Statistiques

Cours de Licence

Bernard Delyon

2 octobre 2018

1. IRMAR, Université Rennes-I, Campus de Beaulieu, 35042 Rennes cedex, France ; bernard.delyon@univ-rennes1.fr

Table des matières

I

Introduction

I.1 Statistiques et probabilités

Le point de départ de la statistique est l'échantillon (ou les données, ou les observations...) ; c'est un tableau de chiffres ou de symboles (p. ex. « âge », « revenus » et « santé » de 1000 personnes). Le but essentiel de la statistique est de mesurer les dépendances entre variables , voire leur indépendance, ainsi que la variabilité : elle cherche à démêler ce qui est *tendances systématiques* (p. ex. le revenu augmente avec l'âge) des *variations aléatoires* (fluctuations).

La statistique va donc dans la sens contraire des probabilités :

Probabilités : inférer des propriétés des variables aléatoires à partir de la connaissance de leur distribution. Essentiellement des théorèmes-limite (lois des grands nombres, théorème-limite central...).

Statistiques : inférer de l'information sur des distributions à partir de l'observation de variables aléatoires.

Ex : on suppose que chaque électeur a une certaine probabilité p_i de voter pour le candidat i ; p_i représente donc la proportion de votants pour ce candidat et la connaissance de tous les p_i permettrait de connaître à l'avance le résultat des élections. On dispose d'un échantillon de 1000 électeurs et de leur vote, peut-on en déduire les p_i ?

I.2 Les branches des statistiques

Statistique descriptive (ou exploratoire) : organisation des données, représentation, et extraction de caractéristiques géométriques. Une approche géométrique avec peu de probabilités (Études des nuages de points, classification...).

Ex : Pour chaque pays, on considère le revenu moyen, l'âge moyen, le nombre de médecins par habitant, le nombre de livres par habitant, et le PIB par habitant. Ceci fait un point de $\mathbb{R}^5$ par pays. La structure du nuage de points ainsi obtenu (présence de groupes...) peut être une source d'informations intéressantes.

Planification d'expériences : comment générer les variables de sorte à inférer au mieux la distribution.

Ex : choix d'échantillons « représentatifs » pour des sondages, mise en œuvre d'expériences avec placebos (utilisation du double aveugle, randomisation des échantillons).

Estimation et test : inférer certains paramètres d'une distribution (moyenne,...) à partir de v.a. tirées selon celle-ci et de certaines hypothèses (restriction de la distribution à une certaine famille....). Étudier les propriétés de ces estimateurs. Mise au point de tests.

Ex : Étude de la dépendance entre certaines variables (p. ex. « âge », « sexe », et « revenus »), exploitation de sondages, reconnaissance des formes...

Ex : En 1854 une épidémie de choléra sévissait sur une partie de Londres. John Snow par une étude attentive de la distribution spatiale des cas de choléra put détecter que le centre de la transmission était une pompe à eau particulière (Broad Street pump). La pompe fut remplacée [1].

1. Cette version, communément répandue, est en réalité très enjolivée, voir l'article :
H. Brody, M. Russell Ripe, b, Peter Vinten-Johansenc, N. Panethb, S. Rachmand, "Map-making and myth-making in Broad Street : the London cholera epidemic, 1854", *The Lancet*, Volume 356, Issue 9223 , 1 July 2000, Pages 64-68.

II

Statistique exploratoire univariée et bivariée

II.1 Rappels sur les variables aléatoires numériques

On rappelle quelques notions de base en se concentrant sur le cas purement continu avec densité et le cas purement discret.

II.1.1 Poids et densité

Considérons une variable aléatoire $X(\omega)$ définie sur un espace Ω.

Si cette variable est discrète et prend les valeurs sur $\{v_1, ..v_J\}$, sa répartition est déterminée par l'ensemble des $p_j = P(X = v_j)$. On a $\sum_j p_j = 1$.

Si elle admet une densité $p(x)$, la probabilité qu'elle tombe dans $[a, b]$ vaut l'intégrale de cette densité entre les deux points a et b, cf figure II.1. Pour qu'une fonction $p(x)$ soit une densité, il faut et il suffit qu'elle soit positive et d'intégrale un. La figure II.1 donne un exemple concret de distribution dissymétrique.

