \leq \gamma_{n,1-\frac{\alpha}{2}}]) = 1-\alpha$$

autrement dit :

$$\forall\theta \;\; P_\theta\left([\frac{\gamma_{n,\frac{\alpha}{2}}}{\sum_{i=1}^n X_i} \leq \theta \leq \frac{\gamma_{n,1-\frac{\alpha}{2}}}{\sum_{i=1}^n X_i}]\right) = 1-\alpha\ .$$

L'intervalle de confiance, au niveau de confiance $1-\alpha$, est donc $[\frac{\gamma_{n,\frac{\alpha}{2}}}{\sum_{i=1}^n x_i}\ ,\ \frac{\gamma_{n,1-\frac{\alpha}{2}}}{\sum_{i=1}^n x_i}]$.

Exemple : utilisant les lectures de tables déjà faites en **4.a** ci-dessus, on obtient que, au niveau de confiance $0,95$ (donc pour $\alpha = 0,05$) on a :
- si $n = 15$, $\gamma_{15\ ,\ 0,025} = 8,40$ et $\gamma_{15\ ,\ 0,975} = 23,50$,
- si $n = 30$, $\gamma_{30\ ,\ 0,025} = 20,02$ et $\gamma_{30\ ,\ 0,975} = 41,44$.

6. Modèle bayésien

On rappelle (voir **1**) que, par rapport à la mesure de Lebesgue sur $\mathbb{R}_+^n$, on peut prendre pour densité, en l'observation $(x_1, \ldots, x_n)$ (tous ≥ 0)

$$p_n(x_1, \ldots, x_n, \theta) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right) ;$$

par ailleurs on adopte pour densité a priori de θ (où $\theta > 0$) :

$$g_{b,\eta}(\theta) = \frac{\eta^b}{\Gamma(b)} \theta^{b-1} \exp(-\eta\theta) .$$

La densité du couple $(\theta, (x_1, \ldots, x_n))$, par rapport à la mesure de Lebesgue sur $\mathbb{R}_+^* \times \mathbb{R}_+^n$, est donc le produit :

$$(\theta, (x_1, \ldots, x_n)) \mapsto p_n(x_1, \ldots, x_n, \theta) g_{b,\eta}(\theta) = \frac{\eta^b}{\Gamma(b)} \theta^{b+n-1} \exp\left(-\theta\left(\eta + \sum_{i=1}^n x_i\right)\right) .$$

La densité marginale

$$h_{b,\eta}(x_1, \ldots, x_n) = \int_0^{+\infty} p_n(x_1, \ldots, x_n, \theta) g_{b,\eta}(\theta) d\theta$$

n'a pas besoin d'être calculée maintenant ; nous importe essentiellement la densité a posteriori, étant observé $(x_1, \ldots, x_n)$, qui est, en remplaçant $\sum_{i=1}^n x_i$ par $n\overline{x}_n$:

$$k_{(x_1,\ldots,x_n),b,\eta}(\theta) = \frac{p_n(x_1, \ldots, x_n, \theta) g_{b,\eta}(\theta)}{h_{b,\eta}(x_1, \ldots, x_n)} = \frac{1}{h_{b,\eta}(x_1, \ldots, x_n)} \frac{\eta^b}{\Gamma(b)} \theta^{b+n-1} \exp(-\theta(\eta + n\overline{x}_n)) ,$$

où on reconnait la forme de la densité de la loi gamma $\mathcal{G}(b+n, \eta+n\overline{x}_n)$.
On a donc $\frac{1}{h_{b,\eta}(x_1,\ldots,x_n)} \frac{\eta^b}{\Gamma(b)} = \frac{(\eta+n\overline{x}_n)^{b+n}}{\Gamma(b+n)}$ (ce qui, accessoirement, fournit $h_{b,\eta}(x_1, \ldots, x_n)$).

L'estimation bayésienne de θ est, par définition, l'espérance mathématique de cette loi a posteriori, c'est-à-dire $\frac{b+n}{\eta+n\overline{x}_n}$; si n tend vers l'infini, à $\overline{x}_n$ fixé, cette estimation converge vers $\frac{1}{\overline{x}_n}$, c'est-à-dire l'estimateur biaisé de θ étudié en question **3**.

7. Observation de n v.a. i.i.d. de loi Gamma de paramètre (a, θ), avec a connu

La somme de ces n v.a. est encore exhaustive et de loi Gamma de paramètre (na, θ) ; donc toute l'étude menée ci-dessus reste valable, en y remplaçant n par na.

▲

Exercice I.3. **1. Le modèle ; une statistique exhaustive**

La loi uniforme sur $[0, \theta]$ admet pour densité, par rapport à la mesure de Lebesgue sur $\mathbb{R}$, l'application f_θ définie par :

$$f_\theta(x) = \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x) ,$$

où $\mathbf{1}_{[0,\theta]}$ désigne la fonction indicatrice de l'intervalle $[0,1]$. Sa fonction de répartition, F_θ, vérifie :
- si $x < 0$, $F_\theta(x) = 0$,
- si $0 \leq x \leq \theta$, $F_\theta(x) = \frac{x}{\theta}$,
- si $\theta < x$, $F_\theta(x) = 1$.

Les v.a. X_i étant indépendantes, la loi de $(X_1, \ldots, X_n)$ admet pour densité, par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, l'application f_θ^n définie par :

$$f_\theta^n(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x_i) \ ,$$

autrement écrit

$$f_\theta^n(x_1, \ldots, x_n) = \frac{1}{\theta^n} \mathbf{1}_{[0,\theta]}(\sup(x_1, \ldots, x_n)) \mathbf{1}_{[0,+\infty]}(\inf(x_1, \ldots, x_n)) \ .$$

Donc, par la méthode de Halmos-Savage (aussi dite "de factorisation"), il apparaît que la v.a. réelle $Y = \sup(X_1, \ldots, X_n)$ est une statistique exhaustive dans ce modèle. Ce résultat est cohérent avec l'intuition : la seule signification concrète du paramètre θ étant qu'il borne supérieurement les valeurs observables, une fois connue la valeur de la plus grande des observations, celles qui lui sont inférieures n'apportent aucune information complémentaire sur θ.

Soit $F_{\theta,n}$ la fonction de répartition de Y (avec en particulier : $F_{\theta,1} = F_\theta$) :

$$F_{\theta,n}(y) = P_\theta([\sup(X_1, \ldots, X_n) \leq y]) = P_\theta([\forall i \ X_i \leq y]) = \prod_{i=1}^n P_\theta([X_i \leq y]) = F_\theta(y)^n \ .$$

Donc :
- si $y < 0$, $F_{\theta,n}(y) = 0$,
- si $0 \leq y \leq \theta$, $F_{\theta,n}(y) = \frac{y^n}{\theta^n}$,
- si $\theta < y$, $F_{\theta,n}(y) = 1$.

La densité de la loi de Y, obtenue par dérivation (sauf en 0 et en 1) de $F_{\theta,n}$, est $f_{\theta,n}$ définie par :

$$f_{\theta,n}(y) = \frac{n}{\theta^n} y^{n-1} \mathbf{1}_{[0,\theta]}(y) \ .$$

Il en résulte élémentairement que :
$\mathbb{E}_\theta(Y) = \frac{n}{\theta^n} \int_0^\theta y.y^{n-1} dy = \frac{n}{n+1}\theta$,
$\mathbb{E}_\theta(Y^2) = \frac{n}{\theta^n} \int_0^\theta y^2.y^{n-1} dy = \frac{n}{n+2}\theta^2$,
$\mathrm{Var}_\theta(Y) = \mathbb{E}_\theta(Y^2) - (\mathbb{E}_\theta(Y))^2 = \frac{n}{(n+2)(n+1)^2}\theta^2$.

2. Estimation du paramètre

a. La vraisemblance s'obtient en considérant la densité de la statistique exhaustive Y. Etant observé $y > 0$, l'estimation par maximum de vraisemblance est le point en lequel prend son maximum la fonction de θ jj définie sur $\mathbb{R}_+^*$ par :
$\theta \mapsto f_{\theta,n}(y) = 0$ si $\theta < y$,
$= \frac{ny^{n-1}}{\theta^n}$ si $\theta \geq y$.

Le maximum est atteint en y (et vaut $\frac{n}{\theta}$). L'estimateur par maximum de vraisemblance est donc la v.a. Y.

N.B. *Cette situation, où le maximum est atteint en un point en lequel la vraisemblance n'est pas continue (et donc a fortiori pas dérivable) met en évidence la nocivité du "réflexe" qui consisterait à effectuer systématiquement la recherche du maximum par annulation de la dérivée.*

On constate que $\mathbb{E}(Y) < \theta$. L'estimateur m.v. est donc ici biaisé inférieurement, ce qui était prévisible puisque Y prend presque sûrement des valeurs strictement inférieures à θ.

Mais on remarque que : $\mathbb{E}(\frac{n+1}{n}Y) = \theta$. Notons $Z = \frac{n+1}{n}Y$; c'est un estimateur sans biais de θ.

Le risque quadratique de Y est :

$$R_Y(\theta) = \mathbb{E}_\theta((Y-\theta)^2) = \mathrm{Var}_\theta(Y) + (\theta - \mathbb{E}_\theta(Y))^2$$

d'où ici :

$$R_Y(\theta) = \theta^2[\frac{n}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2}] = \frac{2}{(n+2)(n+1)}\theta^2 \ .$$

Le risque quadratique de Z, estimateur sans biais, est :

$$R_Z(\theta) = \mathrm{Var}_\theta(Z) = (\frac{n+1}{n})^2 \mathrm{Var}_\theta(Y) = (\frac{n+1}{n})^2 \frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{1}{n(n+2)}\theta^2 \ .$$

On vérifie que, pour tout $n > 1$, on a : $\forall \theta \ R_Z(\theta) < R_Y(\theta)$: Z est meilleur que Y au sens du risque quadratique.

b. Pour tout i, on a $\mathbb{E}(X_i) = \frac{\theta}{2}$; donc $\mathbb{E}(\frac{2}{n}\sum_{i=1}^n X_i) = \theta$. Autrement dit $U = \frac{2}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de θ.

Le risque quadratique de U est :

$$R_U(\theta) = \mathrm{Var}_\theta(U) = \frac{4}{n^2} n \, \mathrm{Var}_\theta(X_1) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n} \ .$$

Cet estimateur est bien plus mauvais que Z (et même Y) ; asymptotiquement, son risque quadratique est de l'ordre de $\frac{1}{n}$ alors que ceux de Y et Z sont de l'ordre de $\frac{1}{n^2}$; ces mauvaises performances ne sont pas étonnantes puisqu'il ne se factorise pas à travers la statistique exhaustive.

3. Tests

a. Hypothèse nulle $[\theta \leq 1]$

Afin de tester l'hypothèse nulle $[\theta \leq 1]$ contre l'hypothèse alternative (dite unilatérale) $[\theta > 1]$, on remarque que, plus θ est grand, plus Y a tendance à prendre de grandes valeurs : précisément, pour tout n et tout $y > 0$, l'application $\theta \mapsto 1 - G_{\theta,n}(y)$ (probabilité de dépasser y) tend vers 0 en décroissant quand θ tend vers $+\infty$.

Il apparaît donc naturel de rejeter l'hypothèse nulle $[\theta \leq 1]$ quand θ est assez grand. Au niveau de signification α, la région de rejet est donc $]c, +\infty[$, où $P_1([Y > c]) = \alpha$; en d'autres termes, c est le quantile d'ordre $1 - \alpha$ de la loi de Y pour la valeur frontière (égale à 1) du paramètre .
c vérifie : $F_{1,n}(c) = c^n = 1 - \alpha$; donc $c = (1 - \alpha)^{1/n}$.

Exemple : $\alpha = 0,05$, n = 15 ; alors $c = (0,95)^{1/15} = 0,9966$; il ne faut pas s'étonner de voir ici $c < 1$: si y est "un tout petit peu" en dessous de 1, on a "tout lieu de penser" que $\theta > 1$.

b. Hypothèse nulle $[\theta = 1]$

Selon les mêmes considérations qu'en **a** ci-dessus, il apparaît naturel de rejeter l'hypothèse nulle $[\theta = 1]$ (l'hypothèse alternative, dire bilatérale, étant $[\theta \neq 1]$) quand y est trop faible ou trop élevé ; la région de non-rejet (en y) est donc de la forme $[c_1, c_2]$, où c_1 et c_2 vérifient : $P_1([c_1 \leq Y \leq c_2]) = 1 - \alpha$. Pour des raisons de symétrie, on prend pour c_1 le quantile d'ordre $\frac{\alpha}{2}$ et pour c_2 le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Y pour la valeur 1 du paramètre. Donc $c_1 = (\frac{\alpha}{2})^{1/n}$ et $c_2 = (1 - \frac{\alpha}{2})^{1/n}$. On remarque que 1 est dans la région de rejet, ce qui pouvait être attendu : si la plus grande valeur observée est égale à 1, c'est que la borne supérieure des valeurs observables, θ, est strictement plus grande que 1.

Exemple : $\alpha = 0,05$, n = 15 ; alors $c_1 = (0,025)^{1/15} = 0,7820$ et $c_2 = (0,975)^{1/15} = 0,9983$.

4. Intervalle de confiance

Si θ est la valeur du paramètre, la loi de $\frac{Y}{\theta}$ est celle de paramètre 1 (dont la fonction de répartition, $F_{1,n}$, a déjà été utilisée en **3** ci-dessus). Donc, avec les mêmes notations qu'en **3.b**, il vient :

$$\forall \theta \;\; P_\theta([c_1 \leq \frac{Y}{\theta} \leq c_2]) = 1 - \alpha$$

d'où

$$\forall \theta \;\; P_\theta([\frac{Y}{c_2} \leq \theta \leq \frac{Y}{c_1}]) = 1 - \alpha \; .$$

L'intervalle de confiance, au niveau de confiance $1-\alpha$, est donc $[\frac{\sup(x_1,\ldots,x_n)}{c_2}\ ,\ \frac{\sup(x_1,\ldots,x_n)}{c_1}]$. ▲

Exercice I.4.

1. Le modèle ; une statistique exhaustive.

Les observations étant indépendantes et de même loi de Poisson $\mathcal{P}_\theta$, où $\theta > 0$ la probabilité d'observer $(x_1,\ldots,x_n) \in \mathbb{N}^n$ est :

$$\prod_{i=1}^n \mathcal{P}_\theta(\{x_i\}) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{x_i}}{x_i!} = e^{-n\theta}\ \theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n \frac{1}{x_i!}$$

(on convient que $0^0 = 1$ et, comme $0^z = 0$ si $z > 0$, on obtient comme loi $\mathcal{P}_0$ la probabilité de Dirac en 0).

Cette probabilité est donc de la forme $g(\theta, \sum_{i=1}^n x_i)h(x_1,\ldots,x_n)$, ce qui assure l'exhaustivité de la statistique $\sum_{i=1}^n x_i$, par la méthode de Halmos-Savage qui est applicable car l'application $(x_1,\ldots,x_n) \mapsto \prod_{i=1}^n \mathcal{P}_\theta(\{x_i\})$ est la densité par rapport à la mesure de dénombrement sur l'ensembe infini dénombrable $\mathbb{N}^n$.

La loi de la somme de n variables aléatoires indépendantes et de même loi $\mathcal{P}_\theta$ est la loi de Poisson $\mathcal{P}_{n\theta}$.

Remarque : On peut aussi démontrer l'exhaustivité en revenant à sa définition et calculant explicitement, pour tout entier $y \geq 0$, la probabilité conditionnelle de l'observation $(x_1,\ldots,x_n)$, sachant que $\sum_{i=1}^n x_i = y$ et en vérifiant qu'elle ne dépend pas du paramètre ; en effet elle est nulle si $\sum_{i=1}^n x_i \neq y$, sinon elle vaut :

$$\frac{e^{-n\theta}\ \theta^y \prod_{i=1}^n \frac{1}{x_i!}}{e^{-n\theta}\ (n\theta)^y \frac{1}{y!}} = \frac{1}{n^y}\frac{y!}{\prod_{i=1}^n x_i!}\ .$$

On retrouve (voir VII.2 Lois de variables aléatoires remarquables), la loi multinomiale $\mathcal{M}_{y,\underline{p}}$, où $\underline{p}$ est la suite de longueur n dont tous les éléments sont égaux à $\frac{1}{n}$.

2. Estimation

Notons, comme il est traditionnel en calcul des probabilités, X_i la v.a. résultant en l'observation x_i. On sait que, pour tout θ, $\mathbb{E}_\theta(X_i) = \theta$ (et donc $\mathbb{E}_\theta(\sum_{i=1}^n X_i) = n\theta$). Il en résulte que $\frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de θ, fondé sur la statistique exhaustive mise en évidence à la question précédente. C'est l'estimateur dit *moyenne empirique.*

Vérifions que c'est un estimateur par maximum de vraisemblance (dit aussi ici, puisqu'il s'agit de lois discrètes, estimateur par maximum de probabilité). A $(x_1,\ldots,x_n)$ fixé, l'application définie sur $\mathbb{R}_+^*$ par :

$$\theta \mapsto e^{-n\theta}\ \theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n \frac{1}{x_i!}$$

admet son maximum (s'il existe et est unique) au même point que

$$\theta \mapsto -n\theta + (\sum_{i=1}^{n} x_i)\log(\theta) \ .$$

On constate (calcul élémentaire par annulation de la dérivée) que ce maximum est unique et atteint en $\frac{1}{n}\sum_{i=1}^{n} x_i$.

Cet estimateur étant sans biais, son risque quadratique est égal à sa variance. Or la variance d'une loi de Poisson est, comme son espérance mathématique, égale à son paramètre. On a donc

$$\mathrm{Var}_\theta(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} n.\theta = \frac{\theta}{n}$$

Cette variance tend vers 0 quand la taille n de l'échantillon tend vers l'infini, ce qui assure que l'estimateur de la moyenne empirique est consistant en loi (et en probabilité) ; autrement dit la loi de cet estimateur de θ tend en probabilité vers la vraie valeur du paramètre quand n tend vers l'infini.

3. Test

L'indication fournie dans l'énoncé exprime que les lois de Poisson sont telles que, plus θ est élevé, plus la probabilité de prendre de grandes valeurs (formellement : la probabilité de dépasser une valeur fixée) est élevée. On dit que les lois de Poisson sont *stochastiquement croissantes* en fonction de leur paramètre.

Ceci incite, si on dispose d'un estimateur de θ, à rejeter une hypothèse nulle du type $\theta \leq \theta_0$ (contre l'hypothèse alternative $\theta > \theta_0$) quand l'estimation de θ est strictement supérieure à une valeur frontière c, qui doit être déterminée en fonction de la taille n (ici 15), de la borne supérieure de l'hypothèse nulle θ_0 (ici 1) et du seuil de signification α (ici $0,01$).

Utilisant l'estimateur de moyenne empirique introduit en **2**, nous rejetterons donc l'hypothèse nulle $\theta \leq 1$ si l'observation $\sum_{i=1}^{15} x_i$ dépasse strictement la valeur entière d $(= 15c)$ définie de la manière suivante :
• si θ vaut 1, la probabilité que $\sum_{i=1}^{15} X_i > d$ est inférieure ou égale à α, c'est-à-dire ici $0,01$;
• d est le plus petit entier compatible avec la condition précédente.

Comme, si $\theta = 1$, $\sum_{i=1}^{15} X_i$ suit la loi de Poisson de paramètre 15, la première de ces deux conditions équivaut à $F_{15}(d) \geq 0,99$, où F_λ désigne la fonction de répartition de la loi de Poisson de paramètre λ. On remarque que, si $\theta < 1$, on a a fortiori $F_{15\theta}(d) > 0,99$ (et donc la probabilité de rejet à tort de l'hypothèse nulle strictement inférieure à 0,01), ceci résultant du fait que les lois de Poisson sont stochastiquement strictement croissantes en fonction de leur paramètre.

La lecture de la table de la loi de Poisson de paramètre 15 conduit au résultat : $d = 25$. Comme on a observé $\sum_{i=1}^{n} x_i = 26$, on rejette l'hypothèse nulle $\theta \leq 1$.

Remarque : nous avons ici détaillé la construction du test mais il n'est pas indispensable de déterminer la valeur de d pour s'assurer que l'observation de $\sum_{i=1}^{n} x_i = 26$ conduit au rejet

de l'hypothèse nulle ; il est clair que 26 est dans la région de rejet du fait que la probabilité que $\sum_{i=1}^{n} x_i \geq 26$ (c'est-à-dire $1 - F_{15}(26-1) = 0,0062$) est inférieure ou égale à 0,01.

▲

Exercice I.5.

Non fournie.

▲

Exercice I.6.

a. Exhaustivité

La densité de la loi de n v.a. indépendantes toutes régies par la loi de paramètre (α, β) est donnée par :

$$\begin{aligned} \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad f_{\alpha,\beta}^{(n)}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \mathbf{1}_{[\alpha,+\infty[}(x_i) \beta e^{-\beta(x_i - \alpha)} \\ &= \beta^n \mathbf{1}_{[\alpha,+\infty[}(\inf_{1 \leq i \leq n} x_i) \exp\left(-\beta(\sum_{i=1}^{n} x_i - n\alpha)\right). \end{aligned}$$

On en déduit, en appliquant le théorème de factorisation de Halmos-Savage, que la statistique bidimensionnelle $(\inf_{1 \leq i \leq n} X_i, \sum_{i=1}^{n} X_i)$ est exhaustive.

b. Estimation par maximum de vraisemblance

L'observation $x = (x_1, \ldots, x_n)$ étant fixée, on cherche l'argument sur $\mathbb{R} \times \mathbb{R}^{+*}$ du maximum (s'il existe) de l'application vraisemblance :

$$(\alpha, \beta) \mapsto p_n(x; \alpha, \beta) = f_{\alpha,\beta}^{(n)}(x_1, \ldots, x_n).$$

Notons d'abord que pour tout $\beta > 0$ fixé, l'application $\alpha \mapsto p_n(x; \alpha, \beta)$ peut aussi s'écrire

$$\alpha \mapsto \beta^n \mathbf{1}_{]-\infty, \inf_{1 \leq i \leq n} x_i]}(\alpha) \exp(n\beta\alpha) \exp(-\beta(\sum_{i=1}^{n} x_i)),$$

et admet alors clairement un maximum global en $\inf_{1 \leq i \leq n} x_i$, où elle vaut

$$g(\beta) = \beta^n \exp[-\beta(\sum_{i=1}^{n} (x_i - \inf_{1 \leq i \leq n} x_i))].$$

Pour rechercher le max global de $p_n(x; \alpha, \beta)$ il suffit de maximiser dans un second temps la fonction $g(\beta)$ ou $\ln g(\beta)$. En se placant en dehors du cas (qui est de probabilité nulle quelle que soit la valeur du paramètre) où $\sum_{i=1}^{n} (x_i - \inf_{1 \leq i \leq n} x_i) = 0$ (ce qui signifie que tous les x_i sont égaux), on considère l'application

$$\ln g \ : \beta \mapsto n \ln \beta - \beta \sum_{i=1}^{n} (x_i - \inf_{1 \leq i \leq n} x_i).$$

Un calcul simple montre que cette application atteint son maximum en $\beta = \dfrac{n}{\sum_{i=1}^{n}(x_i - \inf_{1\le i\le n} x_i)}$.
L'estimateur du maximum de vraisemblance du couple (α, β) est donc :

$$(\hat{\alpha}_n, \hat{\beta}_n) = (\inf_{1\le i\le n} X_i, \frac{n}{\sum_{i=1}^{n}(X_i - \inf_{1\le i\le n} X_i)}).$$

c. Estimateurs de l'espérance et de la variance

On rappelle que l'espérance et la variance de la loi exponentielle de paramètre β valent respectivement $\frac{1}{\beta}$ et $\frac{1}{\beta^2}$. Donc la loi exponentielle décalée de paramètre (α, β) a respectivement pour espérance et variance les quantités $\frac{1}{\beta} + \alpha$ et $\frac{1}{\beta^2}$.