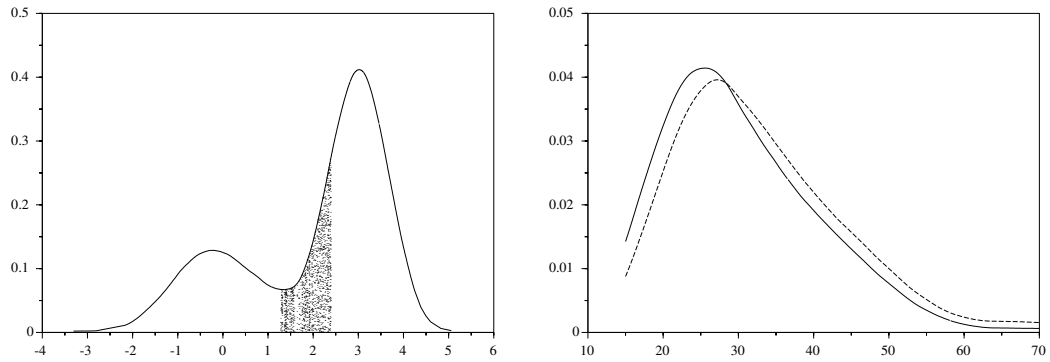


Figure II.1 – Un exemple de densité : la probabilité de l'intervalle $[1,3\ ;\ 2,4]$ est la surface de la zone noircie ; la surface totale (intégrale) fait 1. A droite : Densités des variables « âge du (de la) marié(e) le jour du mariage » en Alaska en 1995 (estimation à partir des données de l'Alaska Bureau of Vital Statistics). Les femmes sont en trait plein et les hommes en pointillés .

Rappelons la loi des grands nombres qui est un des résultats les plus importants de la théorie des probabilités

Théorème 1. *Soit X_n une suite de variables indépendantes de même loi et d'espérance finie, alors, avec probabilité 1 :*

$$\lim_{n\to\infty} \frac{X_1(\omega) + X_2(\omega) + ...X_n(\omega)}{n} = E[X_1].$$

La variable gaussienne

La densité de la variable gaussienne de moyenne m et variance σ^2 est $p(x) = (\sigma\sqrt{2\pi})^{-1} e^{-(x-m)^2/2\sigma^2}$; elle a un pic unique centré sur sa moyenne m et dont la largeur est approximativement 4σ.

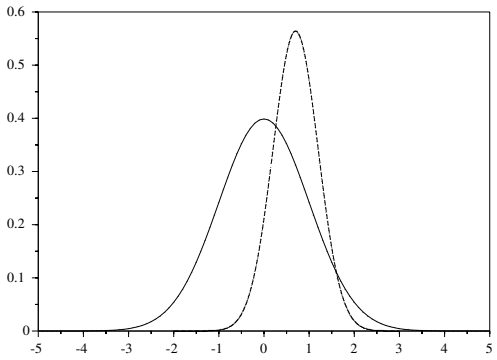


FIGURE II.2 – Densité de la gaussienne centrée réduite, et de la gaussienne de moyenne 0,7 et de variance 1/4.

La loi gaussienne joue un rôle particulier en raison de la propriété fondamentale suivante :

Théorème 2. *Toute somme finie de variables aléatoires gaussiennes indépendantes suit une loi gaussienne, les moyennes et les variances s'ajoutant.*

mais aussi en raison du théorème-limite central :

Théorème 3. *Soit X_n une suite de variables indépendantes de même loi, centrées, de variance σ^2, alors les variables*

$$Y_n(\omega) = \frac{X_1(\omega) + X_2(\omega) + ...X_n(\omega)}{\sqrt{n}}$$

convergent en loi vers la variable normale centrée de variance σ^2.

On a donc pour toute fonction f bornée : $E[f(Y_n)] \longrightarrow \int f(x) e^{-\frac{x^2}{2\sigma^2}} \frac{dx}{\sqrt{2\pi}\sigma}$ (il suffit en réalité que $f(x)/(1+x^2)$ soit borné).

Mentionnons que ce résultat reste vrai sous des hypothèses plus faibles, par exemple en remplaçant l'hypothèse d'identité des lois par $\sup_i E[|X_i|^3] < +\infty$. Un théorème analogue existe également pour le cas où les variances sont distinctes. Ceci explique pourquoi la distribution gaussienne se rencontre souvent dans la nature, dès qu'un phénomène observé est la somme d'un grand nombre de facteurs indépendants (bruit thermique, prix d'un produit...).