Ceci suggère de considérer les statistiques

- $\frac{\sum_{i=1}^{n}(X_i - \inf_{1\le i\le n} X_i)}{n} + \inf_{1\le i\le n} X_i = \frac{\sum_{i=1}^{n} X_i}{n}$ pour estimer $\frac{1}{\beta} + \alpha$, et
- $\left(\frac{\sum_{i=1}^{n}(X_i - \inf_{1\le i\le n} X_i)}{n}\right)^2$ pour estimer $\frac{1}{\beta^2}$.

On retrouve pour l'espérance l'estimateur usuel (moyenne empirique) ; par contre l'estimateur de la variance diffère de celui utilisé en première question.

d. Cette question étant libre, nous nous contentons ici de critiquer les estimateurs de maximum de vraisemblance obtenus à la question b .

Comme, presque sûrement, tout X_i est strictement supérieur à α, il en est de même de $\inf_{1\le i\le n} X_i$. Il est donc évident que $\inf_{1\le i\le n} X_i$ est un estimateur biaisé (supérieurement) de α.

Essayons d'en déduire un estimateur non biaisé. Pour cela calculons l'espérance mathématique $\mathbb{E}_{\alpha,\beta}(\inf_{1\le i\le n} X_i)$.

Il est aisé de voir que si $(Y_1, \ldots, Y_n)$ sont des v.a. indépendantes de même loi exponentielle de paramètre 1, alors $\inf_{1\le i\le n} Y_i$ suit la loi exponentielle de paramètre n. En effet :

$$\forall t \ge 0, \quad P[\inf_{1\le i\le n} Y_i \ge t] = P[\forall i = 1, \ldots, n; \quad Y_i \ge t] = (e^{-t})^n.$$

Donc ici $\inf_{1\le i\le n} Y_i$ est de loi exponentielle décalée de paramètre $(\alpha, \beta n)$; il en résulte que $\mathbb{E}_{\alpha,\beta}(\inf_{1\le i\le n} X_i) = \alpha + \frac{1}{n\beta}$. Le biais de l'estimation de α vaut $\frac{1}{n\beta}$ et n'est donc pas connu de nous. Nous n'avons donc pas obtenu une technique nous permettant de modifier l'estimateur proposé pour le débiaiser.

▲

Exercice I.7.

Non fournie ▲

Chapitre II

Modèle linéaire gaussien

II.1 Énoncés

Exercice II.1.
On s'interroge sur la comparaison des tailles moyennes des garçons et des filles de 6 ans dans une population ; pour cela on a pris comme échantillon, jugé représentatif de cette tranche d'âge, une classe d'école primaire (niveau CP en France), et on a observé :

- 16 garçons : moyenne 126,5 cm, écart-type 12,9 cm
- 15 filles : moyenne 136,9 cm, écart-type 11,9 cm.

On admet que la distribution des tailles dans chacune des sous-populations (garçons, filles) suit une loi gaussienne.

1. Donner des intervalles de confiance pour les tailles moyennes des garcons et des filles.
2. Donner un intervalle de confiance pour l'écart type de la taille des garçons. Même question pour les filles.
3. Les écarts-types observés permettent-ils de déduire que les variances des deux populations sont différentes ?
4. Sur la base de la réponse à la question précédente, on suppose que la variance est la même dans les deux populations. Par ailleurs, au vu de cet échantillon, un observateur avance l'opinion : *dans la population, la taille moyenne des filles dépasse de plus de 2 cm celle des garçons.*
   Les données confirment-elles significativement, au niveau $\alpha = 0.05$, cette opinion ? (autrement dit quelle est la conclusion, au niveau $\alpha = 0.05$, du test de l'hypothèse nulle : *dans la population, la taille moyenne des filles dépasse de moins de 2 cm celle des garçons* ?).

△

Exercice II.2.
On souhaite tester, pour une chaîne de magasins, les politiques de publicité suivantes :

A : aucune publicité
B : tracts distribués dans le voisinage
C : tracts distribués et annonces dans les journaux.

On sélectionne 18 magasins divisés au hasard en 3 groupes de 6, et chaque groupe applique l'une des politiques de publicité. On enregistre ensuite les ventes cumulées sur un mois pour chaque magasin, et l'on obtient les moyennes et écart-types empiriques suivants (en milliers de francs) :

	A	B	C
$\bar{X}$	130.17	139.5	169.17
S	8.57	14.71	18.23

où, par exemple, pour le groupe A d'effectif n_A,

$$\bar{X}_A = \frac{1}{n_A}\sum_{j=1}^{n_A} X_{A,j}, \quad S_A = \sqrt{\frac{1}{n_A - 1}\sum_{j=1}^{n_A}(X_{A,j} - \bar{X}_A)^2}.$$

On suppose que les observations pour chaque groupe sont gaussiennes, de moyennes μ_A, μ_B, μ_C et de même variance σ^2.

1. Donner l'estimateur de σ^2 pour ce modèle. Proposer un test de niveau 5% pour l'hypothèse nulle "il n'existe aucune différence entre les politiques de publicité".
2. Tester l'hypothèse "$\mu_A = \mu_C$" contre "$\mu_A \neq \mu_C$" au niveau 5%. Evaluer approximativement la p-valeur.

△

Exercice II.3.
Avec les données ci-dessous, on considère les modèles de régression suivants :

$$\begin{aligned} Y_k &= \beta + \gamma \log(x_k) + \varepsilon_k \\ Y_k &= \beta + \gamma x_k + \varepsilon_k, \end{aligned}$$

où les ε_k sont des v.a. gaussiennes indépendantes et centrées de variance σ^2.

x	1	2	3	4	5	6	7	8
Y	0.39	1.06	0.89	1.15	1.56	1.77	0.94	0.98
x	9	10	11	12	13	14	15	16
Y	1.9	1.59	1.26	1.68	1.25	1.8	1.77	1.72

1. Dans chacun de ces modèles, proposer une estimation sans biais pour σ^2, γ et β.
2. Effectuer un test de "$\gamma = 0$" contre "$\gamma \neq 0$" pour ces modèles. On donnera à chaque fois la p-valeur.
3. Discuter de la valeur respective des deux modèles : Lequel choisir ?

△

Exercice II.4.
(Cet exercice est extrait de la session d'examen de septembre 2002 du module de Statistique et Analyse de données de l'ENPC)

Un industriel fait appel à un statisticien pour le problème suivant : une même fabrication s'effectue sur quatre machines différentes ; un indicateur numérique de la qualité de la production peut être observé sur chaque pièce produite ; l'industriel désire savoir s'il y a un "effet machine" sur la qualité.

L'industriel et la statisticien se sont mis d'accord sur le protocole expérimental et le modèle suivants : pour chaque machine i (où $1 \leq i \leq 4$) on effectuera n_i observations ; les variables aléatoires correspondantes sont notées $X_{i,j}$ (où $1 \leq j \leq n_i$) ; elles sont indépendantes ; la loi de $X_{i,j}$ est la loi normale (autrement dit gaussienne) de moyenne (inconnue) μ_i et de variance (inconnue mais commune pour les 4 machines) σ^2. On testera, au niveau $0,05$, l'hypothèse nulle $[\mu_1 = \mu_2 = \mu_3 = \mu_4]$. Les effectifs n_i n'ont pu être fixés à l'avance, car ils dépendent des conditions de production.

Se souvenant de son cours de statistique en école d'ingénieurs, l'industriel veut faciliter le travail du statisticien en ne l'encombrant pas avec les observations brutes ; après l'expérimentation, il calcule donc lui-même et communique seulement au statisticien la *variabilité totale* de l'échantillon, qui vaut 3,42 et sa *variabilité intraclasses*, qui vaut 1,14.

1. Quelle donnée manque au statisticien pour effectuer le test ?
2. Le statisticien répond à l'industriel en lui indiquant, en fonction de cette donnée manquante (dont l'industriel, lui, dispose), ce qu'est la conclusion du test. Donnez cette réponse (voir la table de la loi de Fisher).

△

Exercice II.5.
La durée d'une maladie semble liée au nombre de bactéries dans l'organisme et à la température du patient lors de son admission à l'hôpital. On détermine pour $n = 10$ malades leur décompte en milliers de bactéries, Φ^1, et leur température Φ^2, et on observe la durée Y de persistance des symptômes de la maladie en jours :

Φ^1	Φ^2	Y
8	37.6	29
7	39.2	29
4	38.5	19
6	37.4	23
9	38.1	32
7	39.1	28
8	39.0	30
3	37.8	18
8	38.2	30
7	39.1	31

1. On propose tout d'abord un modèle (noté vectoriellement)
$$Y = \alpha \mathbf{1}_n + \beta \Phi^1 + \varepsilon,$$
où $\mathbf{1}_n$ est un vecteur de 1 de taille n, et ε est un n-échantillon de $\mathcal{N}(0, \sigma^2)$. Donner des estimateurs sans biais de α, β, σ^2. Proposez un test de la pertinence de ce modèle au niveau 5% (autrement dit de la significativité du régresseur) ; qu'en concluez-vous ?
2. On propose ensuite le modèle
$$Y = \gamma \mathbf{1}_n + \beta_1 \Phi^1 + \beta_2 \Phi^2 + \varepsilon.$$
Donner des estimateur sans biais de $\gamma, \beta_1, \beta_2, \sigma^2$. Tester, au niveau 5%, l'hypothèse "$\beta_2 = 0$" contre "$\beta_2 \neq 0$" *(attention, ce modèle est assez lourd à traiter numériquement ; il peut être plus pratique d'effectuer les calculs sous* `Scilab`*).*

3. Quel modèle conseillez-vous ?

Pour faciliter les calculs numériques, on donne la matrice des sommes de produits croisés ; on a par exemple $\sum_{j=1}^{10} \Phi_j^2 \Phi_j^1 = 2574.9$.

	$\mathbf{1}_{10}$	Φ^1	Φ^2	Y
Φ^1	67	481		
Φ^2	384	2574.9	14749.72	
Y	269	1884	10341.4	7465

△

Exercice II.6.
Une ville veut mettre en place un réseau d'alerte à la pollution par l'ozone. Elle met en concurrence 3 appareils de détection et leur fait prendre à chacun 20 mesures, dans des conditions identiques de faible pollution. Voici les résultats obtenus, en micro-grammes d'Ozone par mètre cube d'air (moyennés sur une heure).

Appareil 1	23,5	38,7	31,5	26,9	42,0	40,5	29,6	22,2	45,3	42,4
	22,3	36,9	28,2	41,1	36,4	45,5	41,6	52,9	41,0	27,7
Appareil 2	22,1	36,9	30,1	25,3	40,2	39,0	27,8	21,0	43,4	40,4
	21,2	35,0	26,6	39,7	35,0	43,5	40,0	50,8	39,2	26,2
Appareil 3	10,8	43,7	28,8	18,1	51,1	48,4	23,5	08,4	58,2	51,6
	09,0	39,6	20,8	50,1	39,6	58,5	50,6	74,6	49,1	20,1

1. Compte tenu des avis des experts sur la variabilité naturelle des teneurs en ozone, le cahier des charges de l'appel d'offres exigeait : *en situation de faible pollution (inférieure à 80), la précision de l'appareil doit assurer un écart-type de la loi des mesures inférieur ou égal à 10 micro-grammes d'ozone par mètre cube d'air* .

a. Pour chacun des appareils, testez, au seuil 0,05, l'hypothèse que l'appareil satisfait à cette clause du cahier des charges. On admettra pour cela que, pour chaque appareil, les 20 mesures suivent une même loi normale et sont indépendantes.

Les étudiants qui le désirent pourront utiliser les résultats intermédiaires suivants (où $x_{i,j}$, où $1 \leq i \leq 3$ et $1 \leq j \leq 20$) désigne la mesure numéro j faite avec l'appareil numéro i :

$$\sum_{j=1}^{20} x_{1,j} = 716,2 \qquad \sum_{j=1}^{20} x_{2,j} = 683,4 \qquad \sum_{j=1}^{20} x_{3,j} = 754,6$$

$$\sum_{j=1}^{20} x_{1,j}^2 = 27104,72 \qquad \sum_{j=1}^{20} x_{2,j}^2 = 24741,78 \qquad \sum_{j=1}^{20} x_{3,j}^2 = 35332,32$$

b. Le choix d'une autre valeur du seuil, plus faible (test plus sévère) ou plus forte (test moins sévère) changerait-il certaines des conclusions retenues à la sous-question précédente ? Si oui, pouvez-vous donner des indications sur les valeurs du seuil qui conduiraient à de telles modifications ?

2. Seuls les appareils 1 et 2 restent en concurrence. L'étude menée en question **1** justifie de considérer que leurs lois (toujours supposées normales) ont même variance. On veut savoir s'il y a une différence significative entre les résultats qu'ils fournissent. Aucune indication

supplémentaire ne nous ayant été fournie à ce stade sur les conditions de recueil des mesures, cela signifie qu'on veut tester l'hypothèse $\mu_1 = \mu_2$, où μ_i (avec i égal à 1 ou à 2) désigne l'espérance mathématique de la loi des observations faites avec l'appareil i. Sur la base du tableau de mesures fourni précédemment, effectuez ce test au seuil 0,05.

3. On nous indique que les mesures ont été effectuées pendant 20 jours consécutifs, à la même heure (de 9h. à 10h. du matin), les 3 appareils ayant été posés côte à côte. L'indice j désignant alors le jour, il y a lieu de considérer qu'il s'agit de mesures appariées, la "vraie pollution" pouvant varier de jour en jour. Reprendre avec cette indication nouvelle le test, au seuil 0,05, d'identité de comportement des appareils 1 et 2, autrement dit tester que les variables aléatoires, toutes de loi gaussienne et de même variance , $(X_{1,j} - X_{2,j})$, ont leurs espérances mathématiques nulles.

Les étudiants qui le désirent pourront utiliser le résultat intermédiaire suivant :

$$\sum_{j=1}^{20}(x_{1,j} - x_{2,j})^2 = 55,26$$

4. Les précisions des appareils 1 et 2 étant analogues, on envisage, par application du "principe de précaution", de passer plutôt le marché avec le fabricant de l'appareil 1, qui est systématiquement un peu plus pessimiste que l'appareil 2. Mais on veut aussi tester cet appareil en situation de pic de pollution (alors que les mesures précédentes étaient faites en période de faible pollution). On veut aussi tester la capacité du fabricant de fournir en nombre des appareils de même qualité.

On indique que la valeur de 180 micro-grammes d'Ozone par mètre cube d'air est utilisée dans la région Ile-de-France pour déclencher les informations au public et celle de 360 est utilisée pour déclencher les actions restrictives telles que des interdictions de circulation.

On demande donc au fabricant de fournir 10 appareils ; on choisit un jour et une heure où d'autres appareils, extrêmement fiables mais plus chers que ceux dont la ville veut se doter en grand nombre, ont annoncé une pollution égale à 340 ; voici les résultats fournis alors par les 10 appareils de type 1 testés ; on les notera $x_{4,j}$ (avec $1 \leq j \leq 10$) et on les considèrera comme indépendants et issus d'une même loi normale d'espérance mathématique μ :

330,5	345,8	336,4	351,0	345,8	355,2	351,3	363,3	350,5	336,0

a. A quels seuils (parmi ceux que vous pouvez lire sur les tables fournies) ces résultats conduisent-ils à l'acceptation de l'hypothèse $\mu = 340$ (qui exprime que ce type d'appareil détecte bien le pic de pollution à sa vraie valeur) ?

Les étudiants qui le désirent pourront utiliser les résultats intermédiaires suivants :

$$\sum_{j=1}^{10} x_{4,j} = 3465,8 \qquad \sum_{j=1}^{10} x_{4,j}^2 = 1202063,36$$

b. Estimez la probabilité qu'un appareil de type 1 conduise, si la vraie pollution est 340, à une "fausse alarme de mesures restrictives", c'est-à-dire affiche un résultat supérieur à 360.

△

Exercice II.7.
Les données de taux d'équipement des ménages pour un certain produit sont reproduites dans le tableau suivant :

i (année)	1	2	3	4	5	6	7	8	9	10
Y_i (en %)	2.9	4.4	6.0	8.4	11.8	14.6	18.3	24.1	30.8	40.0

On souhaite ajuster les données sur une courbe logistique i.e. de la forme :

$$y(t) = \frac{1}{1 + b\, e^{-at}}$$

On note :

$$X_i = \ln\left(\frac{1 - Y_i}{Y_i}\right)$$

On choisit un modèle de régression de la forme :

$$X_i = \beta + \alpha\, i + \varepsilon_i \text{ pour i=1,2,...,n,}$$

On note $\sigma^2 = E(\varepsilon_i^2)$

1. Vérifier que ce modèle permet d'ajuster $(Y_1, \ldots, Y_n)$ sur une courbe logistique de paramètre a et b.

2. Calculer les estimations de α et β et le coefficient de détermination R^2 de la régression.

3. Calculer une estimation sans biais de σ^2 et des intervalles de confiance pour β et α.

4. En déduire des estimations et des intervalles de confiance pour a et b.

5. Effectuer un test de $(a = 0)$ contre $(a \neq 0)$ pour ce modèle.

△

II.2 Corrections

Exercice II.1.

1. C'est l'application directe de l'IC pour la moyenne d'un échantillon gaussien dont la variance inconnue est estimée par la variance empirique (version sans biais). La loi utilisée est donc celle de Student (Chapitre III, § 2.3), et l'IC est $\bar{X} \pm t_{n-1,1-\alpha/2} S/\sqrt{n}$. Le niveau n'étant pas précisé, on propose de prendre 95% de niveau de confiance, soit $\alpha = 5\%$.
   - Pour les garçons, on a observé $(X_1, \dots, X_{n_G})$ i.i.d. de $\mathcal{N}(\mu_G, \sigma_G^2)$. la table donne $t_{15,0.975} = 2.13$, et on trouve $\mu_G \in [119.63; 133.37]$.
   - Pour les filles, on a observé $(Y_1, \dots, Y_{n_F})$ i.i.d. de $\mathcal{N}(\mu_F, \sigma_F^2)$. La table donne $t_{14,0.975} = 2.15$ et on trouve $\mu_F \in [130.52; 143.28]$.
2. On utilise le fait que dans le cas gaussien, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, donc
$$\mathbb{P}\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha,$$
d'où l'IC de niveau $(1-\alpha)$ pour la variance
$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right].$$
Si on prend $\alpha = 0.05$ (IC de niveau de confiance 95%), on trouve pour les garçons $\sigma_G \in [9.53; 19.97]$ et pour les filles $\sigma_F \in [8.71; 18.77]$.
3. On souhaite tester H_0 : "$\sigma_G^2 = \sigma_F^2$" contre H_1 : "$\sigma_G^2 \neq \sigma_F^2$" (test bilatéral). Ce test n'est pas donné dans le chapitre III. On sait les estimateurs appropriés pour σ_G^2 et σ_F^2 sont S_G^2 et S_F^2 dont les valeurs numériques des racines sont données. Les lois de ces estimateurs sont accessibles via la normalisation $(n_G-1)S_G^2/\sigma_G^2 \sim \chi^2(n_G-1)$ (et idem pour les filles). Elles dépendent chacune de la vraie valeur de la variance, mais sous $H_0 : \sigma_G^2 = \sigma_F^2 = \sigma^2$ inconnu, donc le rapport des deux χ^2 (indépendants) normalisés élimine le paramètre inconnu et suit une loi de Fisher. On choisit comme numérateur par exemple l'estimateur qui a donné la plus grande valeur :
$$\frac{S_G^2}{S_F^2} \sim F(n_G - 1, n_F - 1) \quad \text{sous } H_0,$$
et on rejette H_0 si $\{S_G^2/S_F^2 > F_{n_G-1,n_F-1,1-\alpha/2}\}$ ou si $\{S_G^2/S_F^2 < F_{n_G-1,n_F-1,\alpha/2}\}$. On trouve $S_G^2/S_F^2 = 1.18$ pour $F_{15,14,0.975} = 2.95$, donc on ne rejette pas H_0 (l'autre quantile vaut 0.35, cf la loi ci-dessous). La p-valeur est $2\mathbb{P}(F > 1.18) = 0.76$, donc on est conduit à accepter l'égalité des variances.
4. Il s'agit de tester H_0 : "$\mu_F - \mu_G \leq 2$" contre H_1 : "$\mu_F - \mu_G > 2$". Avec l'hypothèse d'égalité des variances que l'on vient d'admettre, c'est le test donné au chapitre III, § 2.5 avec ici un décalage de 2 (au lieu de 0 dans le cours). La statistique de test est donc
$$T = \frac{(\bar{X}_F - \bar{X}_G - 2)\sqrt{n_F + n_G - 2}}{V\sqrt{1/n_F + 1/n_G}} \sim t(n_F + n_G - 2) \quad \text{sous } H_0,$$

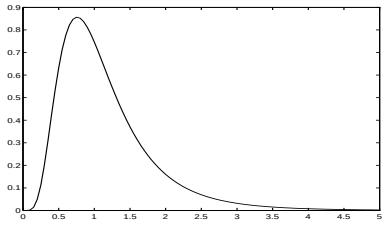


FIG. II.1 – Densité de la loi sous H_0, $F(15, 14)$

où $V^2 = (n_F - 1)S_F^2 + (n_G - 1)S_G^2$. On rejette si $\{T > t_{n_F+n_G-2,1-\alpha} = 1.70\}$. On trouve $T = 1.88$ donc on rejette H_0 au niveau 5%. Remarquons que la p-valeur vaut ici 0.0351, donc le même test au niveau 1% ne rejetait pas H_0.

▲

Exercice II.2.

C'est le modèle d'ANOVA à 1 facteur qui est ici la politique de publicité. C'est donc une application directe du cours. les moyennes et écarts-types empiriques empiriques par groupe, ainsi que la connaissance des effectifs des groupes ($n_A = n_B = n_C = 6$) suffisent à faire les calculs.

1. L'estimateur de σ^2 pour le modèle linéaire est

$$\hat{\sigma}^2 = \frac{||X - X_E||^2}{n - 3}, \quad \text{avec} \quad ||X - X_E||^2 = \sum_{i=1}^{3}(n_i - 1)S_i^2 = 3110.8,$$

d'où $\hat{\sigma}^2 = MSE = 207.38$. Le test de "non effet du facteur" utilise la statistique de Fisher

$$F = \frac{||X_E - X_H||^2/3 - 1}{||X - X_E||^2/n - 3},$$

avec $||X_E - X_H||^2 = \sum_{i=1}^{3} n_i(X_{i\cdot} - X_{\cdot\cdot})^2$, où $X_{1\cdot}$ dénote par exemple la moyenne du groupe A notée $\bar{X}_A$ dans le texte. On calcule la moyenne générale à partir des 3 moyennes par groupes : $X_{\cdot\cdot} = (\sum_{i=1}^{3} 6X_{i\cdot})/18$. Cela donne $||X_E - X_H||^2 = 4976.7$, d'où $F = 12$ et $F_{2,15,0.05} = 3.68$. On rejette H_0 : le facteur "politique de publicité" est significatif.

2. Test de "$\mu_A = \mu_C$" : c'est le test de Student de comparaison de moyennes de 2 populations. La différence avec le cours est que sous l'hypothèse d'homoscédasticité, on estime la variance sur les observations des trois groupes, donc par la MSE du modèle linéaire, plutôt que sur les deux groupes concernés par le test. La statistique de test est

$$T = \frac{(X_{A\cdot} - X_{C\cdot})\sqrt{n-3}}{||X - X_E||\sqrt{1/n_A + 1/n_C}} \sim t(n-3) \text{ sous } H_0.$$

On rejette H_0 si $\{T < -t_{n-3,\alpha}\}$. On trouve T = -4.691, et $-t_{15,0.05} = -1.753$ donc rejet de H_0. On peut approcher la p-valeur avec une table de Student usuelle qui donne par exemple $-t_{15,0.0005} = -4.07$, donc la p-valeur est $< 5.\ 10^{-4}$.

▲

Exercice II.3.

Il s'agit de comparer deux modèles de régression possibles, l'un sur le régresseur $X = (x_1, \ldots, x_n)$, l'autre sur le régresseur $Z = (\log(x_1), \ldots, \log(x_n))$. C'est l'observation du nuage de points du modèle $Y = \beta + \gamma X$ qui a suggéré l'essai de l'autre modèle.

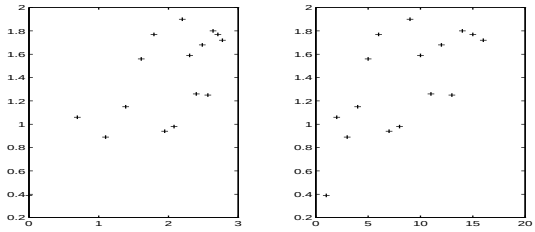


FIG. II.2 – Nuages du modèle $Y = \beta + \gamma \log(X)$ (gauche), et $Y = \beta + \gamma X$ (droite)

(1) Il s'agit de modèles de régression simples, cas traité complètement dans le cours, § 4. Les matrices des régresseurs sont $M1 = [\mathbf{1}_n \; Z]$ (modèle logarithmique) et $M2 = [\mathbf{1}_n \; X]$. Les estimateurs de $(\beta, \gamma, \sigma^2)$ pour chacun des modèles sont

$$\begin{array}{rcl} M1 & : & \beta_1 = 0.58, \quad \gamma_1 = 0.41, \quad \sigma_1^2 = 0.09 \\ M2 & : & \beta_2 = 0.84, \quad \gamma_2 = 0.06, \quad \sigma_2^2 = 0.11 \end{array}$$

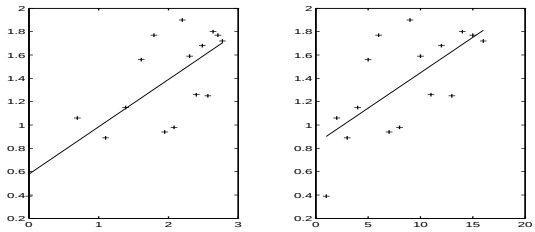


FIG. II.3 – Nuages et droites de régression pour le modèle $M1$ (gauche), et $M2$ (droite)

(2) Il s'agit du test de H_0 : "le régresseur n'a pas d'effet", test de Fisher de statistique

$$F = \frac{||Y_E - \bar{Y}\mathbf{1}_n||^2}{||Y - Y_E||/(n-2)} \sim F(1, n-2) \quad \text{sous } H_0.$$

Le calcul donne les Fisher et p-valeurs suivantes :

$$\begin{array}{rcl} M1 & : & F_1 = 17.22, \quad p_1 = 0.001 \\ M2 & : & F_2 = 11.42, \quad p_2 = 0.005 \end{array}$$

Dans les deux cas, on rejette clairement H_0 pour tout niveau classique : il faudrait en M2 un niveau plus petit que 0.005 (non pratiqué sauf exception) pour ne pas rejeter l'hypothèse nulle.

(3) Comme les deux modèles sont significatifs, on peut les comparer d'une part avec le critère du plus grand Fisher (i.e. de la plus petite p-valeur), d'autre part avec le critère du coefficient de détermination R^2. Ici pour le modèle logarithmique, $R_1^2 = 55.2\%$ de variation expliquée, et pour le modèle $M2$, $R_2^2 = 44.9\%$. Pour les deux critères, le modèle logarithmique est donc préférable.

▲

Exercice II.4.

(1) Il manque évidemment le nombre total d'observations $n = n_1 + n_2 + n_3 + n_4$). On peut remarquer que l'on doit avoir pour appliquer les principes de l'analyse de la variance $n > 4$

(2) La table d'analyse de la variance peut être constituée de la manière suivante en fonction du paramètre n.

Variabilité	SS	DF	MS	Fisher
Interclasse	$1,14$	3	$0,38$	$f = 0,17(n-4)$
Intraclasse	$2,28$	$n-4$	$\frac{2,28}{n-4}$	
Totale	165.82	$n-1$		

Sous l'hypothèse H_0 d'égalité des moyennes la statistique de Fisher suit une loi de Fisher $\mathcal{F}(3, n-4)$. Pour $\alpha = 0,05$, on rejette l'hypothèse H_0 si $f > \mathcal{F}_{3;n-4;0,05}$, soit $n > \nu(n)$ avec $\nu(n) = 4 + \dfrac{\mathcal{F}_{3;n-4;0,05}}{0,17}$.

D'après la table des quantiles de la loi de Fisher- Snedecor, on peut remarquer que si $n = 5$, on a $\mathcal{F}_{3;1;0,05} = 215,71$ et $\nu(5) = 1273$, donc on accepte H_0 alors que si $n \longrightarrow +\infty$, $\nu(+\infty)$ a une valeur finie donc on rejette H_0.
On peut traduire cela par le fait que si n est petit on a pas assez d'observations, donc d'information pour rejeter l'hypothèse H_0 alors que si $n \longrightarrow +\infty$ on a au contraire l'information complète sur les paramètres et on accepte H_0 uniquement si les 4 moyennes empiriques sont égales.

De plus $\mathcal{F}_{3;n-4;0,05}$ décroit avec n.

On en déduit qu'il existe une valeur critique n_c telle que pour $n < n_c$ on accepte H_0 et pour $n \geq n_c$ on rejette H_0.

En calculant $\nu(n)$ pour quelques valeurs de n, on peut évaluer aisément n_c. En particulier on constate que $\nu(22) = 22.59$ et $\nu(23) = 22,41$, donc $n_c = 23$.

▲

Exercice II.5. Non fournie. ▲

Exercice II.6.

1.a. Les estimations sans biais de l'espérance mathématique et de la variance pour chacun des trois appareils sont les suivantes :

$$m_1 = 35,81 \qquad m_2 = 34,17 \qquad m_3 = 37,73$$

$$s_1^2 = 76,72 \qquad s_2^2 = 73,16 \qquad s_3^2 = 361,12$$

Afin de tester, pour chaque i ($1 \leq i \leq 3$), l'hypothèse $\sigma_i^2 \leq 100$, on calcule $19\frac{s_i^2}{100}$ qu'on compare au quantile supérieur d'ordre $0,05$ de la loi du χ^2 à 19 degrés de liberté, qui vaut 30,144. Il vient :

$$19\frac{s_1^2}{100} = 14,58 < 30,144 \quad , \quad 19\frac{s_2^2}{100} = 13,90 < 30,144 \quad , \quad 19\frac{s_3^2}{100} = 68,61 > 30,144$$

Donc, au seuil 0,05, on rejette l'hypothèse que le cahier des charges est respecté pour l'appareil 3, mais on l'accepte pour les appareils 1 et 2 (et dans les trois cas ces conclusions sont fort nettes).

1.b. Si on diminue le seuil α (test plus sévère), on ne peut que confirmer les décisions de non-rejet pour les appareils 1 et 2 ; pour l'appareil 3 on constate en regardant la ligne 19 dans la table des fractiles des lois du χ^2 que toutes les valeurs qui s'y trouvent sont inférieures à 68,61 ; donc, quelque sévère que soit le seuil choisi parmi ceux fournis, on continue à rejeter l'hypothèse.

Si on augmente le seuil α (test moins sévère), on ne peut que confirmer la décision de rejet pour l'appareil 3 ; pour les appareils 1 et 2, on constate en regardant la ligne 19 dans la table des fractiles des lois du χ^2 que pour l'autre valeur de seuil disponible (0,1) la valeur du quantile est 27,20, supérieure à 14,58 (et donc a fortori à 13,90) ; donc, pour un test de niveau 10%, on continue à ne pas rejeter l'hypothèse.

2. Pour effectuer le test de comparaison des espérances mathématiques pour des lois normales de même variance, on calcule $m_1 - m_2 = 1,64$ puis $\hat{s}^2 = \frac{1}{2}(s_1^2 + s_2^2)$ (car les deux sous-échantillons ont les mêmes effectifs $n_1 = n_2 = 20$), d'où $\hat{s}^2 = \frac{76,72+73,16}{2} = 74,94$, et enfin

$$t = \frac{m_1 - m_2}{\hat{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \sqrt{10}\frac{1,64}{\sqrt{74,94}} = 0,599$$

Pour tester l'hypothèse $\mu_1 = \mu_2$ au seuil $\alpha = 0,05$, on compare t au quantile supérieur d'ordre $0,025$ de la loi de Student à 38 degrés de liberté, qui est de l'ordre de 2,02 ; en effet la valeur 38 pour le nombre de degrés de liberté ne figure pas dans la table, mais on a les valeurs pour 30 (2,042) et pour 40 (2,021). On constate que $0,599 < 2,02$ (et ce très largement) et donc on ne rejette pas, au seuil 0,05, l'hypothèse $\mu_1 = \mu_2$.

3. Pour effectuer le test de comparaison des espérances mathématiques pour des échantillons appariés de lois normales, on utilise les $y_j = x_{1,j} - x_{2,j}$, dont voici la liste (*calcul pouvant être évité en utilisant le résulat intermédiaire figurant dans l'énoncé*)

1,4	1,8	1,4	1,6	1,8	1,5	1,8	1,2	1,9	2,0
1,1	1,9	1,6	1,4	1,4	2,0	1,6	2,1	1,8	1,5

On calcule $m = \frac{1}{20}\sum_{j=1}^{20} y_j = m_1 - m_2 = 1,64$ puis $\sum_{j=1}^{20} y_j^2 = 55,26$, d'où l'estimation sans biais de la variance de la loi commune des y_j, $\hat{s}^2 = 0,0773$ (**attention :** ce n'est pas le même $\hat{s}^2$ qu'en question **2**).

On en déduit $z = \sqrt{20}\frac{m}{\hat{s}} = 26,38$ que l'on compare au quantile supérieur d'ordre 0,025 de la loi de Student à 19 degrés de liberté, qui vaut 2,093. On constate que $26,38 > 2,093$ et donc ici on rejette très nettement, au seuil 0,05, l'hypothèse d'identité de loi des appareils 1 et 2. On pouvait s'y attendre car dans les 20 couples de valeurs, celle enregistrée par l'appareil 1 est toujours supérieure à celle enregistrée par l'appareil 2 (la différence variant entre 1,1 et 2).

4.a. Pour tester l'hypothèse que l'espérance mathématique de la loi des enregistrements lors du pic de pollution vaut 340, on calcule les estimateurs sans biais de l'espérance mathématique et de la variance, c'est-à-dire $m = 346,58$ et $s^2 = 98,49$.

On en déduit $\sqrt{n}\frac{|m-340|}{s} = \sqrt{10}\frac{|346,58-340|}{9,92} = 2,10$.

Considérons les quantiles supérieurs de la loi de Student à 9 degrés de liberté. La valeur calculée 2,10 se situe entre celui d'ordre 0,05 (qui vaut 1,833) et celui d'ordre 0,025 (qui vaut 2,262). Donc, parmi les valeurs du seuil α classiques, celles supérieures ou égales à 0,10 conduisent au rejet de l'hypothèse $\mu = 340$ et celles inférieures ou égales à 0,05 conduisent au non-rejet de cette hypothèse.

4.b. On estime la loi des observations par la loi normale d'espérance mathématique 346,58 et de variance 98,49 (donc d'écart-type 9,92). Φ désignant la fonction de répartition de la loi normale centrée réduite, la probabilité de dépasser 360 est donc estimée par :

$$1 - \Phi(\frac{360 - 346,58}{9,92}) = 1 - \Phi(1,35) = 1 - 0,9115 = 0,0885$$

La probabilité de "fausse alarme de mesures restrictives" est donc proche de 9%.

▲

Exercice II.7. 1. On a :

$$h(t) = \ln(\frac{1 - y(t)}{y(t)}) = \ln b - at$$

Donc la courbe logistique est transformée par l'application h, en une droite : l'ajustement à la courbe logistique devient une régression linéaire sur les données transformés par h avec $\alpha = -a$ et $\beta = \ln b$.

2. Les estimateurs de α et β, ainsi que les variances de ces 2 estimateurs sont explicités dans le polycopié en **4.2**. On obtient : $\hat{\alpha} = -0.332$; $\hat{\beta} = 3.763$. Le coefficient de détermination vaut : $R^2 = 0.9973$

3. L'estimation sans biais de la variance vaut : $\hat{\sigma}^2 = 3.1\ 10^{-3}$
On en déduit : $V(\hat{\alpha}) = 3.79\ 10^{-5}$ et $V(\hat{\beta}) = 1.5\ 10^{-3}$.
Au niveau 95% , on obtient les intervalles de confiances :

$$I(\alpha) = [-0.347; -0.317] \text{ et } I(\beta) = [3.67; 3.86]$$

4. Les estimations de a et b valent : $\hat{a} = -\hat{\alpha} = 0.332$ et $\hat{b} = e^{\hat{\beta}} = 43.1$
On obtient les intervalles de confiance suivants :

$$I(a) = -I(\alpha) = [0.317; 0.347] \text{ et } I(b) = [e^{3.67}; e^{3.86}] = [23.80; 39.26]$$

5. Tester $(a = 0)$ contre $(a \neq 0)$ revient à tester $(\alpha = 0)$ contre $(\alpha \neq 0)$, dans la régression linéaire.
On a la table suivante :

Fisher	p-valeur
2906	0.000

On rejette évidemment l'hypothèse (a=0).

▲

Chapitre III

Modèles discrets

III.1 Énoncés

Exercice III.1.
On a croisé 2 types de plantes, différant par deux caractères ; le premier prend les valeurs A et a, le second prend les valeurs B et b. On s'est assuré de l'homogénéité des plantes de la première génération : pour chaque type de plante, chacun des deux phénotypes représente la moitié de l'échantillon sur lequel on effectue les croisements. On s'interroge sur le modèle suivant :

- A est dominant et a est récessif,
- B est dominant et b est récessif.

Par les lois de Mendel ce modèle conduirait, à la seconde génération, pour les 4 phénotypes AB , Ab , aB et ab, à des probabilités égales respectivement à 9/16 , 3/16 , 3/16 et 1/16.

Or, à partir d'un échantillon de 160 plantes, on a observé des effectifs respectifs de 100 , 18 , 24 et 18.

1. Testez, au niveau de signification $\alpha = 0,05$, le modèle envisagé.

2. Que pouvez-vous dire (à l'aide de la table de quantiles de la loi du χ^2 fournie à l'appui de ce cours) sur la p-valeur associée au résultat observé (autrement dit en dessous de quelle valeur pour $\alpha = 0,05$ ce résultat ne conduit pas au rejet du modèle proposé) ?

3. Reprendre la question 1 dans le cas où l'expérience aurait porté sur deux fois moins de plantes, soit 80, et conduit aux effectifs respectifs de 50, 9, 12 et 9 (c'est-à-dire les mêmes proportions que dans l'expérience initiale).

△

Exercice III.2.
On se propose de comparer les réactions produites par deux vaccins B.C.G. désignés par A et B. Un groupe de 348 enfants a été divisé par tirage au sort en deux séries qui ont été vaccinées, l'une par A, l'autre par B. La réaction a été ensuite lue par une personne ignorant le vaccin utilisé. Les résultats figurent dans le tableau suivant :

Vaccin	Réaction légère	Réaction moyenne	Ulcération	Abcès	Total
A	12	156	8	1	177
B	29	135	6	1	171
Total	41	291	14	2	348

On désire tester l'hypothèse selon laquelle les réactions aux deux vaccins sont de même loi.

1. Expliquez pourquoi cette situation relève d'un test du χ^2 d'indépendance.

2. Les effectifs observés permettent-ils d'effectuer le test ? Si non, procédez aux opérations nécessaires sur ces données, puis effectuez le test au niveau de signification $\alpha = 0,05$. Discutez selon le choix d'autres valeurs de α.

△

Exercice III.3.
Nous disposons des résultats d'une enquête réalisée auprès de 200 femmes américaines mariées sur leur activité. Parmi les questions, deux vont nous intéresser pour cet exercice. La première, notée **A**, est la suivante : *Avez-vous une activité professionnelle actuellement ?* alors que la seconde, notée **B**, est : *Avez-vous des enfants de moins de deux ans ?* . L'objectif de cette étude est de savoir si la présence d'enfants très jeunes influe sur le fait d'avoir une activité professionnelle.

La répartition des réponses fournies aux questions **A** et **B** se trouve dans le tableau suivant :

A — B	OUI	NON	Total
OUI	32	103	135
NON	30	35	65
Total	62	138	200

1. Testez (en discutant sur le niveau de signification choisi) l'hypothèse selon laquelle les deux variables sont indépendantes ? Indication : *on utilisera le test du χ^2 d'indépendance (chapitre 3, section 1).*

2. Testez (en discutant sur le niveau de signification choisi) l'hypopthèse selon laquelle, dans la population totale, les proportions de femmes exerçant une activité professionnelle sont égales parmi celles qui ont des enfants de moins de deux ans et celles qui n'en ont pas ? Indication :*on utilisera le test de comparaison des proportions dans deux grands échatillons appariés (chapitre 3, section 3).*

3. On désire modéliser le fait d'avoir une activité professionnelle (la variable à expliquer) par la présence d'enfants de moins de deux ans (la variable candidate à l'explication). Pour cela on choisit d'effectuer une régression logistique *(chapitre 3, section 2).* Mais ici la variable candidate à l'application est qualitative (alors que dans le modèle général de la régression logistique elle est numérique). On doit donc adopter un codage arbitraire pour cette variable, par exemple 0 pour "pas d'enfant de moins de 2 ans" et 1 pour "présence d'au moins un enfant de moins de 2 ans" (ou bien respectivement -1 et 1).

a) Justifiez le choix de la régression logistique en montrant que le codage n'aura aucun effet sur le test statistique.

b) Ecrivez le modèle et sa vraisemblance, avec le codage par 0 et 1 proposé ci-dessus.

c) Estimez les paramètres du modèle.

d) On veut tester l'hypothèse selon laquelle *la présence d'enfants de moins de deux ans n'influerait pas sur le fait d'avoir une activité professionnelle*. Pour répondre à cette question, déroulez les étapes suivantes :

(i) Exprimez l'hypothèse nulle dans le modèle choisi en b) ci-dessus ?
(ii) Calculez la statistique du rapport de vraisemblances.
(iii) Effectuez le test avec un risque de première espèce de 5%.
(iv) Qu'en déduisez-vous ?
(v) Interprétez le modèle.

e) Calculez le coefficient de détermination de Mc Fadden.

4. Comparez les différents résultats obtenus (test d'indépendance, test d'égalité des proportions et test du rapport de vraisemblance du modèle logistique).

△

Exercice III.4.
On désire étudier la répartition des naissances suivant le type du jour de semaine (jours ouvrables ou week-end) et suivant le mode d'accouchement (naturel ou par césarienne). Les données proviennent du "National Vital Statistics Report" et concernent les naissances aux USA en 1997.

Naissances	Naturelles	César.	Total
J.O.	2331536	663540	2995076
W.E.	715085	135493	850578
Total	3046621	799033	3845654

Naissances	Naturelles	César.	Total
J.O.	60.6 %	17.3 %	77.9%
W.E.	18.6 %	3.5 %	22.1%
Total	79.2 %	20.8 %	100.0%

On note $p_{J,N}$ la probabilité qu'un bébé naisse un jour ouvrable et sans césarienne, $p_{W,N}$ la probabilité qu'un bébé naisse un week-end et sans césarienne, $p_{J,C}$ la probabilité qu'un bébé naisse un jour ouvrable et par césarienne, $p_{W,C}$ la probabilité qu'un bébé naisse un week-end et par césarienne.