II.1.2 Fonction de répartition

C'est la fonction

$$F(y) = P(X \leqslant y)$$

qui n'est donc définie que pour les variables scalaires. Elle possède les propriétés suivantes :

- ► $F(y) \in [0,1], \quad y \in \mathbb{R}$
- ► F est croissante et continue à droite.

Pour tout intervalle $]a,b]$, on a donc $P(]a,b]) = F(b) - F(a)$. Si une variable aléatoire est discrète, sa fonction de répartition est constante par morceaux et saute de p_i en y_i, cf. la figure II.3.

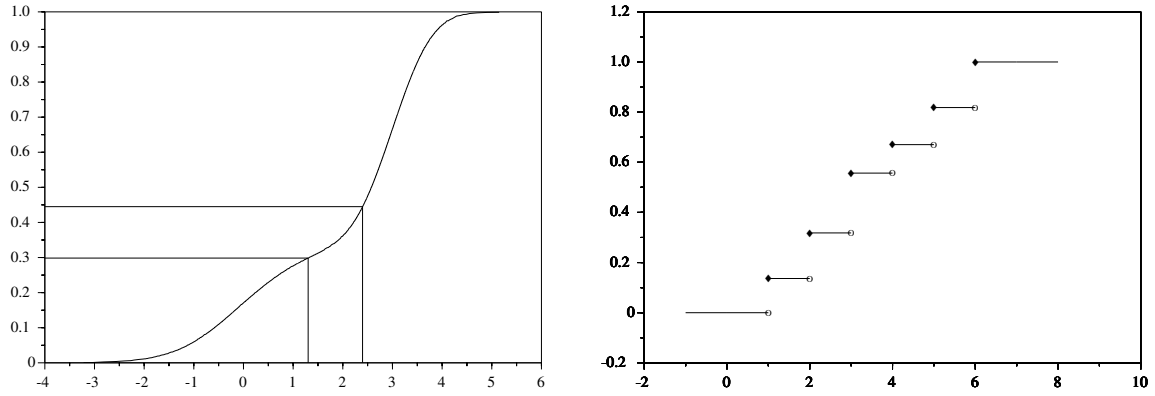


FIGURE II.3 – Fonction de répartition correspondant à la densité de la figure II.1. On voit que la probabilité de l'intervalle $[1,3\ ;\ 2,4]$ vaut environ 0,15. Fonction de répartition d'une variable prenant les valeurs $\{1,2,3,4,5,6\}$: $P(X=4) \simeq 0,1$

Dans le cas discret, si l'on suppose que les y_i sont ordonnés par ordre croissant, on a (figure de droite)

- $F(y) = 0$ si $y \leqslant y_1$, et $F(y) = 1$ si $y > y_n$
- $F(y) = p_1 + ... + p_i$ si $y_i \leqslant y < y_{i+1}$.

Si X admet une densité $p(x)$ on a

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

La dérivée de F est donc p.

II.1.3 Simulation des variables à partir d'une variable uniforme

Variables continues

À partir d'une suite de variables uniformes sur $[0,1]$, $U_1(\omega)$,... $U_n(\omega)$, produites par exemple sur ordinateur, il est facile de produire des variables dont la fonction de répartition est une fonction $F(x)$ donnée.

Supposons d'abord F strictement croissante, continue et donc inversible, définissons alors la fonction quantile Q comme la fonction inverse de F ; alors si l'on pose

$$X_k(\omega) = Q(U_k(\omega))$$

on a

$$P(X_k \leqslant y) = P(Q(U_k) \leqslant y) = P(U_k \leqslant F(y)) = F(y)$$

ce qui prouve que les variables X_k ont bien la distribution voulue. On peut montrer que cette propriété reste vraie, même pour les variables numériques discrètes, en définissant en toute généralité la fonction quantile par :

$$Q(u) = \min\{y : F(y) \geqslant u\}.$$

Variables discrètes

Soit à simuler la distribution donnant la probabilité p_i à la valeur y_i, $i = 1, ..n$.

Découpons l'intervalle $[0, 1]$ en intervalles successifs $I_1, I_2, ...I_n$ de longueur $p_1, p_2, ...p_n$. Comme la somme des p_i fait 1, on recouvre ainsi exactement $[0, 1]$. Alors la variable définie par

$$X_k(\omega) = y_i, \quad \text{pour le } i \text{ tel que } \ U_k(\omega) \in I_i$$

a bien la loi désirée. En effet, pour tout i

$$P(X_k = y_i) = P(U_k \in I_i) = p_i$$

car U_k est uniforme.