1. Rappeler l'estimateur du maximum de vraisemblance de

$$p = (p_{J,N}, p_{W,N}, p_{J,C}, p_{W,C})$$

.

2. À l'aide d'un test du χ^2, pouvez-vous accepter ou rejeter l'hypothèse d'indépendance entre le type du jour de naissance (jour ouvrable ou week-end) et le mode d'accouchement (naturel ou césarienne) ?

3. On désire savoir s'il existe une évolution significative dans la répartition des naissances par rapport à 1996. À l'aide d'un test du χ^2, pouvez-vous accepter ou rejeter l'hypothèse $p = p_0$, où p_0 correspond aux données de 1996 ? On donne les valeurs suivantes pour p_0 :

Naissances	Naturelles	Césariennes
J.O.	60.5 %	17.0 %
W.E.	18.9 %	3.6 %

△

Exercice III.5.
On souhaite vérifier la qualité du générateur de nombres aléatoires d'une calculatrice scientifique. Pour cela, on procède à 250 tirages dans l'ensemble $\{0, \ldots, 9\}$ et on obtient les résultats suivants :

x	0	1	2	3	4	5	6	7	8	9
$N(x)$	28	32	23	26	23	31	18	19	19	31

À l'aide du test du χ^2, vérifier si le générateur produit des entiers indépendants et uniformément répartis sur $\{0, \ldots, 9\}$.

△

III.2 Corrections

Exercice III.1. 1. On va procéder à un test du χ^2. Les "effectifs théoriques" sous l'hypothèse à tester (selon les notations du polycopié ce sont les valeurs $n.p_j^0$, où $1 \leq j \leq 4$) sont respectivement 90, 30, 30 et 10, d'où le calcul de la distance du χ^2 :

$$\frac{10^2}{90} + \frac{(-12)^2}{30} + \frac{(-6)^2}{10} + \frac{8^2}{90} = 13,51.$$

Or le quantile d'ordre $1-\alpha$ de la loi du χ^2 à 3 degrés de liberté est ici :

$$\chi^2_{3,0,95} = 7,815.$$

On a $13,51 > 7,815$ donc l'hypothèse nulle est rejetée (on dit que la différence entre la répartion observée et la répartion théorique est "significative" au niveau 0,05).

2. La table fournie avec ce cours nous montre que 13,51 est compris entre les quantiles d'ordres 0,09 et 0,009 de la loi du χ^2 à 3 degrés de liberté. Donc on sait que :
- si $\alpha \geq 0,01$ on est conduit au rejet de l'hypothèse nulle,
- si $\alpha \leq 0,001$ on n'est pas en situation de rejeter l'hypothèse nulle.

3. La conservation de toutes les proportions (théoriques et observées), avec division de l'effectif par 2, conduit à diviser aussi par 2 la valeur calculée de la statistique du χ^2, qui vaut donc maintenant 6,75. Cette valeur est inférieure à 7,815 et cette fois on ne peut rejeter l'hypothèse nulle au niveau 0,05.

▲

Exercice III.2.

1. Reprenons les notations du cours sur le test de χ^2 d'indépendance (**chapitre IV, 2.**). Nous observons ici 348 v.a. i.i.d. $X_i = (Y_i, Z_i)$, où les Y_i sont à valeurs dans un ensemble à 2 éléments (les 2 vaccins) et les Z_i sont à valeurs dans un ensemble à 4 éléments (les 4 réactions). Le paramètre est donc de la forme $\underline{p} = (p_{j,h})_{1 \leq j \leq k, 1 \leq h \leq m}$.

Si on pose pour tout j $(1 \leq j \leq 2)$ $q_j = \sum_{h=1}^m p_{j,h}$ et, pour tout h $(1 \leq h \leq 4)$, $r_h = \sum_{j=1}^k p_{j,h}$, les q_j caractérisent la loi commune des v.a. Y_i et les r_h caractérisent la loi commune des v.a. Z_i ; ces lois sont appelées aussi première et seconde **lois marginales** des X_i.

Considérons les deux hypothèses suivantes :

A : les 2 composantes sont indépendantes, autrement dit : $\forall (j,h)\ \ p_{j,h} = q_j.r_h$
B : la loi, conditionnellement au vaccin, de la réaction est la même pour chacun des deux vaccins, autrement dit : $\forall h\ \ \frac{p_{1,h}}{q_1} = \frac{p_{2,h}}{q_2}$.

Vérifions que ces deux hypothèses sont en fait équivalentes. Il est évident que A implique B. Inversement, B étant satisfaite, notons, pour tout h, s_h la valeur commune de $\frac{p_{1,h}}{q_1}$ et $\frac{p_{2,h}}{q_2}$; il vient alors :

$$r_h = \sum_{j=1}^{2} p_{j,h} = q_1.s_h + q_2.s_h = (q_1+q_2).s_h = s_h$$

et donc on retrouve $p_{j,h} = q_j.r_h$.

2. Les effectifs, dans la colonne "abcès", sont trop faibles (inférieurs à 5) pour que l'on puisse appliquer le test du χ^2 dont on rappelle qu'il a une justifiction asymptotique. On va donc regrouper les modalités 3 et 4 de la variable "réaction" (ce qui est raisonnable vu la proximité de leurs interprétations). On obtient le tableau modifié :

Vaccin	Réaction légère	Réaction moyenne	Ulcération ou Abcès	Total
A	12	156	9	177
B	29	135	7	171
Total	41	291	16	348

On dresse alors un tableau comprenant dans chaque case (j,h) (avec désormais $1 \leq h \leq 3$), l'une au dessus de l'autre, les deux valeurs suivantes :
• l'estimation par m.v. de $p_{j,h}$ sans faire l'hypothèse d'indépendance, c'est-à-dire la proportion de couples (j,h) observée dans l'échantillon (notée $\frac{n_{j,h}}{n}$ dans le cours),
• l'estimation par m.v. de $p_{j,h}$ sous l'hypothèse d'indépendance, c'est-à-dire le produit des proportions, observées dans l'échantillon, de modalités j pour le vaccin et de modalités h pour la réaction, après regroupement (notée $\frac{n'_j n''_h}{n}$ dans le cours).

Vaccin	Réaction légère	Réaction moyenne	Ulcération ou Abcès	Total
A	0,0345	0,4483	0,0259	0,5086
	0,0599	0,4253	0,0234	0,5086
B	0,0833	0,3879	0,0201	0,4914
	0,0579	0,4109	0,0226	0,4914
Total	0,1178	0,8362	0,0460	1

La valeur de la statistique du χ^2 est alors :

$$n \sum_{j=1}^{k} \sum_{h=1}^{m} \frac{(\frac{n_{j,h}}{n} - \frac{n'_j . n''_h}{n^2})^2}{\frac{n'_j . n''_h}{n^2}} = 8,81$$

La loi (approchée asymptotiquement) de cette statistique est la loi du χ^2 à $(2-1)(3-1) = 2$ degrés de liberté. Le quantile d'ordre $0,95$ de cette loi vaut 5,991, que dépasse la valeur observée 8,81 : on rejette donc l'hypothèse d'indépendance au niveau 0,05, autrement dit les deux séries de réactions observées diffèrent significativement.

On remarque par ailleurs que 8,81 est compris entre les quantiles d'ordres 0,98 et 0,99 de la loi du χ^2 à 2 degrés de liberté ; donc, au niveau de signification 0,01, l'hypothèse d'indépendance n'aurait pu être rejetée.

▲

Exercice III.3.

Non fournie. ▲

Exercice III.4.

1. L'estimateur du maximum de vraisemblance, $\hat{p}$, de p est le vecteur des **fréquences empiriques**. On a donc $\hat{p} = (0,606; 0,186; 0,173; 0,035)$.

2. Le nombre de degrés de liberté pour ce test du χ^2 d'indépendance est (voir le polycopié) $(2-1)(2-1)=1$.

   Rappelons une argumentation heuristique couramment employée pour justifier ce nombre de degrés de liberté : la dimension du vecteur p est 4 ; mais il faut tenir compte de la contrainte $p_{J,N}+p_{W,N}+p_{J,C}+p_{W,C}=1$; enfin l'hypothèse d'indépendance revient à dire que $p=h(p_J,p_N)$, où p_J est la probabilité de naître un jour ouvrable et p_N la probabilité pour que l'accouchement soit sans césarienne ; en particulier, on a $p_{J,N}=p_Jp_N$, $p_{W,N}=(1-p_J)p_N$, $p_{J,C}=p_J(1-p_N)$ et $p_{W,C}=(1-p_J)(1-p_N)$ et il faut tenir compte des deux estimations : celle de p_J et celle de p_N ; le nombre de degrés de liberté du test du χ^2 est donc q**=4-1-2=1**.

   L'estimateur du maximum de vraisemblance $\hat{p}_J$, de p_J, et $\hat{p}_N$, de p_N, est celui des fréquences empiriques. On a donc $\hat{p}_J=0,779$, $\hat{p}_N=0,792$, $\hat{p}_W=1-\hat{p}_J$ et $\hat{p}_C=1-\hat{p}_N$. La statistique du χ^2 est :

$$\zeta_n=n\left(\frac{(\hat{p}_{J,N}-\hat{p}_J\hat{p}_N)^2}{\hat{p}_J\hat{p}_N}+\frac{(\hat{p}_{W,N}-\hat{p}_W\hat{p}_N)^2}{\hat{p}_W\hat{p}_N}+\frac{(\hat{p}_{J,C}-\hat{p}_J\hat{p}_C)^2}{\hat{p}_J\hat{p}_C}+\frac{(\hat{p}_{W,C}-\hat{p}_W\hat{p}_C)^2}{\hat{p}_W\hat{p}_C}\right).$$

   On obtient $\zeta_n\simeq 15594$

   On lit dans la table du χ^2 que $\mathbb{P}(X>11)\leq 0,1\%$, où la loi de X est $\chi^2(1)$.

3. Ici on teste l'hypothèse simple $p=p^0$, avec $p^0=(0,605;0,189;0,17;0,036)$. Le nombre de degrés de liberté de ce test du χ^2 d'adéquation est $4-1=3$ (voir le polycopié).

   La statistique du χ^2 est

$$\zeta_n=n\left(\frac{(\hat{p}_{J,N}-p^0_{J,N})^2}{p^0_{J,N}}+\frac{(\hat{p}_{W,N}-p^0_{W,N})^2}{p^0_{W,N}}+\frac{(\hat{p}_{J,C}-p^0_{J,C})^2}{p^0_{J,C}}+\frac{(\hat{p}_{W,C}-p^0_{W,C})^2}{p^0_{W,C}}\right).$$

   On obtient $\zeta_n\simeq 409$.

   On lit dans la table du χ^2 que $\mathbb{P}(X>17)\leq 0,1\%$, où la loi de X est $\chi^2(3)$. On rejette donc l'hypothèse au niveau de 99,9%. Il y a donc une évolution entre 1996 et 1997.

▲

Exercice III.5.

Non fournie ▲

Chapitre IV

Tests non paramétriques

IV.1 Énoncés

Exercice IV.1.
Des pharmacologues étudient l'effet d'une nouvelle molécule chez l'homme. Ils pensent que cette molécule permettrait l'augmentation de certains globules blancs appelés neutrophiles. Pour leur étude, ils disposent d'un groupe de 24 volontaires, parmi lesquels 12 sont effectivement traités par la nouvelle molécule et 12 reçoivent un placebo. On mesure la quantité (en milliers par millimètre cube) de ces neutrophiles pour chacun des 24 individus :

gp traité	4.8	4.5	4.4	5.0	4.9	5.1	5.3	5.3	5.4	5.5	5.6	5.3
gp témoin	4.6	4.9	4.2	4.6	4.5	4.3	4.5	5.0	5.2	5.3	5.4	5.2

On supposera que les volontaires sont choisis au hasard dans un large groupe, et que, si la molécule a un effet, il est nécessairement dans le sens d'une augmentation des neutrophiles.

1. En listant clairement les hypothèses que vous faites, proposez d'abord un test de Student (aux niveaux 1% et 5%) pour répondre à la question "y a-t-il une augmentation significative de neutrophiles chez les sujets traités ?". Commentez vos résultats.

2. Proposez ensuite un test de Mann-Whitney, en comparant hypothèses et résultats avec la question précédente. Discutez.

D'autres chercheurs se posent la même question mais ils ne disposent que de 12 individus pour leur étude. Ils décident donc de traiter tout le groupe et de mesurer la quantité de neutrophiles, pour chaque patient, avant et après le traitement. Ils obtiennent les résultats suivants :

avant traitement	4.2	4.3	4.5	4.5	4.5	4.6	4.9	5.0	5.2	5.2	5.3	5.4
après traitement	4.4	4.6	4.8	4.9	5.0	5.1	5.3	5.3	5.3	5.4	5.5	5.6

3. En quoi ce nouveau plan d'expérience change-t-il le problème statistique ?

4. Proposez, pour ces nouvelles données, un test non paramétrique pour répondre à la question des pharmacologues.

△

Exercice IV.2.
On dispose de 10 résultats de simulation de la loi uniforme sur l'intervalle $[0, 1]$ (obtenus par usage d'un ordre RANDOM sur un ordinateur ou calculatrice) :

0.134	0.628	0.789	0.905	0.250	0.563	0.790	0.470	0.724	0.569

A l'aide d'un test de Kolmogorov au niveau 0.20, étudiez si cet échantillon conduit à rejeter l'hypothèse nulle selon laquelle "le tirage a bien eu lieu selon la loi uniforme $[0, 1]$" (en l'occurence, le rejet serait bien sûr une conclusion erronnée).

△

Exercice IV.3.
Un statisticien s'est perdu en pleine brousse. Dans le but de construire un ballast avec des cailloux, il doit choisir entre deux carrières celle dont les cailloux sont les plus durs. Pour déterminer quel est le plus dur de deux cailloux, il ne dispose que d'un seul moyen : les frotter l'un contre l'autre. Soit n le nombre d'expériences qu'il réalise (portant chaque fois sur des couples de cailloux distincts), et N^+ le nombre d'entre elles qui donnent un caillou plus dur pour la première carrière.

1. Sous l'hypothèse $\mathcal{H}_0$: "il n'y a pas de différence entre les carrières", quelle est la loi de N^+ ? Quelle est sa loi "asymptotique", lorsque n tend vers l'infini ? On supposera que $n^+ > n/2$; le statisticien pense donc qu'il devrait choisir la première carrière.

2. En admettant que n est "assez grand" (précisez le sens de cette expression), déduisez-en un test non paramétrique simple pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$: "la première carrière contient des cailloux plus durs que la deuxième". Ce test est connu sous le nom de "test des signes".

3. Comment peut-on utiliser un test des signes pour tester l'égalité des lois de deux échantillons appariés (cas de la question 4 de l'exercice 1 par exemple) ? Quel inconvénient a-t-il par rapport au test de Wilcoxon ?

△

Exercice IV.4.
Une étude de marketing vise à révéler si la présence d'une étiquette sur une bouteille de champagne influe sur son appréciation par les consommateurs. On effectue donc des tests de consommation : 271 dégustateurs sont invités à noter sur une échelle de 1 à 11 deux champagnes supposés différents (1 est la moins bonne note et 11 la meilleure). Il s'agit en fait du même vin mais servi, dans un ordre aléatoire, par une bouteille sans étiquette pour l'un et par une bouteille avec étiquette pour l'autre.

Les résultats vous sont présentés de la manière suivante : on effectue pour chaque consommateur la différence entre la note du champagne sans étiquette et celle du champagne avec

étiquette. On observe 177 différences strictement négatives, 14 différences strictement positives et 80 différences nulles. De plus, on vous dit que la somme des rangs (dans l'ordre croissant des valeurs absolues des différences) des 14 différences strictement positives est de $656,2$.

1. Commentez brièvement le protocole expérimental et les résultats obtenus. Quel(s) test(s) proposez-vous pour éclairer les conclusions de l'étude ?

2. Effectuez ce test et concluez, en commentant les éventuelles limites.

△

Exercice IV.5.
L'étude de $N = 688$ familles ayant 7 enfants s'est traduite par la distribution suivante :

nb de garçons	7	6	5	4	3	2	1	0
nb de filles	0	1	2	3	4	5	6	7
nb de familles	8	38	106	190	188	110	40	8

On veut comparer cette distribution à la distribution théorique qui correspond à l'équiprobabilité des naissances d'un garçon et d'une fille. Proposez deux tests différents pour réaliser cette comparaison, en précisant bien les hypothèses à vérifier pour chacun. Concluez.

△

Exercice IV.6.
1. On considère l'échantillon i.i.d. suivant, pour lequel la loi commune des observations est supposée de densité continue inconnue :

$$-0.35\ -0.15\ -0.14\ 0.28\ -0.60\ 0.75\ -1.80\ 0.35\ 0.17\ 1.33\ -0.40\ -2.31\ -0.82\ -1.05$$

En vous inspirant de la construction du test de Wilcoxon vue en cours, proposez un test non paramétrique de l'hypothèse : "la densité de Z est symétrique par rapport à zéro".

2. On considère l'échantillon i.i.d. suivant, pour lequel la loi commune des observations est supposée admettre une fonction de répartition continue et strictement croissante :

$$4.65\ 4.86\ 4.40\ 3.20\ 5.17\ 4.60\ 4.18\ 4.85\ 5.28\ 5.75\ 5.35\ 6.33\ 2.69\ 3.95$$

En vous inspirant de la construction du test des signes de l'exercice 3, proposez un test non paramétrique de l'hypothèse : "la médiane est égale à 5".

△

Exercice IV.7.
Sur un échantillon de femmes on a mesuré les rythmes cardiaques suivants :

$$66\quad 74\quad 69\quad 76\quad 72\quad 73\quad 75\quad 67\quad 68$$

Sur un échantillon d'hommes les valeurs suivantes sont été relevées :

58 76 82 74 79 65 74 86

Comparez les deux distributions à l'aide d'un test non paramétrique. Indication : *on pourra utiliser un est de Kolmogorov-Smirnov à deux échantillons.*

△

IV.2 Corrections

Exercice IV.1. 1. Il s'agit de savoir si la différence entre les données du groupe traité et celles du groupe témoin est due au hasard, ou si elle provient de l'action de la molécule. On utilise dans un premier temps une approche paramétrique, avec un test unilatère de Student. On se place donc dans le cadre d'un modèle linéaire gaussien. Les hypothèses nécessaires sont :

- On modélise par une loi normale la loi du nombre de neutrophiles dans la population, les paramètres de cette loi pouvant éventuellement être modifiés par le traitement.
- Homoscédasticité (même variance en présence ou en absence de traitement).
- Indépendance des données au sein de chaque groupe et entre les deux groupes.

On reprend les notations utilisées dans le chapitre 2. Dans ce cadre, les données du premier groupe suivent une loi $\mathcal{N}(\mu_1, \sigma^2)$ et celles du second groupe une loi $\mathcal{N}(\mu_2, \sigma^2)$. L'hypothèse à tester est donc $\mathcal{H}_0$: "$\mu_1 = \mu_2$" contre "$\mu_1 > \mu_2$".
On effectue le test décrit en détail dans le cours (chap 2, 2.5) , dans lequel la statistique de Student sous $\mathcal{H}_0$ est :

$$T = \frac{(\overline{X}_1 - \overline{X}_2)\sqrt{n_1+n_2-2}}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$$

Avec $n_1 = n_2 = 12$, la zone de rejet de $\mathcal{H}_0$ pour le niveau α est ici : $[T > t_{22,\alpha}]$
On calcule les statistiques usuelles du modèle linéaire gaussien :

$$\overline{X}_1 = 5.09 \ ; \ S_1 = 0.38 \ ; \ \overline{X}_2 = 4.81 \ ; \ S_2 = 0.42$$

On trouve alors :

$$T = 1.75$$

alors que les tables statistiques donnent :

$$t_{22,5\%} = 1.72 \quad \text{et} \quad t_{22,1\%} = 2.51$$

On voit que pour un niveau de confiance de 5%, le test de Student conduit à rejeter l'hypothèse $\mathcal{H}_0$ alors que pour un test à 1%, il conduit à l'accepter. Il ne permet donc pas de conclusion "franche". De plus, des hypothèses fortes ont été faites alors qu'elles ne sont pas acquises : on ne sait rien de la réelle distribution des données qui sont peut-être loin de la normalité. De même, rien ne laisse penser que l'hypothèse d'homoscédasticité est raisonnable (rarement le cas en pharmacologie). Ces limites du test de Student nous conduisent à proposer un test non paramétrique.

2. Nous effectuons maintenant un test unilatère de Mann-Whitney pour ces deux échantillons non appariés. Il s'agit donc de tester $\mathcal{H}_0$: "la molécule n'a pas d'effet sur la quantité de neutrophiles" contre $\mathcal{H}_1$: "la molécule tend à augmenter la quantité de neutrophuiles" Ce test ne nécessite plus d'hypothèse de normalité sur les données, ni celle d'homoscédasticité. Il faut par contre garder celle d'indépendance des données, ce qui paraît raisonnable.

Calculons la statistique de Mann-Whitney. On classe les données suivant leur rang :

x_1 (traités)	x_2 (témoins)	rang
	4.2	1
	4.3	2
4.4		3
	4.5	5
	4.5	5
4.5		5
	4.6	7.5
	4.6	7.5
4.8		9
	4.9	10.5
4.9		10.5
5		12.5
	5	12.5
5.1		14
	5.2	15.5
	5.2	15.5
5.3		18.5
5.3		18.5
5.3		18.5
	5.3	18.5
	5.4	21.5
5.4		21.5
5.5		23
5.6		24

d'où, avec les notations du cours ($R_{\underline{x}_1}$ désignant ici la somme des rangs des sujets traités), et en utilisant l'approximation normale (taille de l'échantillon supérieure à 10) :

$$U_{\underline{x}_1,\underline{x}_2} = R_{\underline{x}_1} - \frac{n_1(n_1+1)}{2}$$
$$V = \frac{U_{\underline{x}_1,\underline{x}_2} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \quad \sim \quad \mathcal{N}(0,1)$$

On trouve ici que $R_{\underline{x}_1} = 178$, $U_{\underline{x}_1,\underline{x}_2} = 100$, $V = 1.61$. Or, nous sommes dans le cadre d'un test unilatère dont la zone de rejet est de la forme $[V > \phi_\alpha]$ où ϕ_α est le quantile d'ordre $(1-\alpha)$ de la loi normale centrée réduite. Les tables donnent : $\phi_{5\%} = 1.64$ et $\phi_{1\%} = 2.32$, et donc on ne peut dans aucun cas conclure à un effet significatif du traitement.

On voit sur cet exemple qu'un test non paramétrique est plus conservateur qu'un test paramétrique dans la mesure où le rejet de l'hypothèse H_0 nécessite que les données contredisent plus nettement H_0.

3. Nous avons toujours deux échantillons de même taille, mais contrairement à la question précédente, ils sont appariés : le facteur "individu" peut influer sur la valeur mesurée.

4. On propose un test unilatère de Wilcoxon pour tester la même hypothèse $\mathcal{H}_0$ qu'à la question 2. On suppose que les observations sont indépendantes mais on ne fait aucune hypothèse sur le modèle ni sur l'homoscédasticité.

x_1 (avant traitement)	x_2 (après traitement)	$x_2 - x_1$	rangs de $\lvert x_2 - x_1 \rvert$
4.2	4.4	+0.2	3.5
4.3	4.6	+0.3	7
4.5	4.8	+0.3	7
4.5	4.9	+0.4	9.5
4.5	5.0	+0.5	11.5
4.6	5.1	+0.5	11.5
4.9	5.3	+0.4	9.5
5.0	5.3	+0.3	7
5.2	5.3	+0.1	1
5.2	5.4	+0.2	3.5
5.3	5.5	+0.2	3.5
5.4	5.6	+0.2	3.5

Toutes les différences sont positives, la statistique T^+ s'obtient donc ici comme la somme de la dernière colonne du tableau, on obtient : $T^+ = 78$. Nous pouvons encore utiliser l'approximation normale :

$$V = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \quad \sim \quad \mathcal{N}(0,1)$$

on trouve $V = 3.06$. La zone de rejet a ici la même forme que pour le test de Mann-Whitney, puisque l'on procède à un test unilatère et que T^+ devient grand sous $\mathcal{H}_1$. Ainsi, V se trouve dans les zones de rejet $[V > 1.64]$ et $[V > 2.32]$ pour les deux niveaux de confiance 5% et 1%. On conclut cette fois à un effet de la molécule. Les résultats et la conclusion trouvés sont franchement différents de ceux de la question 2., alors que les différences entre les données avec et sans traitement ne sont pas plus grandes. La différence cruciale réside donc ici dans l'appariement des données. Il réduit la différence entre les données due à la variabilité entre les individus. La différence constatée est donc "plus facilement attribuable" à un effet de la molécule que dans le cas sans appariemment.

▲

Exercice IV.2.

Ce graphique représente :
- la fonction de répartition F de la loi uniforme sur $[0,1]$ (en pointillés)
- la fonction de répartition empirique $F_{\underline{x}}$ de l'échantillon $\underline{x}$ (en trait plein)

Il y apparait que la distance maximale entre elles est atteinte à gauche en 0.470 (3ème valeur) et vaut $0.470 - 0.2 = 0.270$. Or, pour $n = 10$, le quantile d'ordre 0.8 de la loi de Kolmogorov de paramètre 10 vaut 0.322. Il n'est pas dépassé par la valeur observée, il n'y a donc pas rejet de l'hypothèse nulle.

Autre manière de calculer la valeur de la statistique de Kolmogorov :

L'application $t \longmapsto \left|F_{\underline{x}}(t) - F(t)\right|$ est maximale en l'une des valeurs observées ; on les ordonne par ordre croissant :

$$0.134\,;\ 0.250\,;\ 0.470\,;\ 0.563\,;\ 0.569\,;\ 0.628\,;\ 0.724\,;\ 0.789\,;\ 0.790\,;\ 0.905$$

En la $i$ème valeur ainsi classée (notons la y_i), la fonction de répartition empirique saute de $\frac{i-1}{10}$ à $\frac{i}{10}$; donc la valeur maximale de la statistique est la plus grande des 20 valeurs : $\left|y_i - \frac{i-1}{10}\right|, \left|y_i - \frac{i}{10}\right|$ où $i = 1.10$. Voici le tableau de ces valeurs, avec en gras la plus forte valeur :

i	1	2	3	4	5	6	7	8	9	10
$\left\|y_i - \frac{i-1}{10}\right\|$	0.134	0.150	**0.270**	0.263	0.169	0.128	0.124	0.089	0.010	0.005
$\left\|y_i - \frac{i}{10}\right\|$	0.034	0.050	0.170	0.163	0.069	0.068	0.024	0.011	0.110	0.095

▲

Exercice IV.3.

1. Sous $\mathcal{H}_0$, N^+ suit une loi binomiale $\mathcal{B}(n, 1/2)$, puisqu'il s'agit alors d'une somme de n v.a. de Bernoulli indépendantes et de paramètre $1/2$. Lorsque n tend vers l'infini, la loi de N^+ tend vers une loi normale (d'après le théorème de la limite centrale) d'espérance $E(N^+) = \frac{n}{2}$ et de variance $V(N^+) = n(1/2)(1-1/2) = \frac{n}{4}$.

2. On veut maintenant tester $\mathcal{H}_0$: "il n'y a pas de différence entre les carrières", contre $\mathcal{H}_1$: "la première carrière contient des cailloux plus durs que la deuxième". Il est clair que sous $\mathcal{H}_1$, N^+ tend à être significativement plus grand que $\frac{n}{2}$. On peut donc proposer de rejeter $\mathcal{H}_0$ au niveau α pour des N^+ tels que $\frac{N^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}} > \phi_\alpha$, où ϕ_α est le quantile d'ordre $(1-\alpha)$ de la loi normale centrée réduite. Cette approximation normale est valide, avec une précision jugée en général satisfaisante, si $n(\frac{1}{2})(1-\frac{1}{2}) \geq 5$, c'est-à-dire $n \geq 20$. Ce test non paramétrique très simple ne requiert aucune hypothèse sur la forme du modèle. Il est adapté à ce genre de situation où l'on ne dispose pas de deux échantillons appariés complets mais seulement de leur comparaison paire par paire.

3. Dans les situations où l'on dispose des données chiffrées pour deux échantillons appariés, on peut se ramener au cas précédent en ne considérant que le signe des différences entre les valeurs de chaque paire. Si X_1 et X_2 sont les deux échantillons appariés, on note $Z = signe(X_1 - X_2)$, constitué de " + " et de " − ". Le test se base alors uniquement sur Z et est indépendant des valeurs quantitatives prises par X_1 et X_2. L'hypothèse $\mathcal{H}_0$ devient "il y a autant de chances d'observer un signe " + " qu'un signe " − ", et le même test des signes s'applique avec N^+ = nombre de signes " + " (d'où le nom de test des signes).

Ce test semble intuitivement moins efficace que le test de Wilcoxon qui exploite, lui, à la fois le signe et la valeur des différences. Le test des signes exploite donc moins d'information que le test de Wilcoxon. En pratique, on observe en effet que le test de Wilcoxon est très souvent bien plus puissant que le test des signes, surtout pour de petits échantillons. En revanche, la théorie montre que la différence des puissances tend à s'annuler quand la taille n des échantillons tend vers l'infini. Il est donc préférable de n'utiliser ce test des signes que lorsque l'on ne dispose pas des données chiffrées des deux échantillons à comparer.

▲

Exercice IV.1.

1. Le protocole d'étude permet de mesurer les préférences de chaque consommateur. De plus, l'ordre de dégustation des champagnes est aléatoire pour chaque consommateur, car sinon, la dégustation du premier pourrait influer sur l'appréciation du second. Ainsi, le recueil des deux notes fournit deux échantillons appariés : nous pouvons effectuer un test de Wilcoxon avec les données fournies par l'étude (la décimale sur la somme des rangs provient de l'application de la règle du rang moyen). Néanmoins, on constate que le nombre d'ex-aequo est relativement élevé, ceci pouvant provenir du fait que l'échelle de notation n'est pas utilisée entièrement, ou bien que les dégustateurs sont de véritables experts.

2. Le test de Wilcoxon permet donc de tester l'hypothèse $\mathcal{H}_0$: "l'étiquette sur la bouteille de champagne n'a pas d'influence sur son appréciation par le consommateur". On se place dans le modèle non-paramétrique de décalage suivant : la fonction de répartition commune des notes données aux bouteilles avec étiquettes est $F_\mu(t) = F(t-\mu)$ avec $\mu \in \mathbb{R}$ où F est la fonction de répartition commune des notes données aux bouteilles sans étiquette. L'hypothèse nulle est $\mathcal{H}_0 = \{\mu = 0\}$. L'énoncé laisse une certaine incertitude sur le choix de l'hypothèse alternative $\mathcal{H}_1$. Un choix raisonnable pourrait être d'admettre que, de toute façon, la présence d'une étiquette ne peut qu'influencer favorablement le dégustateur i.e. $\mathcal{H}_1 = \{\mu > 0\}$. La statistique de test T^+ est définie comme la somme des rangs des différences positives fournies par l'étude. Elle a tendance à prendre des valeurs d'autant plus faibles que l'on est plus nettement dans l'hypothèse alternative, c'est-à-dire que μ est grand. Le rejet de l'hypothèse nulle se fait donc si la valeur prise par T^+ est assez petite. De plus, la grande taille de l'échantillon nous autorise à utiliser l'approximation normale du test de Wilcoxon vue en cours. Ainsi, on a sous $\mathcal{H}_0$:

$$S = \frac{T^+ - \frac{271.272}{4}}{\sqrt{\frac{271.272.543}{24}}} \sim \mathcal{N}(0,1)$$

Dans notre cas, on trouve $s = -13,7$ ce qui nous conduit à rejeter fortement $\mathcal{H}_0$ pour tous les niveaux de signification usuels. En effet $-13,7$ est inférieur à tous les quantiles inférieurs d'ordres $0,05$, $0,01$, $0,001\ldots$ de la loi normale centrée réduite ; autrement dit la p-valeur associée à cette observation est $\Phi(-13,7)$ où Φ désigne comme il est usuel la fonction de répartition de la loi normale centrée réduite ; il s'agit d'une valeur extrêmement faible : la table de Φ donnée dans le polycopié nous apprend que $\Phi(-10) = 7,6 \times 10^{-24}$.

Si on pense que l'influence de l'étiquette peut conduire à influencer le dégustateur aussi bien défavorablement que favorablement (imaginez que l'étiquette désigne une marque connue

comme médiocre !) on choisit $\mathcal{H}_1 = \{\mu \neq 0\}$. Alors le rejet de l'hypothèse nulle se fait si $|T^+|$ est assez grand ; on compare sa valeur au quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite ; ceci ne change rien à nos conclusions en l'occurence.

Il faut cependant rester très critique sur ces résultats numériques, au vu du grand nombre d'ex-aequo qui induit forcément une erreur non négligeable dans les calculs. Une autre manière simple de procéder, plus élémentaire mais peut-être plus prudente dans ce cas, consiste à réaliser un test des signes, à la manière de l'exercice précédent. Dans ce cadre, sous $\mathcal{H}_0$, le nombre $n^+ = 14$ de différences strictement positives est une réalisation d'une loi binomiale $\mathcal{B}(n, p)$, avec $n = 271 - 80 = 191$ et $p = 1/2$, et pour laquelle on peut encore utiliser l'approximation normale. On trouve donc sous $\mathcal{H}_0$:

$$S' = \frac{N^+ - 191/2}{\sqrt{191/4}} \sim \mathcal{N}(0, 1)$$

avec $s' = -11,8$. Cela ne change pas notre conclusion.

▲

Exercice IV.5.

Non fournie. ▲

Exercice IV.6.

Non fournie ▲

Exercice IV.7.

Non fournie. ▲

Chapitre V

Analyse des données

V.1 Énoncés

Exercice V.1.

Présentation des données

La table ci-dessous est le résultat d'une étude (ancienne) de dépenses annuelles de ménages français. Les individus sont des ménages, et sont identifiés par :

- les caractéristiques professionnelles du chef de famille, autrement dit la "CSP" du ménage (MA=travailleur manuel, EM=employé non manuel, CA=cadre) ;
- le nombre d'enfants du ménage (2, 3, 4 ou 5).

Les 7 variables quantitatives (Pain, Légume,. . .) correspondent aux principaux types de produits achetés.

Ménage	Pain	Légume	Fruit	Viande	Volaille	Lait	Vin
MA2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
CA2	372	767	562	1948	927	235	433
MA3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
CA3	438	843	689	2345	1148	243	341
MA4	534	660	367	1620	0638	414	407
EM4	460	699	484	1856	762	400	416
CA4	385	789	621	2366	1149	304	282
MA5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
CA5	515	1097	887	2630	1167	561	284

Objectif

Il s'agit de "résumer" ce tableau de données à l'aide d'une ACP, afin de tenter d'expliquer les habitudes de consommation des ménages.

Il est important de souligner que les individus ne sont pas anonymes dans cette étude, ils ont un sens en terme de CSP et de nombre d'enfants. Il est, pour cette raison, important de repérer les individus par leurs identifiants sur les plans factoriels. Ainsi, si c'est possible, on tâchera de repérer des tendances à la consommation de certains produits en fonction de la CSP, ou du nombre d'enfant. On pourra aussi essayer de repérer des "classes" homogènes d'individus le long des axes, et d'interpréter ces classes en terme des nouveaux caractères.

Statistiques descriptives

Voici les statistiques descriptives élémentaires pour les 7 variables :

Statistique	Pain	Légume	Fruit	Viande	Volaille	Lait	Vin
Moyenne	446.7	732.0	505.0	1886.7	803.2	358.3	368.6
Écart-type	107.15	189.18	165.09	395.75	249.56	117.13	71.78

On peut aussi réaliser une étude descriptive des liens entre les 7 caractères à l'aide de la matrice des coefficients de corrélation empiriques :

	Pain	Légume	Fruit	Viande	Volaille	Lait	Vin
Pain	1.0000	0.5931	0.1961	0.3213	0.2480	0.8556	0.3038
Légume	0.5931	1.0000	0.8563	0.8811	0.8268	0.6628	-0.3565
Fruit	0.1961	0.8563	1.0000	0.9595	0.9255	0.3322	-0.4863
Viande	0.3213	0.8811	0.9595	1.0000	0.9818	0.3746	-0.4372
Volaille	0.2480	0.8268	0.9255	0.9818	1.0000	0.2329	-0.4002
Lait	0.8556	0.6628	0.3322	0.3746	0.2329	1.0000	0.0069
Vin	0.3038	-0.3565	-0.4863	-0.4372	-0.4002	0.0069	1.0000

Question 1. Quels groupes de caractère homogènes pouvez-vous proposer ?
Question 2. Pensez-vous qu'une ACP sur ces données donnera de bons résultats ?

Éléments de l'ACP

On réalise une ACP **non normée** sur ces données, i.e. on ne réduit pas les variables. Ceci revient à effectuer la diagonalisation de la matrice de variances-covariances.
Question 3. Pourquoi est-ce raisonnable dans cet exemple ?
On obtient les résultats suivants :

Axe	Valeur propre ($\times 10^5$)	% d'inertie	% d'inertie cumulée
1	2.748	88.003	88.003
2	0.264	08.459	96.462
3	0.063	02.003	98.465
4	0.023	00.736	99.201
5	0.021	00.669	99.871
6	0.003	00.108	99.979
7	0.001	00.021	100.00

Matrice des vecteurs propres :

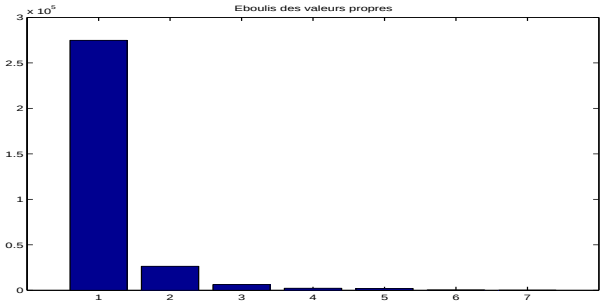


FIG. V.1 – *Eboulis des valeurs propres.*

	V^1	V^2	V^3	V^4	V^5	V^6	V^7
Pain	-0.0728	0.5758	0.4040	0.1140	-0.1687	0.6737	0.0678
Légume	-0.3281	0.4093	-0.2917	0.6077	0.4265	-0.1828	-0.2348
Fruit	-0.3026	-0.1001	-0.3402	-0.3965	0.5682	0.4320	0.3406
Viande	-0.7532	-0.1082	0.0681	-0.2942	-0.2848	-0.0011	-0.4987
Volaille	-0.4653	-0.2439	0.3809	0.3299	-0.0645	-0.2076	0.6503
Lait	-0.0911	0.6316	-0.2254	-0.4135	-0.2366	-0.4390	0.3498
Vin	0.0588	0.1444	0.6599	-0.3068	0.5705	-0.3005	-0.1741

Cartographie des individus

On représente seulement les plans 1–2 et 1–3.

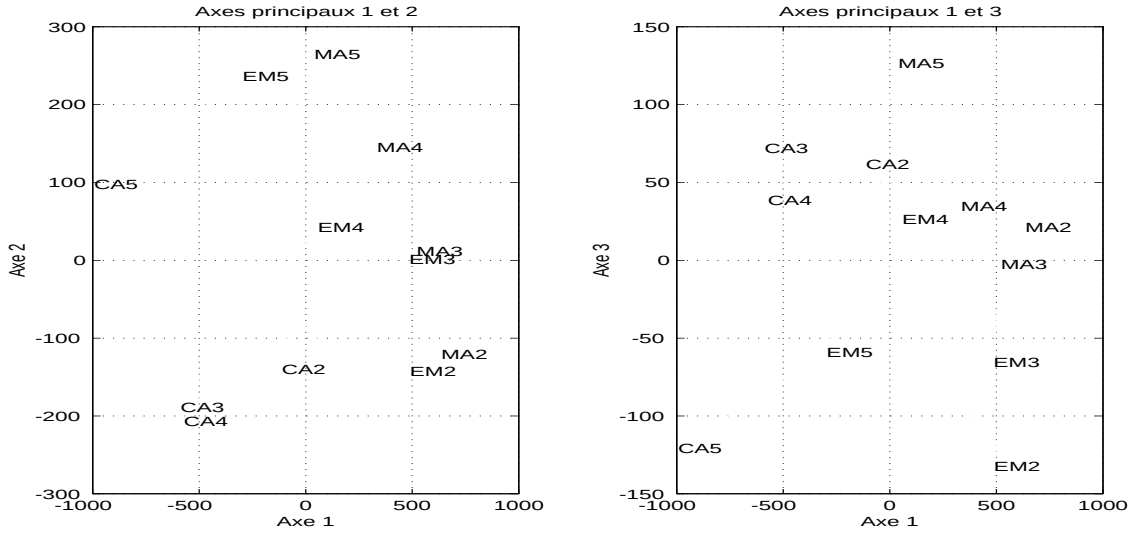


FIG. V.2 – Projections des individus dans le plan principal 1–2 (gauche) et dans le plan 1–3 (droite).

Qualités de représentation des individus dans le plan 1–2

MA2	0.9874
EM2	0.9285
CA2	0.6618
MA3	0.9979
EM3	0.9781
CA3	0.9801
MA4	0.9818
EM4	0.6168
CA4	0.9709
MA5	0.8124
EM5	0.9271
CA5	0.9786

Cartographie des caractères : cercles de corrélation

Pour faciliter la lecture des représentations, on a codé les noms des variables de la manière suivante :

Pain	P
Légume	Le
Fruit	F
Viande	Vi
Volaille	Vo
Lait	L
Vin	W

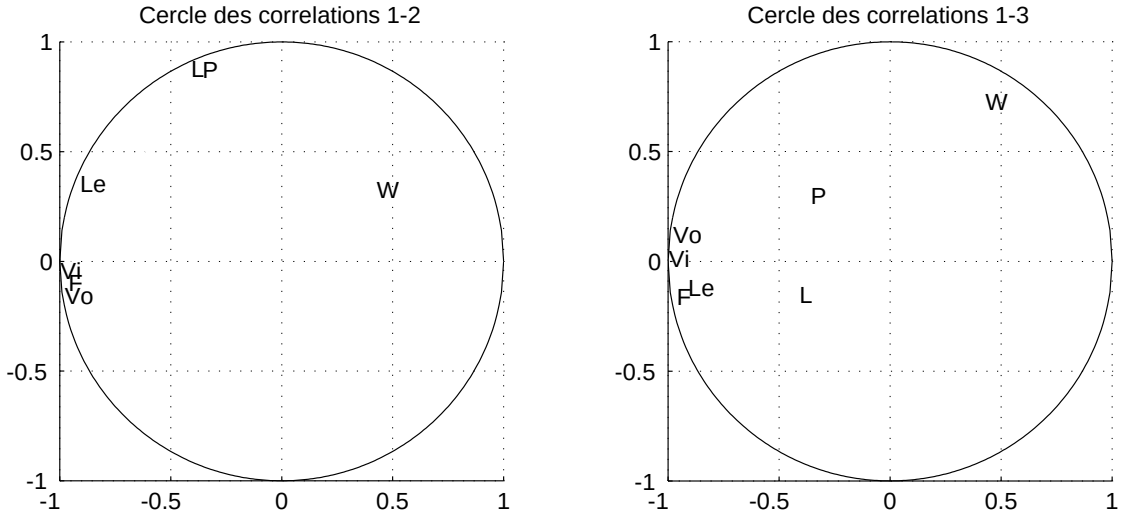


FIG. V.3 – *Cercle de corrélation 1–2 (gauche) et 1–3 (droite).*

Question 4. A partir des cercles de corrélation, donnez une signification concrète aux nouveaux caractères. Pouviez-vous déduire cette interprétation de l'examen

de la matrice des vecteurs propres ?
Question 5. A l'aide des éléments fournis, interprétez les résultats de l'ACP.

△

Exercice V.2.
Présentation du problème

La pollution de l'eau, de l'air, ... est un des problèmes les plus importants dans le domaine de l'environnement. De nombreuses études relatives à ce type de problème font appel à la Statistique et permettent de répondre à différentes questions sensibles telles que : "Est-ce que la pollution a un impact sur le taux de mortalité ?", "Peut-on construire un indicateur de pollution ?", ou encore "Y a t-il des lieux qui se comportent différemment face à la pollution ?".