II.1.4 Caractéristiques numériques

L'espérance d'une fonction réelle f de X est définie comme (selon le cas)

$$E[f(X)] = \int_{-\infty}^{+\infty} f(y)p(y)\ dy$$
$$E[f(X)] = \sum_i f(y_i)p_i$$

(sous réserve d'existence de l'intégrale, ce qui est le cas si f est bornée) et la probabilité d'un évènement A [1]

$$P(X \in A) = E[1_A(X)] = \int_A p(y)\ dy$$
$$P(X \in A) = E[1_A(X)] = \sum_{y_i \in A} p_i.$$

On a l'inégalité de Cauchy-Schwartz

$$E[f(X)g(X)] \leqslant E[f(X)^2]^{1/2} E[g(X)^2]^{1/2}.$$

avec égalité seulement si les variables $f(X)$ et $g(X)$ sont proportionnelles (il existe $a \in \mathbb{R}$ tel que $f(X) = ag(X)$ ou $g(X)$ est identiquement nul).

L'espérance (ou la moyenne) d'une variable numérique X est $E[X]$ et sa **variance** est $Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$. Son **écart type est** $\sigma_X = \sqrt{Var(X)}$.

Propriétés de l'espérance. Soit un réel a :

1. On note $1_{x=a}$ ou $1_a(x)$ la fonction de x qui vaut 1 en a et 0 ailleurs. De même $1_{x \in A} = 1_A(x)$ vaut 1 pour $x \in A$ et 0 ailleurs. On parle de fonction caractéristique d'ensemble

- $\sigma_X \geqslant 0$, et $\sigma_X = 0$ ssi X est constante (d'où le terme de variance)
- multiplicativité : $\sigma_{aX} = |a|\sigma_X$
- invariance par translation : $\sigma_{X+a} = \sigma_X$
- Si X_i est une suite de v.a. indépendantes : $Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n Var(X_i)$

Par analogie avec la gaussienne, on considère que le domaine de variation typique d'une variable aléatoire continue est l'intervalle centré en $E[X]$ et de largeur $4\sigma_X$. Ceci est bien sûr très faux si X a une distribution éloignée de la gaussienne.

Un mode de X, variable continue, est un point où la densité $p(y)$ est maximum. Deux modes apparaissent donc sur la figure II.1 à gauche.

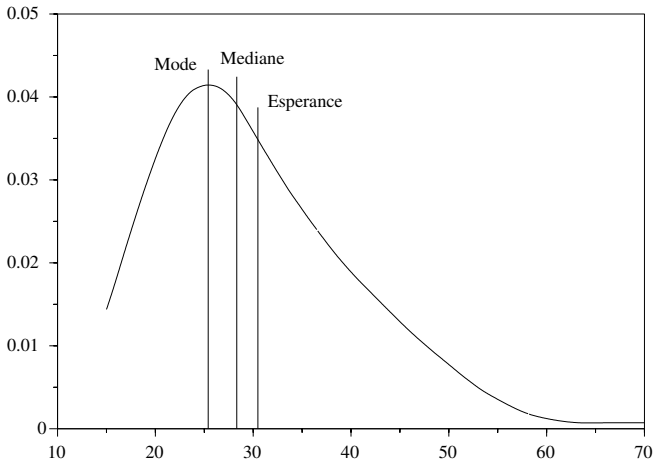
La médiane d'une variable numérique X est le point où F vaut $1/2$.

Pour une variable discrète, cette définition est insuffisante car F est constante par morceaux : Si F ne prend pas la valeur $1/2$, c'est le point où F traverse cette valeur (p.ex. 3 dans la figure II.3). Sinon F vaut $1/2$ sur un plateau $[a, b]$ et la médiane est par convention $(a + b)/2$ (tout point de $]a, b[$ pourrait être choisi car il sépare les valeurs en deux ensembles de probabilité $1/2$).

Pour une variable possédant une densité, si la médiane est m, on a

$$P(X < m) = P(X > m) = \frac{1}{2}$$

Exemple. On voit sur l'exemple suivant (l'âge du marié, cf figure II.1) l'effet typique de la dissymétrie de la densité sur les trois indices de localisation.