Pour cela, sur un échantillon de 40 villes des Etats-Unis en 1960, 11 mesures ont été relevées, en plus du taux de mortalité :
- TMR (nombre de décès pour 10000 durant un an)
- GE65 : pourcentage ($\times 10$) de la population des 65 ans et plus,
- LPOP : logarithme (en base 10 et $\times 10$) de la population,
- NONPOOR : pourcentage de ménages avec un revenu au dessus du seuil de pauvreté,
- PERWH : pourcentage de population blanche,
- PMEAN : moyenne arithmétique des relevés réalisés deux fois par semaine de particules suspendues dans l'air ($\mu_g/m^3 \times 10$),
- PMIN : plus petite valeur des relevés réalisés deux fois par semaine de particules suspendues dans l'air ($\mu_g/m^3 \times 10$),
- PMAX : plus grande valeur des relevés réalisés deux fois par semaine de particules suspendues dans l'air ($\mu_g/m^3 \times 10$),
- SMEAN : moyenne arithmétique des relevés réalisés deux fois par semaine de sulfate ($\mu_g/m^3 \times 10$),
- SMIN : plus petite valeur des relevés réalisés deux fois par semaine de sulfate ($\mu_g/m^3 \times 10$),
- SMAX : plus grande valeur des relevés réalisés deux fois par semaine de sulfate ($\mu_g/m^3 \times 10$),
- PM2 : densité de population par mile carré ($\times 0.1$).

Le tableau relatif à ces données est fourni ci-après :

CITY	PMIN	PMEAN	PERWH	NONPOOR	GE65	LPOP
PROVIDEN	56	119	97.9	83.9	109	5.85
JACKSON	27	74	60	69.1	64	5.27
JOHNSTOW	70	166	98.7	73.3	103	5.45
JERSEY C	63	147	93.1	87.3	103	5.79
HUNTINGT	56	122	97	73.2	93	5.41
DES MOIN	61	183	95.9	87.1	97	5.43
DENVER	54	126	95.8	86.9	82	5.97
READING	34	120	98.2	86.1	112	5.44
TOLEDO	52	104	90.5	86.1	98	5.66
FRESNO	45	119	92.5	78.5	81	5.56
MEMPHIS	46	102	63.6	72.5	73	5.80
YORK	28	147	97.7	84.8	97	5.38
MILWAUKE	49	150	94.4	90.4	88	6.08
SAVANNAH	46	82	65.9	72	65	5.27
OMAHA	39	107	94	86.4	90	5.66
TOPEKA	52	101	92.7	84.1	99	5.15
COLUMBUS	74	119	88.1	86.3	79	5.83
BEAUMONT	32	76	79.3	79.9	58	5.49
WINSTON	72	147	75.8	79.9	62	5.28
DETROIT	59	146	84.9	86.5	72	6.58
EL PASO	49	150	96.7	77.9	45	5.50
MACON	22	122	69	73.7	62	5.26
ROCKFORD	36	86	95.8	88.2	85	5.32
JACKSON	39	77	94.3	86.5	90	5.12
FALL RIV	18	102	98.7	82.9	116	5.60
BOSTON	55	141	97.3	88.5	109	6.49
DAYTON	50	132	89.8	87.1	74	5.84
CHARLOTT	62	124	75.4	79.5	57	5.43
MIAMI	33	54	85.1	77.2	100	5.97
BRIDGEPO	32	91	94.7	90.7	94	5.82
SIOUX FA	25	108	99.3	82.4	92	4.94
CHICAGO	88	182	85.2	89.4	86	6.80
SOUTH BE	28	90	94	88.4	84	5.38
NORFOLK	39	89	73.6	73.1	53	5.76
CLEVELAN	86	174	85.5	88.6	89	6.25
AUSTIN	10	78	87.2	75.2	76	5.33
KNOXVILL	28	135	92.5	72.5	74	5.57
INDIANAP	92	178	85.6	87.2	85	5.84
NASHVIL	45	130	80.8	76.5	79	5.60
SEATTLE	32	69	95.2	88.8	96	6.04

TAB. V.1 – Données de pollution de l'air

CITY	**TMR**	**SMIN**	**SMEAN**	**SMAX**	**PMAX**	**PM2**
PROVIDEN	1096	30	163	349	223	116.1
JACKSON	789	29	70	161	124	21.3
JOHNSTOW	1072	88	123	245	452	15.8
JERSEY C	1199	155	229	340	253	1357.2
HUNTINGT	967	60	70	137	219	18.1
DES MOIN	950	31	88	188	329	44.8
DENVER	841	2	61	188	229	25.4
READING	1113	50	94	186	242	31.9
TOLEDO	1031	67	86	309	193	133.2
FRESNO	845	18	34	198	304	6.1
MEMPHIS	873	35	48	69	201	83.5
YORK	957	120	162	488	408	26.2
MILWAUKE	921	65	134	236	299	150.2
SAVANNAH	990	49	71	120	192	42.7
OMAHA	922	20	74	148	198	29.9
TOPEKA	904	19	37	91	158	25.9
COLUMBUS	877	94	161	276	190	127.2
BEAUMONT	728	27	71	144	190	23.5
WINSTON	802	28	58	128	306	44.7
DETROIT	817	52	128	260	235	191.5
EL PASO	618	47	87	207	373	29.8
MACON	869	18	27	128	754	28.6
ROCKFORD	842	33	66	210	143	40.3
JACKSON	928	41	52	138	124	18.7
FALL RIV	1157	62	79	136	254	71.7
BOSTON	1112	42	163	337	252	174.5
DAYTON	847	18	106	241	327	53.9
CHARLOTT	791	43	81	147	234	50.2
MIAMI	897	44	57	68	124	45.5
BRIDGEPO	938	137	205	308	182	103.3
SIOUX FA	795	18	55	121	358	10.6
CHICAGO	1000	75	166	328	296	167.5
SOUTH BE	888	73	77	261	164	51.1
NORFOLK	803	49	112	198	242	86.7
CLEVELAN	969	69	160	282	336	261.1
AUSTIN	689	40	46	58	157	20.9
KNOXVILL	825	56	77	157	302	25.8
INDIANAP	969	50	139	269	275	173.5
NASHVIL	919	54	160	362	310	75.1
SEATTLE	938	1	47	179	141	26.2

Tab. V.2 – Données de pollution de l'air

	GE65	LPOP	NONPOOR	PERWH	PMEAN	PMIN
Minimum	45	4.94	69.1	60	54	10
Q1	73.5	5.38	76.85	85	90.5	32
Médiane	85.5	5.58	84.45	92.6	119.5	46
Q3	97	5.84	87.15	95.85	146.5	57.5
Maximum	116	6.79	90.7	99.3	183	92
moyenne	84.28	5.65	82.22	88.29	119.23	47.1
Ecart-type	17.18	0.4	6.33	10.54	33.33	19.24

TAB. V.3 – Statistiques descriptives

	SMAX	SMEAN	SMIN	PM2	PMAX	TMR
Minimum	58	27	1	6.1	124	618
Q1	137.5	59.5	28.5	25.85	190	833
Médiane	193	80	45.5	44.75	238.5	911.5
Q3	272.5	136.5	63.5	109.7	305	969
Maximum	488	229	155	1357.2	754	1199
Moyenne	209.9	98.1	50.23	100.76	257.33	912.2
Ecart-type	94.73	50.13	33.18	212.57	113.6	124.37

TAB. V.4 – Statistiques descriptives

Étude descriptive

Le premier réflexe lorsque que l'on étudie des données est de les regarder, notamment à l'aide de quelques statistiques descriptives sur l'ensemble des variables, comme ci-dessous :

La figure suivante visualise la distribution empirique de chaque variable sous forme d'un histogramme. La discrétisation a été réalisée automatiquement sans volonté d'optimisation de largeur des barres et de nombre d'individus par pas de discrétrisation.

Question 1 : Que tire t-on de l'ensemble de ces informations ?

Les nuages d'individus des variables croisées 2 à 2 (scatter-plots) et la matrice de corrélations permettent d'aller plus loin dans l'étude descriptive car elles exhibent les relations entre variables.

Question 2 : Est-ce que ces deux représentations sont redondantes, ou au contraire complémentaires, et pourquoi ?

Variables	**GE65**	**LPOP**	**NONPOOR**	**PERWH**	**PMEAN**	**PMIN**
GE65	1	0.1592	0.4789	0.6655	0.084	0.0226
LPOP	0.1592	1	0.4304	0.0612	0.3613	0.4725
NONPOOR	0.4789	0.4304	1	0.5771	0.2639	0.2904
PERWH	0.6655	0.0612	0.5771	1	0.2172	-0.0134
PMEAN	0.084	0.3613	0.2639	0.2172	1	0.7088
PMIN	0.0226	0.4725	0.2904	-0.0134	0.7088	1
SMAX	0.2864	0.4252	0.475	0.3156	0.4868	0.3472
SMEAN	0.2842	0.53	0.4229	0.2092	0.4906	0.4546
SMIN	0.2611	0.1655	0.1989	0.1834	0.2588	0.18
PM2	0.2089	0.2673	0.2545	0.0574	0.2592	0.2979
PMAX	-0.1453	-0.0735	-0.169	-0.0279	0.5576	0.0839
TMR	0.8079	0.2606	0.3386	0.335	0.2379	0.2485

Tab. V.5 – Matrice de corrélations

Variables	**SMAX**	**SMEAN**	**SMIN**	**PM2**	**PMAX**	**TMR**
GE65	0.2864	0.2842	0.2611	0.2089	-0.1453	0.8079
LPOP	0.4252	0.53	0.1655	0.2673	-0.0735	0.2606
NONPOOR	0.475	0.4229	0.1989	0.2545	-0.169	0.3386
PERWH	0.3156	0.2092	0.1834	0.0574	-0.0279	0.335
PMEAN	0.4868	0.4906	0.2588	0.2592	0.5576	0.2379
PMIN	0.3472	0.4546	0.18	0.2979	0.0839	0.2485
SMAX	1	0.8245	0.5862	0.3515	0.181	0.4191
SMEAN	0.8245	1	0.7568	0.5818	0.0616	0.4805
SMIN	0.5862	0.7568	1	0.5754	0.0357	0.4235
PM2	0.3515	0.5818	0.5754	1	-0.0078	0.444
PMAX	0.181	0.0616	0.0357	-0.0078	1	0.0155
TMR	0.4191	0.4805	0.4235	0.444	0.0155	1

Tab. V.6 – Matrice de corrélations

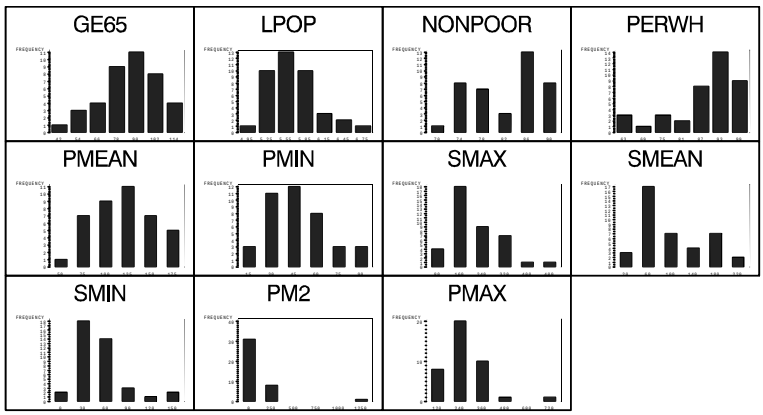


FIG. V.4 – Distributions empiriques des variables

Analyse en Composantes Principales

Une ACP est effectuée sur les données, sauf TMR qui est considérée comme une variable supplémentaire.

Question 3 : Pourquoi est-il légitime de travailler sur la matrice de corrélations pour mettre en oeuvre l'ACP ?

Question 4 : D'après l'histogramme des valeurs propres ci-dessous, combien de composantes est-il raisonnable de retenir ?

Question 5 : Qu'observe t-on sur le cercle de corrélations du plan 1-2 ?

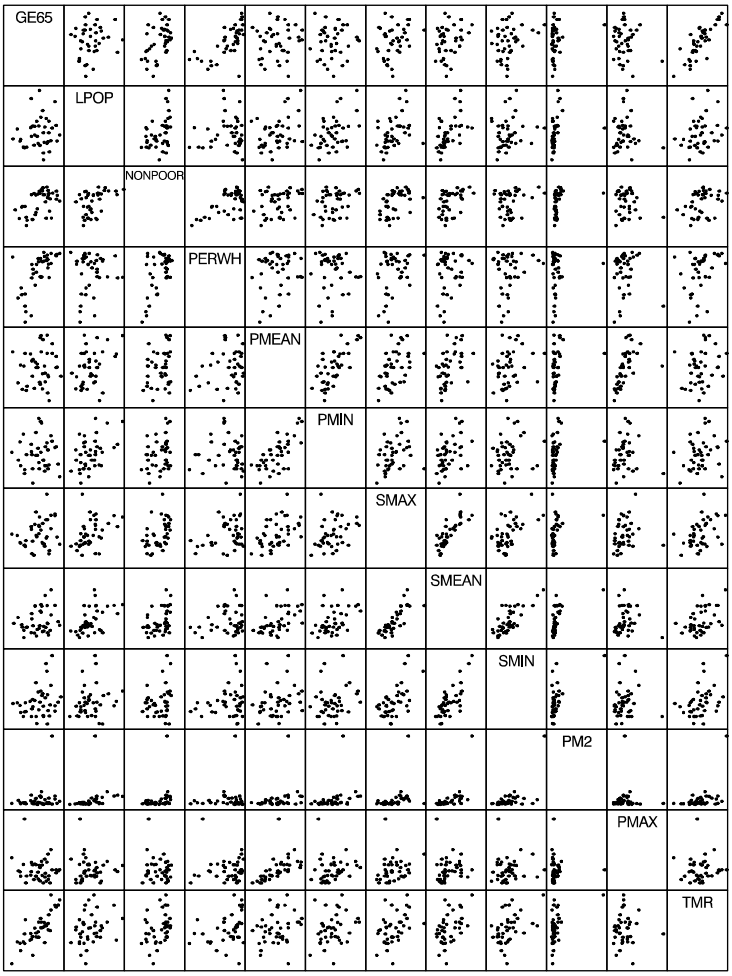


FIG. V.5 – Scatter plots

Question 6 : Que remarque t-on pour la variable supplémentaire TMR ?

Les vecteurs propres associés aux deux premières composantes sont les suivants :

Question 7 : Pour chaque composante principale, calculer quelques corrélations avec les variables actives.

Question 8 : Commenter le plan 1-2 des coordonnées des individus ci-dessous.

Question 9 : Calculer quelques coordonnées d'individus sur le plan 1-2, notamment pour la ville de Jersey City (à droite sur l'axe 1). Les calculs devront être détaillés, le tableau des coordonnées des individus permettant de vérifier les résultats obtenus.

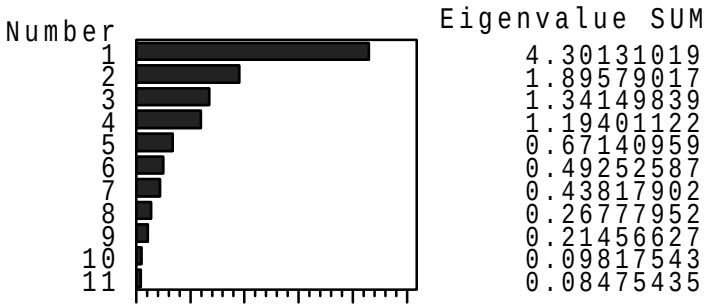


FIG. V.6 – Histogramme des valeurs propres

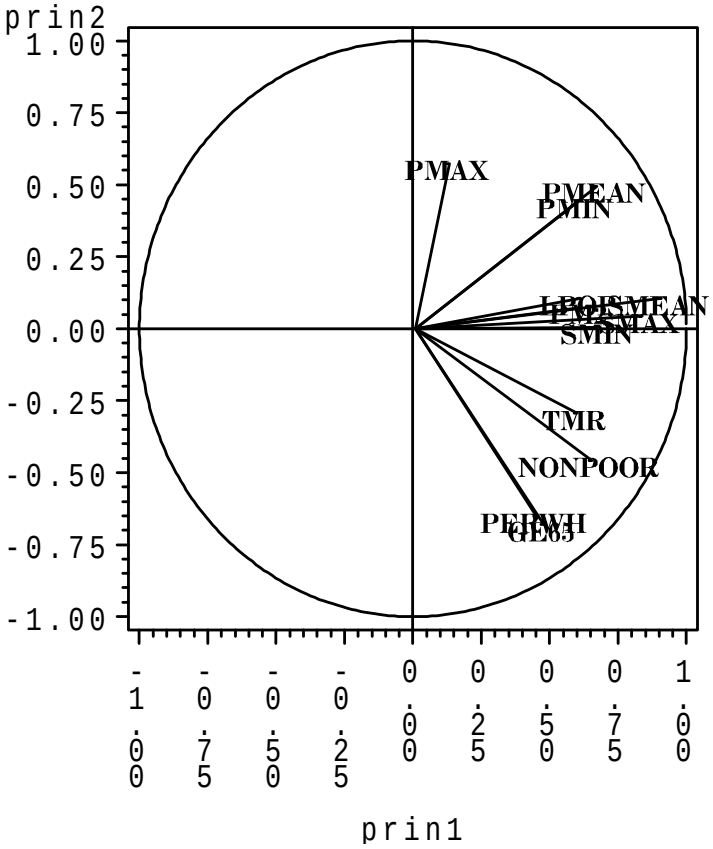


FIG. V.7 – Cercle des corrélations du plan 1-2

Question 10 : Compléter les deux tableaux suivants.

Variables	u_1	u_2	u_3	u_4
GE65	0.2253	-0.4941	0.1536	0.1301
LPOP	0.2904	0.0764	0.0069	-0.5397
NONPOOR	0.3101	-0.3327	0.2062	-0.2514
PERWH	0.2130	-0.4736	0.3829	0.2105
PMEAN	0.3191	0.3596	0.4071	0.0596
PMIN	0.2851	0.3201	0.1658	-0.3970
SMAX	0.3990	0.0319	-0.0324	0.1428
SMEAN	0.4325	0.0759	-0.2481	0.0564
SMIN	0.3249	0.0035	-0.4321	0.3636
PM2	0.2914	0.0536	-0.4224	0.0846
PMAX	0.0599	0.4166	0.4078	0.5101

TAB. V.7 – Vecteurs propres des quatre premières composantes

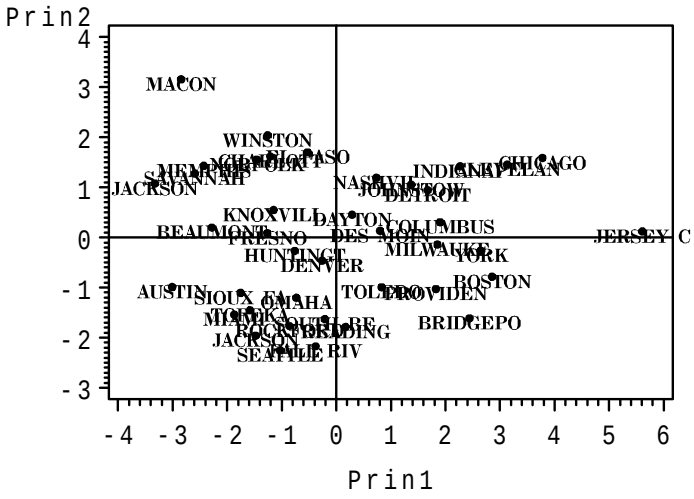


FIG. V.8 – Nuage des individus dans le plan 1-2

Question 11 : Quelles informations supplémentaires apportent ces tableaux par rapport au nuage de points des individus ?

Question 12 : Serait-il judicieux de retirer la ville de Jersey City de l'analyse et pourquoi ?

△

Exercice V.3.
On considère le tableau de données suivant concernant 10 villes françaises (Origine des données : Météofrance).