II.1.5 Mesures de corrélation

Définition 4. *La covariance de X et Y est*

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

Si $Cov(X, Y) = 0$, on dit que X et Y sont décorrélées.

Si la paire X, Y admet une densité $p(x, y)$, ce coefficient vaut donc

$$Cov(X, Y) = \int\int xyp(x, y)\ dx\ dy - \int\int xp(x, y)\ dx\ dy \int\int yp(x, y)\ dx\ dy$$

Si elle prennent des valeurs discrètes (x_i, y_i) avec probabilité p_i, on a

$$Cov(X,Y) = \sum_i x_i y_i p_i - \sum x_i p_i \sum y_i p_i$$

Propriété 5. *Pour tous réels a, b, on a*
- $Cov(aX+b, Y) = aCov(X,Y)$.
- $Cov(X+Y, Z) = Cov(Z, X+Y) = Cov(X,Z) + Cov(Y,Z)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Propriété 6. *La covariance de deux variables aléatoires indépendantes est nulle. L'inverse n'est pas vrai en général.*

En dehors du cas gaussien, l'inverse est vrai dans la situation suivante :

Propriété 7. *Soient U et V deux variables aléatoire à valeurs 0 ou 1, si leur covariance est nulle, alors elles sont indépendantes. Ceci revient à dire que les évènements $\{\omega : U = 1\}$ et $\{\omega : V = 1\}$ sont indépendants.*

Démonstration. Tout va venir de ce que toute fonction définie sur $\{0,1\}$ coïncide avec une fonction linéaire. Soient deux fonction réelles f et g. Remarquons que $f(U) = aU + b$, avec $b = f(0)$ et $a = f(1) - f(0)$, et de même pour V, d'où $Cov(f(U), g(V)) = 0$ et donc

$$E[f(U)g(V)] = E[f(U)]E[g(V)].$$

La dernière affirmation est laissée en exercice. ∎

Définition 8. *Le coefficient de corrélation linéaire entre X et Y non-constantes est*

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Propriété 9. *$|Cor(X,Y)| \leqslant 1$. $|Cor(X,Y)| = 1$ si et seulement s'il existe $(a,b) \in \mathbb{R}_* \times \mathbb{R}$ tels que $Y = aX + b$; dans ce cas $Cor(X,Y) = signe(a)$.*

Démonstration. Noter qu'on a l'inégalité de Cauchy-Schwarz $Cov(X,Y)^2 \leqslant Var(X)Var(Y)$. Si $|Cor(X,Y)| = 1$, on est dans le cas d'égalité dans l'inégalité de Cauchy-Schwarz et donc $X - E[X]$ et $Y - E[Y]$ sont proportionnels ; étant tous deux non-nuls, on en conclut l'existence de a tel que $Y - E[Y] = a(X - E[X])$, ce qui implique bien $Y = aX + b$, avec $b = E[Y] - aE[X]$.

Si réciproquement $Y = aX + b$, alors $E[Y] = aE[X]$ donc par différence $Y - E[Y] = a(X - E[X])$, et l'on vérifie immédiatement que $Cor(X,Y) = \text{signe}(a)$. ∎

Cas de deux Bernoulli. Soient U et V deux Bernoulli corrélés dont la loi liée est donnée par la matrice 2×2, $P(U = i, V = j) = p_{ij}$. On note $p_{i.} = P(U = i) = p_{i0} + p_{i1}$, et de même $p_{.j} = P(V = j)$. On a alors

$$Cov(U,V) = p_{11} - p_{.1}p_{1.}, \quad Cor(U,V) = \frac{p_{11} - p_{.1}p_{1.}}{\sqrt{p_{0.}p_{1.}p_{.0}p_{.1}}}. \tag{II.1}$$

Exemples : On peut raisonnablement supposer que la corrélation entre l'âge et le revenu est positive, celle entre l'âge et l'état de santé est négative, et celle entre le jour de naissance (entre 1 et 365) et l'état de santé est nulle.

II.2 Les données en première analyse

II.2.1 Introduction

En statistique *univariée*, les données sont constituées d'un tableau $(X_i)_{1\leqslant i\leqslant n}$, ensemble de valeurs qui peuvent être soit

- quantitatives continues. Ex : taille d'un individu.
- quantitatives discrètes. Ex : sortie d'un coup de dé.
- qualitatives ou catégorielles. Ex : P ou F, sortie d'un jeu de pile ou face.

Les données quantitatives sont numériques. Les valeurs prises par une variable catégorielles sont appelées les *modalités*.

La statistique *bivariée* se consacre à l'étude d'un tableau de données contenant deux variables, par exemple l'âge des mariés en Alaska en 1995. Ce tableau a donc deux colonnes, une par variable (âge de l'homme et âge de la femme), et un nombre arbitraire de lignes (une par individu).

L'idée qui guide les méthodes de statistique exploratoire est l'interprétation de ces données comme des réalisations indépendantes d'une variable aléatoire de loi inconnue.

II.2.2 Tableaux et tables de contingence

La représentation la plus simple est la suite exhaustive des X_i.

Pour une variable catégorielle, il y a souvent peu de valeurs possibles vis-à-vis du nombre de données, et l'on peut représenter les données par un double tableau valeurs/nombre d'occurrences : La table II.1 permet de manipuler deux tableaux de longueur fixe 6 au lieu d'un grand tableau dont la longueur est celle de l'échantillon (ici 88). De même le tableau II.1 représente une paire de variables à deux modalités.

v_j	1	2	3	4	5	6
n_j	12	16	21	10	13	16

	Hommes	Femmes
fumeurs	72	323
non-fumeurs	56	233

TABLE II.1 – Résultats de 88 jets de dés. Table de contingence : Évaluation du tabagisme sur 684 individus classés par sexe.

Dans le tableau II.2, plutôt que de lister l'ensemble des 5514 les paires (âge du marié, âge de la mariée), on a classifié ces paires en 81 classes (discrétisation) ce qui donne le tableau suivant :

II.2.3 Histogrammes (variables quatitatives réelles)

Les tableaux ne sont pas très parlants, surtout s'ils sont grands. La représentation par histogramme pour les observations issues d'une variable continue, permet de visualiser aisément la distribution des données (distribution empirique).

Une méthode naturelle consiste à les discrétiser sur des intervalles et à tracer un **histogramme par effectifs**, également appelé « diagramme bâtons », où les ordonnées figurent le nombre n_i (ou la proportion p_i en %) de points observés dans la classe, fig.II.4.

On voit tout de suite le défaut de cette méthode : le résultat ne correspond pas à ce que l'on attend si les intervalles sont inégaux. Pour pallier à cela on va diviser la hauteur par la largeur de l'intervalle, et pour une raison qui va apparaître, on va également diviser par le nombre total de points ; la hauteur au dessus du i-ième intervalle sera donc $h_i = n_i/(nl_i)$ où l_i est la largeur de l'intervalle. On obtient alors la figure II.5.

F\ H	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55+	TOTAL
15-19	152	323	68	20	4	3	1			571
20-24	56	733	452	146	42	12	7	4	1	1453
25-29	3	157	417	312	136	49	31	7	1	1113
30-34	2	34	141	273	194	107	41	8	11	811
35-39		10	55	116	180	157	70	32	13	633
40-44		4	11	52	80	118	88	40	25	418
45-49			4	18	36	41	79	58	41	277
50-54			1	4	9	16	28	35	48	141
55+							7	8	82	97
TOTAL	213	1261	1149	941	681	503	352	192	222	5514

TABLE II.2 – L'âge des mariés en Alaska en 1995 (Alaska Bureau of Vital Statistics). L'âge du marié varie en ligne et celui de la mariée en colonne. Chaque case indique un nombre de mariés.

Dans la représentation de la figure II.5, noter que l'ordonnée h_i ne correspond pas à la probabilité empirique de la classe, mais c'est la surface ; on a donc $h_i l_i = p_i = n_i/n$. La surface totale fait donc 1. L'HISTOGRAMME FOURNIT À LA FOIS UN ESTIMATEUR DE LA DENSITÉ DE PROBABILITÉ ET UNE DESCRIPTION DES DONNÉES.

On voit deux modes se dessiner, laissant penser à deux types d'activité différents.

Effet de la fusion ou de la scission de classes. Si l'on fusionne deux classes, le bloc résultant aura une hauteur *moyenne* entre les deux blocs initiaux (et sa surface est la somme des surfaces). L'histogramme garde donc le même aspect, figure II.6. Si les classes sont trop petites vis-à-vis du nombre d'échantillons, l'histogramme obtenu peut être assez mauvais car illusoirement précis, figure II.6.