CITY	C^1	C^2	C^3	C^4
PROVIDEN	1.8271	-1.0266	0.4432	-0.2295
JACKSON	-3.3231	1.0784	-2.2568	-0.3177
JOHNSTOW	1.3756	1.0548	1.2606	1.9276
JERSEY C	5.611	0.1215	-3.5828	1.5539
HUNTINGT	-0.7598	-0.2651	0.2703	0.514
DES MOIN	0.8042	0.1352	2.1204	0.2342
DENVER	-0.2578	-0.4623	1.4198	-1.3648
READING	0.1717	-1.7851	0.7404	0.6743
TOLEDO	0.8333	-0.9909	-0.3021	-0.0977
FRESNO	-1.2631	0.0885	1.0768	0.1012
MEMPHIS	-2.4293	1.4315	-1.205	-1.0873
YORK	2.649	-0.2639	0.0785	2.9008
MILWAUKE	1.8552	-0.1379	0.5936	-0.2777
SAVANNAH	-2.5916	1.2797	-1.7274	-0.2167
OMAHA	-0.7302	-1.2026	0.6377	-0.6178
TOPEKA	-1.5805	-1.4507	0.706	-0.3972
COLUMBUS	1.9036	0.3068	-0.8873	-0.6442
BEAUMONT	-2.2797	0.1997	-0.927	-0.5426
WINSTON	-1.2556	2.0396	0.621	-0.4992
DETROIT	1.677	0.9454	-0.0967	-1.7003
EL PASO	-0.5239	1.7006	0.8574	0.7087
MACON	-2.8437	3.1563	1.3595	2.4874
ROCKFORD	-0.857	-1.7641	0.0598	-0.2275
JACKSON	-1.4774	-1.9564	-0.1284	-0.0905
FALL RIV	-0.3776	-2.1686	0.2342	1.0331
BOSTON	2.8581	-0.7814	0.6782	-0.9519
DAYTON	0.2914	0.4553	1.0194	-0.5328
CHARLOTT	-1.2087	1.618	-0.4007	-0.6774
MIAMI	-1.8682	-1.5371	-1.0872	-0.9446
BRIDGEPO	2.4437	-1.6104	-1.8443	0.7919
SIOUX FA	-1.7527	-1.1038	1.339	1.5509
CHICAGO	3.7796	1.5897	0.583	-1.869
SOUTH BE	-0.2093	-1.6265	-0.566	0.4442
NORFOLK	-1.4551	1.5556	-1.6232	-0.2894
CLEVELAN	3.1351	1.455	0.5516	-0.9822
AUSTIN	-3.0075	-0.9877	-0.9295	0.4448
KNOXVILL	-1.1495	0.5485	0.1277	1.0644
INDIANAP	2.2722	1.4201	0.8783	-1.0799
NASHVIL	0.7372	1.192	-0.5604	0.738
SEATTLE	-1.0237	-2.2513	0.4685	-1.5316

TAB. V.8 – Coordonnées des individus

CITY	CTR_1	CTR_2	CTR_3	CTR_4	CTR
PROVIDEN	0.0199	0.0143	0.0038	0.0011	0.018
JACKSON	0.0658	0.0157	0.0973	0.0022	0.0448
JOHNSTOW	0.0113	0.015		0.0798	0.0293
JERSEY C			0.2454	0.0519	0.1332
HUNTINGT	0.0034	0.001	0.0014	0.0057	0.0113
DES MOIN	0.0039	0.0002	0.0859	0.0012	0.0165
DENVER	0.0004	0.0029	0.0385	0.04	0.0111
READING	0.0002	0.0431	0.0105	0.0098	0.0112
TOLEDO	0.0041	0.0133	0.0017		0.0072
FRESNO	0.0095	0.0001	0.0222	0.0002	0.0083
MEMPHIS	0.0352	0.0277	0.0278	0.0254	0.0288
YORK	0.0418	0.0009	0.0001	0.1807	0.0471
MILWAUKE	0.0205	0.0003	0.0067	0.0017	0.0116
SAVANNAH	0.04		0.057	0.001	0.0283
OMAHA	0.0032	0.0196	0.0078	0.0082	0.0071
TOPEKA	0.0149	0.0285	0.0095	0.0034	0.0181
COLUMBUS	0.0216	0.0013	0.015	0.0089	0.0159
BEAUMONT	0.031	0.0005	0.0164	0.0063	0.0173
WINSTON	0.0094	0.0563		0.0054	0.02
DETROIT	0.0168	0.0121	0.0002	0.0621	0.0191
EL PASO	0.0016	0.0391	0.0141	0.0108	0.0199
MACON	0.0482	0.1347	0.0353	0.1329	0.0757
ROCKFORD	0.0044	0.0421	0.0001	0.0011	0.0121
JACKSON		0.0518	0.0003	0.0002	0.0174
FALL RIV	0.0009	0.0636	0.001	0.0229	0.0183
BOSTON	0.0487	0.0083	0.0088	0.0195	0.0289
DAYTON	0.0005	0.0028	0.0199		0.0067
CHARLOTT	0.0087	0.0354	0.0031	0.0099	0.0136
MIAMI	0.0208	0.032		0.0192	0.0255
BRIDGEPO	0.0356		0.065	0.0135	0.0393
SIOUX FA	0.0183	0.0165	0.0343	0.0517	0.022
CHICAGO	0.0852	0.0342	0.0065	0.075	0.0504
SOUTH BE	0.0003	0.0358	0.0061	0.0042	0.012
NORFOLK	0.0126	0.0327	0.0504	0.0018	0.0199
CLEVELAN	0.0586	0.0286	0.0058	0.0207	0.0318
AUSTIN	0.0539	0.0132	0.0165	0.0042	0.0281
KNOXVILL	0.0079	0.0041	0.0003	0.0243	0.0115
INDIANAP	0.0308	0.0273	0.0147	0.025	0.0248
NASHVIL	0.0032	0.0192	0.006	0.0117	0.0137
SEATTLE	0.0062	0.0686		0.0504	0.0241

TAB. V.9 – Contributions des individus

CITY	d^2	$CO2_1$	$CO2_2$	$CO2_3$	$CO2_4$
PROVIDEN	7.9373	0.4314	0.1362	0.0254	0.0068
JACKSON	19.7305	0.5741	0.0605		0.0052
JOHNSTOW	12.8964	0.1505	0.0885	0.1264	0.2955
JERSEY C	58.603			0.2247	0.0423
HUNTINGT	4.9548	0.1195	0.0145	0.0151	0.0547
DES MOIN	7.253	0.0915	0.0026	0.6358	0.0078
DENVER	4.8737	0.014	0.045	0.4242	0.392
READING	4.9318	0.0061	0.6627	0.114	0.0946
TOLEDO	3.1639	0.2251	0.3183	0.0296	
FRESNO	3.6564	0.4475	0.0022	0.3252	0.0029
MEMPHIS	12.6598	0.4781	0.166	0.1176	0.0958
YORK	20.735	0.3471	0.0034	0.0003	0.4162
MILWAUKE	5.1249	0.6888	0.0038	0.0705	0.0154
SAVANNAH	12.4309	0.5541		0.2462	0.0039
OMAHA	3.1027	0.1763	0.4781	0.1344	0.1262
TOPEKA	7.9738	0.3213	0.2707	0.0641	0.0203
COLUMBUS	7.0092	0.5303	0.0138	0.1152	0.0607
BEAUMONT	7.6159	0.6999	0.0054	0.1157	0.0396
WINSTON	8.7794	0.1842	0.486		0.0291
DETROIT	8.4163	0.3427	0.1089	0.0011	0.3523
EL PASO	8.772	0.0321	0.3381	0.0859	0.0587
MACON	33.2996	0.2491	0.3068	0.0569	0.1906
ROCKFORD	5.3231	0.1415	0.5997	0.0007	0.01
JACKSON	7.6608		0.5124	0.0022	0.0011
FALL RIV	8.0712	0.0181	0.5976	0.007	0.1356
BOSTON	12.707	0.6593	0.0493	0.0371	0.0731
DAYTON	2.9338	0.0297	0.0725	0.3633	
CHARLOTT	5.9738	0.2508	0.4495	0.0276	0.0788
MIAMI	11.2189	0.3191	0.216		0.0816
BRIDGEPO	17.3116	0.3538		0.2015	0.0372
SIOUX FA	9.6835	0.3254	0.129	0.1899	0.2548
CHICAGO	22.1974	0.6601	0.1168	0.0157	0.1614
SOUTH BE	5.279	0.0085	0.514	0.0622	0.0383
NORFOLK	8.7374	0.2486	0.284	0.3093	0.0098
CLEVELAN	14.0001	0.72	0.1551	0.0223	0.0707
AUSTIN	12.3625	0.7504	0.0809	0.0717	0.0164
KNOXVILL	5.0562	0.268	0.061	0.0033	0.2298
INDIANAP	10.9329	0.4843	0.1892	0.0724	0.1094
NASHVIL	6.0447	0.0922	0.2411	0.0533	0.0924
SEATTLE	10.5856	0.1015	0.4911		0.2273

TAB. V.10 – Cosinus carrés des individus

Numéro	Ville	X_1	X_2	X_3	X_4
1	Biarritz	1474	1921	7.6	19.7
2	Brest	1157	1757	6.1	15.6
3	Clermont	571	1899	2.6	19.4
4	Lille	612	1641	2.4	17.1
5	Lyon	828	2036	2.1	20.7
6	Marseille	533	2866	5.5	23.3
7	Nice	868	2779	7.5	22.7
8	Paris	624	1814	3.4	19.1
9	Perpignan	628	2603	7.5	23.8
10	Strasbourg	719	1696	0.4	19.0
Moyenne		801	2101	4.5	20.0
Ecart-type		285	442	2.51	2.51

Les variables étudiées sont :
- X_1 :Hauteur moyenne des précipitations par an
- X_2 :Durée annuelle d' ensoleillement en heures
- X_3 :Température moyenne du mois de Janvier en degré Celsius
- X_4 :Température moyenne du mois de Juillet en degré Celsius

On a effectué une Analyse en Composantes Principales sur les données normalisées, dont les résultats sont rassemblés en annexe.

- A - Analyse en composantes principales

1. *Calculer la part d'inertie portée par le premier plan factoriel.*
2. *Déterminer les corrélations entre les caractères initiaux et les 2 premiers axes et représenter le cercle des corrélations.*
3. *Quelles réflexions peut-on faire sur les données à partir du cercle des corrélations et de la projection sur premier plan factoriel ?*
4. *Calculer la contribution de la ville de Biarritz à l'inertie des 2 premiers axes et la qualité de sa projection sur ces 2 axes.*
5. *Quelles sont les villes particulièrement caractéristiques sur les 2 premiers axes ?*

- B - Classification

On décide de conserver uniquement les coordonnées des 2 caractères principaux. Le tableau des distances euclidiennes entre les villes se trouve dans l'annexe. On choisit comme stratégie d'aggrégation la stratégie du minimum.

1. *Déterminer la classification ascendante hiérarchique .*
2. *Expliciter la classification en 3 classes associée.*

Annexe

Matrice de corrélation

$$R = \begin{pmatrix} 1.0000 & -0.2178 & 0.4709 & -0.3047 \\ -0.2178 & 1.0000 & 0.6026 & 0.8925 \\ 0.4709 & 0.6026 & 1.0000 & 0.4039 \\ -0.3047 & 0.8925 & 0.4039 & 1.0000 \end{pmatrix}$$

Valeurs propres et vecteurs propres

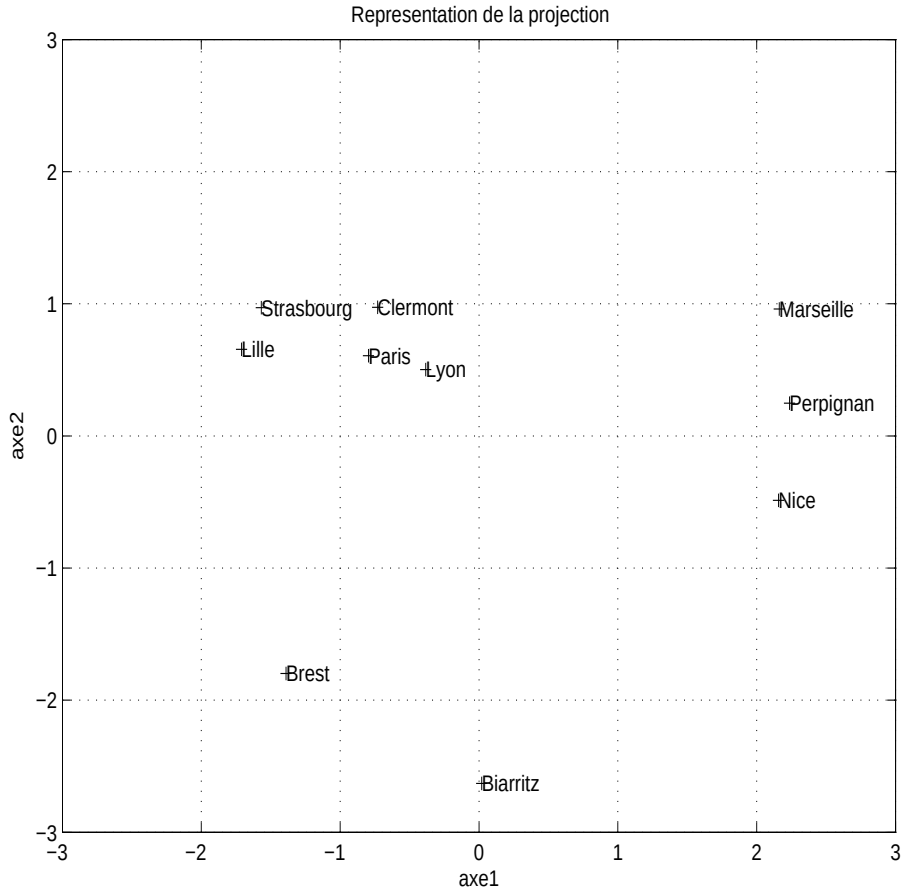
Valeurs propres		Vecteurs propres				
$\boldsymbol{\lambda_1}$	2.30	$\boldsymbol{u_1}$	-0.08	0.64	0.46	0.61
$\boldsymbol{\lambda_2}$	1.43	$\boldsymbol{u_2}$	-0.80	0.08	-0.55	0.22
$\boldsymbol{\lambda_3}$	0.21	$\boldsymbol{u_3}$	0.56	-0.07	-0.58	0.59
$\boldsymbol{\lambda_4}$	0.06	$\boldsymbol{u_4}$	-0.21	-0.76	0.38	0.49

Coordonnées, contributions à l'inertie et qualité de projection

Numéro	Ville	C^α		Ctr_α		$CO2_\alpha$	
		Axe 1	Axe 2	Axe 1	Axe 2	Axe 1	Axe 2
1	Biarritz	0.02	-2.63				
2	Brest	-1.39	-1.80	0.084	0.227	0.340	0.570
3	Clermont	-0.73	0.97	0.023	0.066	0.354	0.625
4	Lille	-1.71	0.65	0.127	0.029	0.812	0.117
5	Lyon	-0.38	0.50	0.006	0.017	0.141	0.245
6	Marseille	2.16	0.96	0.203	0.064	0.815	0.161
7	Nice	2.16	-0.49	0.203	0.017	0.943	0.049
8	Paris	-0.80	0.61	0.028	0.026	0.559	0.325
9	Perpignan	2.24	0.25	0.218	0.004	0.943	0.012
10	Strasbourg	-1.57	0.97	0.107	0.066	0.652	0.250

Distance entre les villes dans le plan principal

—-—-———*1-—*9 Biarritz	1	0									
—-—-———*2-—*8 Brest	2	2.10	0								
—-—-———*3-—*7 Clermont	3	3.74	2.93	0							
—-—-———*4-—*6 Lille	4	3.86	2.50	1.10	0						
—-—-———*5-—*5 Lyon	5	3.19	2.90	1.10	1.86	0					
—-—-———*6-—*4 Marseille	6	4.27	4.53	2.93	3.92	2.74	0				
—-—-———*7-—*3 Nice	7	3.12	3.83	3.25	4.07	2.84	1.45	0			
—-—-———*8-—*2 Paris	8	3.43	2.57	0.43	0.97	1.20	3.04	3.20	0		
—-—-———*9-—*1 Perpignan	9	3.72	4.25	3.08	4.00	2.89	1.07	1.03	3.06	0	
—-—-———*10-— Strasbourg	10	3.95	3.06	1.13	1.17	1.29	3.81	4.06	1.27	4.00	0
		1	2	3	4	5	6	7	8	9	10


Representation de la projection
Strasbourg
Clermont
Marseille
Lille
Paris
Lyon
Perpignan
Nice
Brest
Biarritz
axe1
axe2
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

△

Exercice V.4.
On donne la description d'une analyse appelée "le canidé de Jussac", effectuée sur des données réelles ; on fournit à ce propos des résultats de calculs de statistique descriptive et des éléments sur une analyse en composantes principales normée.

Présentation des données

Le crâne d'un animal préhistorique appartenant à la famille des canidés a été découvert il y a quelques années, dans la région de Jussac (Auvergne). L'une des questions que se posaient les scientifiques était de savoir si cet animal se rapprochait plus d'un chien ou d'un loup.

On a mesuré six grandeurs caractéristiques sur des crânes chiens de même taille que celle de l'animal inconnu (berger allemand, lévrier, doberman,...), et sur des crânes de loups.

Les variables mesurées sont :

- X_1 : longueur condylo-basale (LCB)
- X_2 : longueur de la mâchoire supérieure (LMS)
- X_3 : largeur bi-maxilaire (LBM)
- X_4 : longueur de la carnassière supérieure (LP)
- X_5 : longueur de la première molaire supérieure (LM)
- X_6 : largeur de la première molaire supérieure (LAM)

Les mesures figurent dans la table ci-dessous

Type	X_1	X_2	X_3	X_4	X_5	X_6
Chien	129	064	95	17.5	11.2	13.8
Chien	154	074	76	20.0	14.2	16.5
Chien	170	087	71	17.9	12.3	15.9
Chien	188	094	73	19.5	13.3	14.8
Chien	161	081	55	17.1	12.1	13.0
Chien	164	090	58	17.5	12.7	14.7
Chien	203	109	65	20.7	14.0	16.8
Chien	178	097	57	17.3	12.8	14.3
Chien	212	114	65	20.5	14.3	15.5
Chien	221	123	62	21.2	15.2	17.0
Chien	183	097	52	19.3	12.9	13.5
Chien	212	112	65	19.7	14.2	16.0
Chien	220	117	70	19.8	14.3	15.6
Chien	216	113	72	20.5	14.4	17.7
Chien	216	112	75	19.6	14.0	16.4
Chien	205	110	68	20.8	14.1	16.4
Chien	228	122	78	22.5	14.2	17.8
Chien	218	112	65	20.3	13.9	17.0
Chien	190	093	78	19.7	13.2	14.0
Chien	212	111	73	20.5	13.7	16.6
Chien	201	105	70	19.8	14.3	15.9
Chien	196	106	67	18.5	12.6	14.2
Chien	158	071	71	16.7	12.5	13.3
Chien	255	126	86	21.4	15.0	18.0
Chien	234	113	83	21.3	14.8	17.0
Chien	205	105	70	19.0	12.4	14.9
Chien	186	097	62	19.0	13.2	14.2
Chien	241	119	87	21.0	14.7	18.3
Chien	220	111	88	22.5	15.4	18.0
Chien	242	120	85	19.9	15.3	17.6
Loup	199	105	73	23.4	15.0	19.1
Loup	227	117	77	25.0	15.3	18.6
Loup	228	122	82	24.7	15.0	18.5
Loup	232	123	83	25.3	16.8	15.5
Loup	231	121	78	23.5	16.5	19.6
Loup	215	118	74	25.7	15.7	19.0
Loup	184	100	69	23.3	15.8	19.7
Loup	175	094	73	22.2	14.8	17.0
Loup	239	124	77	25.0	16.8	27.0
Loup	203	109	70	23.3	15.0	18.7
Loup	226	118	72	26.0	16.0	19.4
Loup	226	119	77	26.5	16.8	19.3
Jussac	210	103	72	20.5	14.0	16.7

Statistiques descriptives

Voici les statistiques descriptives élémentaires pour les 6 variables. On donne les moyennes et écarts-type pour l'ensemble des observations à l'exception de celle correspondant au crâne inconnu, puis par groupe (chiens et loups).

Statistique	LCB	LMS	LBM	LP	LM	LAM
Moyenne	204.8333	106.5476	72.5476	21.069	14.3024	16.8119
Écart-type	27.6528	15.1725	9.232	2.6265	1.3659	2.4922
Moyenne chiens	200.6	103.5	71.4	19.7	13.7067	15.8233
Moyenne loups	215.4167	114.1667	75.4167	24.4917	15.7917	19.2833
Écart-type chiens	29.2641	15.9757	10.4142	1.5175	1.0589	1.5662
Écart-type loups	20.5270	9.8242	4.3788	1.3228	0.7810	2.7119

N.B. L"écart-type calculé ici est à chaque fois la racine carrée de l'estimation sans biais de la variance pour la population population concernée.

On peut aussi réaliser une étude descriptive des liens entre les 6 caractères à l'aide de la matrice des coefficients de corrélation empiriques :

	LCB	LMS	LBM	LP	LM	LAM
LCB	1.0000	0.9608	0.3486	0.6145	0.7196	0.5877
LMS	0.9608	1.0000	0.2001	0.6606	0.7356	0.5948
LBM	0.3486	0.2001	1.0000	0.3699	0.3502	0.3547
LP	0.6145	0.6606	0.3699	1.0000	0.8934	0.7629
LM	0.7196	0.7356	0.3502	0.8934	1.0000	0.7895
LAM	0.5877	0.5948	0.3547	0.7629	0.7895	1.0000

Sorties de l'ACP

On réalise une ACP normée sur ces données, à l'exception de l'observation correspondant au crâne inconnu que l'on garde comme élément supplémentaire. On obtient les résultats suivants :

Axe	Valeur propre	% d'inertie	% d'inertie cumulée
1	4.1021	68.3678	68.3678
2	0.8828	14.7132	83.0810
3	0.6387	10.6453	93.7262
4	0.2590	4.3158	98.0421
5	0.0974	1.6235	99.6656
6	0.0201	0.3344	100.0000

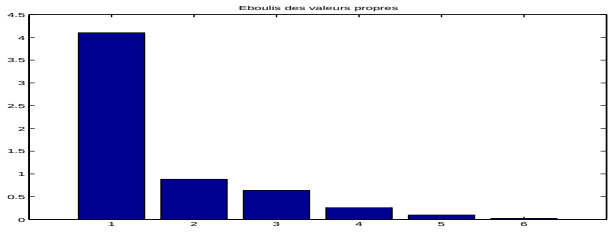


FIG. V.9 – *Eboulis des valeurs propres.*

Matrice des vecteurs propres : la colonne V^j (où $1 \leq j \leq 6$), donne les 6 composantes du j-ème vecteur propre normé (ceux-ci étant classés selon l'ordre décroissant des valeurs propres auxquelles ils sont associés).

Composante	V^1	V^2	V^3	V^4	V^5	V^6
1	0.4313	0.2285	-0.5285	-0.1056	0.0462	0.6850
2	0.4305	0.3807	-0.3924	-0.0125	-0.2018	-0.6891
3	0.2280	-0.8880	-0.3756	0.0212	-0.0034	-0.1336
4	0.4389	-0.0663	0.3969	0.5262	-0.5821	0.1723
5	0.4600	0.0206	0.2730	0.3073	0.7815	-0.0912
6	0.4153	-0.0971	0.4400	-0.7854	-0.0873	-0.0049

Cartographie des individus

On représente seulement les plans 1–2 et 1–3.

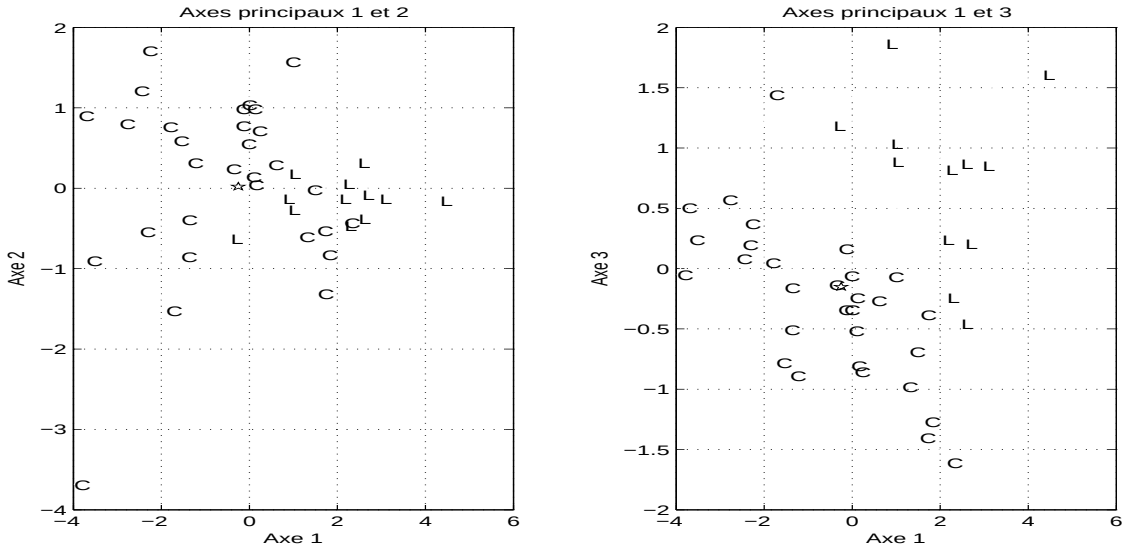


FIG. V.10 – *Projections des individus dans le plan principal 1–2 (gauche) et dans le plan 1–3 (droite) ; C=chien, L=loup, ⋆ =crâne de Jussac.*

Cartographie des caractères : cercles des corrélations

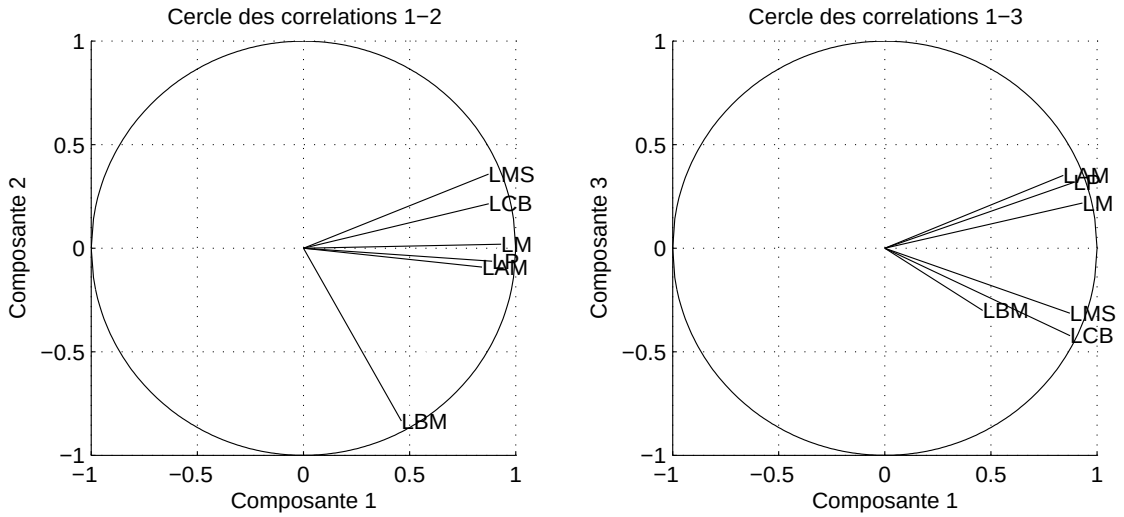


FIG. V.11 – *Cercles des corrélations 1–2 (gauche) et 1–3 (droite).*

1. Analyse en Composantes Principales

a. Calculer les quatre corrélations entre d'une part les deux caractères initiaux X_3 (LBM) et X_5 (LM) et d'autre part les deux premières composantes principales.

b. La cartographie des caractères dans le plan des deux premières composantes principales singularise l'un des caractères par rapport aux autres ; ce phénomène était-il prévisible avant d'avoir effectué le calcul des composantes principales ?

c. Certaines des trois premières composantes principales différencient-elles les chiens des loups ? Lesquelles ?

d. Le canidé de Jussac apparaît-il comme devant être classé plutôt parmi les chiens ou plutôt parmi les loups ?

On va s'intéresser désormais uniquement aux caractères LM (le mieux corrélé avec la première composante principale) et LBM (le mieux corrélé avec la seconde composante principale). On considère bien sûr que pour chaque caractère les 43 observations (30 chiens,12 loups et le canidé de Jussac) sont indépendantes.

2. Etude non paramétrique sur la caractère LM

On donne (en Annexe II, partie 1) la suite croissante des valeurs du caractère LM observées (sauf celle du canidé de Jussac), avec répétition en cas d'observations multiples, en indiquant à chaque fois s'il s'agit d'un chien ou d'un loup.

a. "Visuellement", peut-on considérer que ce caractère différencie bien les chiens et les loups ? Et que peut-on dire alors du canidé de Jussac ?

b. Effectuez un test de Mann-Whitney, au niveau de signification 0,05, de l'hypothèse nulle *"le caractère LM a la même loi pour les chiens et pour les loups"* contre l'hypothèse

alternative *"le caractère LM a tendance à prendre de plus grandes valeurs pour les loups que pour les chiens"*.
N.B. Vous direz vous-même s'il y a lieu ici d'utiliser l'approximation normale de la loi de Mann-Whitney.

3. Etude dans le modèle linéaire pour le caractère LM

a. On admet ici que, dans la population des chiens d'où est extrait l'échantillon de taille 30 observé, la loi du caractère LM est normale, d'espérance mathématique μ_C et de variance σ^2 inconnues et que de même, dans la population des loups d'où est extrait l'échantillon de taille 12 observé, la loi du caractère LM est normale, d'espérance mathématique μ_C et variance σ^2 (même variance pour les chiens et les loups). Effectuez dans ce modèle, au niveau de signification 0,05, le test de l'hypothèse nulle $\mu_C = \mu_L$ contre l'hypothèse alternative $\mu_C < \mu_L$.

b. On admet que le canidé de Jussac ait appartenu à une population dans laquelle le caractère LM suivait une loi normale, d'espérance mathématique μ_J et de même variance σ^2 que pour les populations actuelles de chiens et de loups. Effectuez dans ce modèle, au niveau de signification 0,05, le test de l'hypothèse nulle *"le canidé de Jussac était un chien"* (autrement dit $\mu_J = \mu_C$) contre l'hypothèse alternative $\mu_J \neq \mu_C$.
Attention : *Il s'agit d'un test de comparaison entre deux populations normales de même variance, sur la base de deux échantillons dont l'un est d'effectif 1 ; il n'est pas possible de faire une estimation de variance sur un échantillon de taille 1 et donc les formules classiques du test de Student ne s'appliquent pas ; vous vous assurerez qu'on est cependant bien dans le cadre des tests dans le modèle linéaire gaussien et achèverez l'étude.*

c. Reprendre cette étude pour tester cette fois l'hypothèse nulle *"le canidé de Jussac était un loup"*.

4. Les hypothèses de normalité pour le caractère LM se justifiaient-elles ?

On se demande si, en utilisant, pour le caractère LM, les lois normales comme on l'a fait à la question précédente, on n'aurait pas commis une erreur grossière ; on va donc effectuer des tests assez sommaires de normalité (faute de temps, on ne se préoccupera pas de tester, dans le cas où la normalité serit acceptée, l'égalité des variances relatives aux chiens et aux loups).

a. Pour la population des chiens, on va tester, au niveau de confiance 0,05, l'adéquation de la loi du caractère LM à l'ensemble de toutes les lois normales (à 2 paramètres réels, μ_C et σ_C, dont on rappelle que leurs estimations m_C et s_C se trouvent dans l'annexe I, partie 2) ; on effectue pour cela un test du χ^2 en utilisant les nombres d'observations figurant dans chacun des quatre intervalles de probabilité $\frac{1}{4}$ pour la loi normale d'espérance mathématique m_C et écart-type s_C (on rappelle que, si Φ désigne la fonction de répartition de la loi normale centrée réduite, on a $\Phi(-0,6745) = \frac{1}{4}$, $\Phi(0) = \frac{1}{2}$ et $\Phi(0,6745) = \frac{3}{4}$).

b. Pour la population des loups, un tel test du χ^2 est ici impossible (dites pourquoi). On va effecter, au niveu de confiance 0,05, un test de Kolmogorov d'adéquation de la loi du caractère LM à la loi normale d'espérance mathématique $15,8$ et écart-type $0,78$. A cet effet, notant $(x_{(1)}, \ldots, x_{(12)})$ la suite, ordonnée en croissant, des observations du caractère LM pour les loups, on fournit ci-dessous la liste des valeurs $\Phi(\frac{x_{(i)}-15,8}{0,78})$:

0,102	0,153	0,153	0,153	0,261	0,449
0,500	0,601	0,815	0,898	0,898	0,898

Quelques autres questions intéressantes (Considérer l'une ou l'autre de ces questions peut donner lieu à points suplémentaires)

a. Reprendre la question 2 pour le caractère LBM (utiliser l'annexe II, partie 2)

b. Si en question 4.b vous avez rejeté l'hypothèse de normalité, pouvez-vous avancer des arguments, relatifs au contexte expérimental (ou à la suite des observations sur les chiens) permettant une interprétation de cette circonstance ?

c. Au choix du candidat.

△

V.2 Corrections

Exercice V.1.

Dans cet énoncé, seules les statistiques descriptives élémentaires et la matrice de corrélation sont fournies (pas d'histogrammes ni de scatterplots). Les variables, toutes exprimées dans la même unité, sont assez homogènes, avec des moyennes s'échelonnant entre 358 et 1887 et des écart-types comparables.

Question 1. Etude de la matrice de corrélation

La matrice de corrélation laisse apparaître de forts coefficients (trois d'entre eux sont supérieurs à 90%) ce qui permet de penser qu'il y a une redondance entre les 7 variables du tableau, et qu'une ACP résumera bien les données en projetant convenablement le nuage sur peu d'axes.

Plus précisément on peut, à partir de la matrice de corrélation, construire des groupes de variables fortement corrélés. En notant $\rho(X, Y)$ la corrélation empirique entre X et X, on remarque que :

1. $\rho(Viande, Volaille) \approx 0.98\%$; $\rho(Viande, Fruit) \approx 0.96\%$; $\rho(Fruit, Volaille) \approx 0.93\%$; $\rho(Viande, Legume) \approx 0.88\%$
2. $\rho(Pain, Lait) \approx 0.86\%$
3. Le Vin n'est fortement corrélé avec aucun autre caractère.

Ceci suggère de considérer 3 groupes de variables : (Viande, Volaille, Fruit, Légume), puis (Pain, Lait) et enfin (Vin).

Ces groupes de variables constitués à partir de la matrice de corrélation permettent de penser que 3 caractères suffiraient à résumer convenablement le tableau. En effet, si on admet que les groupes constituent des variables redondantes (dans une certaine mesure), on peut choisir une variable dans chaque groupe, par exemple Viande, Pain, et Vin, et projeter le nuage sur ces 3 axes. On a ainsi réduit "à la main" la dimension de 7 à 3. L'ACP fait essentiellement le même travail, mais de manière optimale, en construisant des combinaisons linéaires de toutes les variables plutôt qu'en en éliminant certaines.

Remarque : ces regroupements visuels sont faciles à faire ici parce que l'exemple est de petite taille (7 variables) et que les corrélations sont très tranchées ; Il ne faut pas en conclure que l'ACP est inutile.

Question 2. ACP non normée

On a vu que les variables, exprimées dans la même unité (le Franc), étaient de plus homogènes (i.e. comparables). Ceci suggère de réaliser plutôt une ACP non normée, qui préservera les valeurs initiales (non réduites) de la table. Techniquement, ceci revient à diagonaliser la matrice de variances-covariances plutôt que la matrice de corrélation.

Questions 4 et 5 : interprétation de l'ACP

Choix du nombre d'axes

Le raisonnement sur la matrice de corrélation laissait penser que 3 variables (une par groupe de variables corrélées) résumeraient assez bien la table. Les % d'inertie cumulés, ainsi

que l'éboulis des valeurs propres, permettent de conclure que 2 ou 3 axes sont suffisants pour résumer respectivement 96.5% et 98.5% de l'inertie. On pourrait dans un premier temps se contenter de 2 axes, puis rajouter le troisième si l'interprétation réalisée sur le plan principal (axes 1 et 2) n'est pas complète ou pas assez satisfaisante .

Interprétation des cercles de corrélations

On constate sur le cercle (1–2) que le premier caractère est très fortement corrélé (de manière négative) avec les variables (Viande, Volaille, Fruit), et également très corrélé avec Légumes. On retrouve ici notre premier groupe constitué grâce à la matrice de corrélation. Le second caractère est, lui, très corrélé avec le second groupe (Pain, Lait), et pratiquement non corrélé avec (Viande, Volaille, Fruit).

Le Vin est faiblement corrélé avec ces 2 premiers nouveaux caractères. Ceci suggère de construire le cercle (1–3) sur lequel on voit que le troisième caractère peut être interprété comme l'axe de la consommation de Vin ; cet axe est de plus pratiquement non corrélé avec notre premier groupe de variables.

On peut interpréter le premier axe factoriel comme l'axe des "produits de consommation chers", par opposition au second axe factoriel qui peut être vu comme l'axe des produits "de consommation courante", et bon marchés.

Il faut aussi, pour l'interprétation des plans factoriels, garder présent à l'esprit le fait que le premier caractère est corrélé de manière négative avec le groupe des "produits chers". Ceci signifie que des individus situés très *à gauche* sur le premier axe factoriel (coordonnées négatives) sont de forts consommateurs de produits chers (plus que la moyenne des individus). De même, des individus situés très à droite sur cet axe sont de faibles consommateurs de produits chers (toujours relativement au barycentre, le "ménage moyen").

L'axe 2 s'interprétera, lui, conformément à l'intuition : des individus situés très *en haut* de l'axe sont de forts consommateurs de Pain et de Lait, et inversement.

Interprétation des plans factoriels

Sur cet exemple, les contributions ne sont pas fournies. Nous allons donc simplement interpréter les plans factoriels (le lecteur pourra les calculer en réalisant l'ACP avec le programme Scilab utilisé en TP et compléter le commentaire).

Sur le premier plan factoriel (plan principal, axes factoriels 1–2), les qualités de représentation des individus sont toutes raisonnables. On remarque d'abord une classification assez nette en les différentes catégories socio-professionnelles (CSP). On peut ainsi délimiter trois classes "convexes" représentant les groupes des CA, des EM et des MA (ceci signifie qu'aucun individu de l'un des groupes de CSP n'est "au milieu" d'un groupe d'une autre CSP).

On peut aussi remarquer que ces 3 groupes de CSP se répartissent le long du premier axe en, de gauche à droite, CA, puis EM, puis MA (la séparation en EM et MA étant moins nette). Ceci s'interprète comme le fait que les CA sont de plus gros consommateurs des produits qualifiés de "chers" (caractère 1), que les EM sont à peu près dans la moyenne et que les MA sont de faibles consommateurs de ces produits. Il n'y a pas une telle répartition le long de l'axe 2 : des représentants des 3 CSP sont présents aussi bien dans les grandes que les petites valeurs de l'axe 2.

On peut aussi s'intéresser à la répartition des nombres d'enfants par ménages, puisque cette information est aussi présente dans les "noms" des ménages. On remarque par exemple que les classes de CSP EM et MA sont ordonnées par nombre d'enfants croissants le long de l'axe 2 (ce n'est pas vrai pour les CA, bien que CA5 soit tout de même le plus en haut de l'axe 2 pour cette classe de CSP). Ceci s'interprète naturellement par le fait que les familles ayant plus d'enfants sont de plus gros consommateurs de produits "de base" tels que Pain et Lait (cf. cercle de corrélation 1–2), à la fois pour des raisons économiques (les MA sont d'ailleurs légèrement au-dessus des EM sur l'axe 2) et pour des raisons alimentaires (les enfants sont en principe plus consommateurs de Lait que les adultes).

Le plan factoriel (1–3), sur lequel apparaît l'axe associé à la consommation de Vin, ne permet pas de tirer de conclusions claires en terme de liens avec les CSP ou le nombre d'enfants.

▲

Exercice V.2.

Non fournie. ▲

Exercice V.3.

- A - Analyse en composantes principales

1. Les taux d'inertie sur les 2 premiers axes sont :

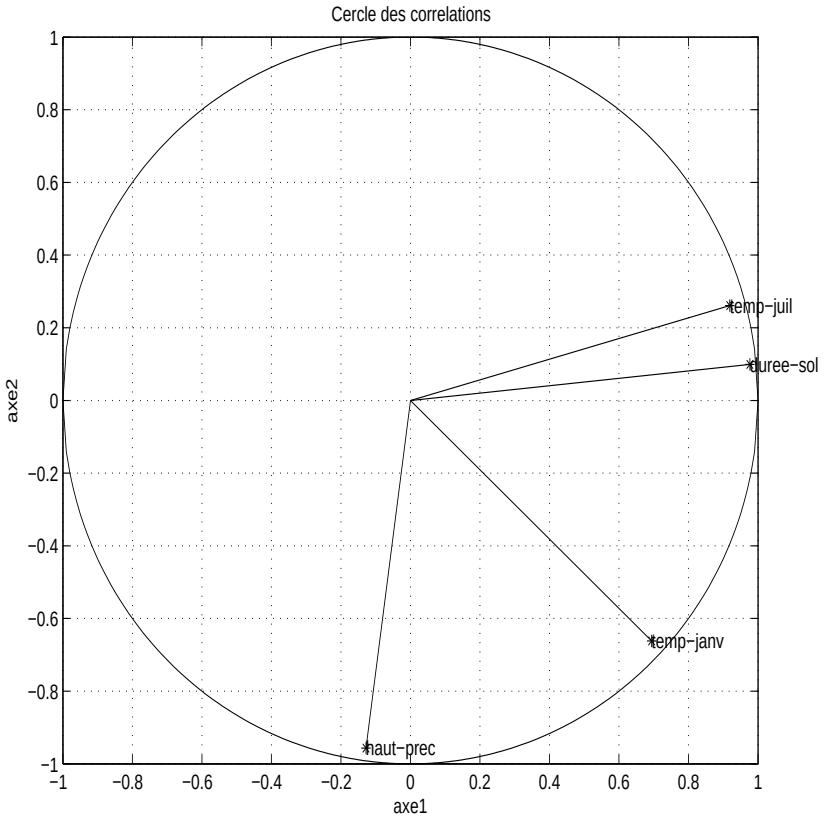
$$\tau_1 = \frac{2.30}{4} = 0.575 \text{ et } \tau_2 = \frac{1.43}{4} = 0.358$$

   D'où la part d'inertie expliquée par le plan principal vaut $\tau_1 + \tau_2 = 0.933$

2. Comme $corr(C_\alpha, X^j) = \sqrt{\lambda_\alpha} u_\alpha^j$, on peut calculer facilement à l'aide des valeurs propres λ_1 et λ_2 et des vecteurs propres u_1 et u_2, les corrélations. On obtient :

$$\begin{array}{ll} corr(C_1, X^1) = -0.12 & corr(C_2, X^1) = -0.96 \\ corr(C_1, X^2) = 0.97 & corr(C_2, X^2) = 0.095 \\ corr(C_1, X^3) = 0.70 & corr(C_2, X^3) = -0.66 \\ corr(C_1, X^4) = 0.93 & corr(C_2, X^4) = 0.26 \end{array}$$

3. L'inertie sur les 2 premiers axes représente 93.3% de l'inertie totale. On a donc une information suffisante pour analyser les données.
   L'axe C_1 est fortement corrélé avec l'ensoleillement et la température en Juillet : on sépare le long de cet axe des villes dont l'ensoleillement et la température en Juillet sont les plus faibles (à gauche) et celles qui sont les plus chaudes en Juillet et les plus ensoleillées (à droite). En particulier on distingue bien les trois villes méditerranéennes détachées sur la droite (Nice, Marseille, Perpignan).
   Sur l'axe C_2, c'est la hauteur des précipitations qui est prépondérante avec une corrélation fortement négative : les villes les plus pluvieuses sont vers le bas et les villes les moins pluvieuses vers le haut. On peut notamment remarquer vers le bas deux villes atlantiques très pluvieuses : Brest et Biarritz.


Cercle des correlations
temp−juil
duree−sol
temp−janv
haut−prec
axe1
axe2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

4. La contribution de Biarritz sur les 2 premiers axes est :

$$CTR_1(Biarritz) = \frac{0.1\ (0.02)^2}{2.3} = 0.00 \text{ et } CTR_2(Biarritz) = \frac{0.1\ (-2.63)^2}{1.43} = 0.484$$

A l'aide du tableau des données on peut calculer le carré de la distance entre Biarritz et le centre de gravité :

$$\begin{aligned} d(Biarritz, G)^2 &= \left(\frac{1474-801}{285}\right)^2 + \left(\frac{1921-2101}{442}\right)^2 + \left(\frac{7.6-4.5}{2.51}\right)^2 + \left(\frac{19.7-20.0}{2.51}\right)^2 \\ &= 7.27 \end{aligned}$$

On en déduit la qualité de la projection :

$$CO2_1(Biarritz) = \frac{0.02^2}{7.27} = 0.00 \text{ et } CO2_2(Biarritz) = \frac{-2.63^2}{7.27} = 0.95$$

On peut remarquer que l'axe 2 contient 95% de l'information sur Biarritz et que cette ville contribue à près de la moitié de l'inertie de cet axe.

5. Pour l'axe 1 les contributions des villes méditerranéennes est significatives (62.4% de l'inertie de l'axe pour ces 3 villes) et on peut noter que 94% de l'information sur Nice et Perpignan est contenue dans l'axe 1. Pour l'axe 2, outre Biarritz, c'est Brest dont la contribution à l'inertie de cet axe est la plus importante (22.7% de l'inertie de l'axe). D'autre part ces deux axes contiennent plus de 90% de l'information sur chacune des villes sauf Paris (87.5%) et surtout Lyon (seulement 38.6% de l'information) qui est la seule ville mieux représentée par les axes 3 et 4, que par les axes 1 et 2.

- B - Classification

1. Le déroulement de la classification ascendante hiérarchique est résumé dans le tableau ci-dessous.

Itération	Classe formée				Distance
	Nom	regroupant		Effectif	
1	A_1	Clermont	Paris	2	0.43
2	A_2	A_1	Lille	3	0.97
3	A_3	Nice	Perpignan	2	1.03
4	A_4	A_3	Marseille	3	1.07
5	A_5	A_2	Lyon	4	1.10
6	A_6	A_5	Strasbourg	5	1.13
7	A_7	Biarritz	Brest	2	2.10
8	A_8	A_6	A_7	7	2.50
9	A_9	A_4	A_8	10	2.74

2. Le regroupement en trois classes est obtenu à l'itération 7. Les trois classes sont :

A_5 ={Marseille, Perpignan, Nice }

A_6 ={Clermont, Lille, Lyon, Paris, Strasbourg }
A_7 ={Biarritz, Brest }

▲

Exercice V.4.
Non fournie.

▲