II.2.4 Digression : un estimateur de la densité.

Pour une variable continue, les représentations du § II.2.3 donnent une estimation de la densité de la variable, i.e. de la dérivée de $F(x)$. Plutôt que d'utiliser un histogramme (figure II.5), qui donne une estimée très irrégulière, il est courant d'estimer la densité par une formule du type

$$p_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

où K est une fonction positive d'intégrale 1 (de sorte que p_n est aussi d'intégrale 1) et h un réel, bien choisis. Si K est l'indicateur de $[-1/2, 1/2]$, $p_n(x)$ n'est autre que la proportion d'échantillons observés dans un voisinage de taille h de x. Cette formule peut s'interpréter comme une façon particulière de régulariser l'histogramme. Le choix de K et du réel h est un domaine difficile des statistiques. Pour K, un bon choix est le noyau d'Epanechnikov

$$K(x) = \frac{3}{4\sqrt{5}} \left(1 - \frac{x^2}{5}\right), \qquad |x| < \sqrt{5}.$$

Quant à h, il doit être assez petit pour n grand : $h \simeq s n^{-1/5}$ est préconisé pour ce choix de K (s est l'écart-type empirique défini plus bas) ; toutefois en pratique, h est souvent choisi à la

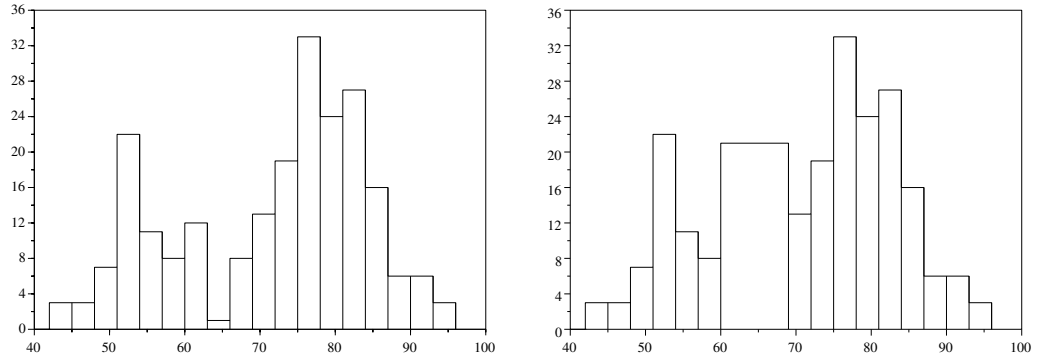


FIGURE II.4 – Histogramme par effectif des intervalles de temps entre deux éruptions du geyser Old Faithful du parc de Yellowstone. Un total de 222 mesures ont été prises. Il y a 22 mesures entre 51 et 54 minutes. Dans le second histogramme, on a fusionné 3 classes de sorte à mettre toutes les données de l'intervalle $[60, 69]$ dans une seule classe.

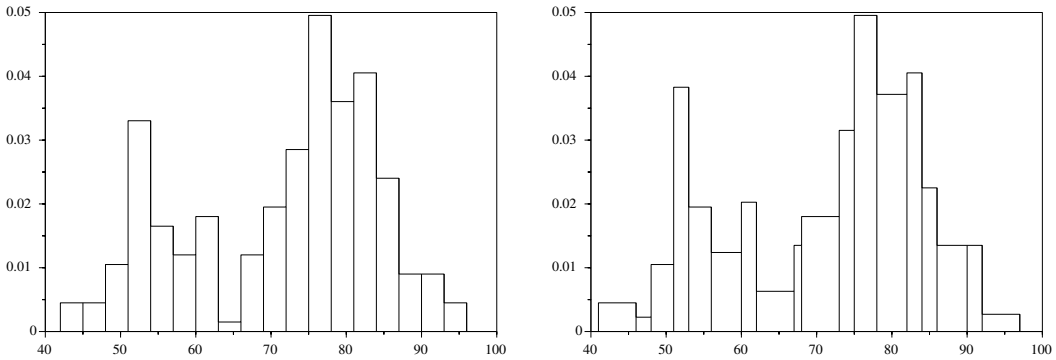


FIGURE II.5 – Histogramme des données « geyser ». Elles ont été discrétisées par intervalles de 3 minutes. L'intégrale fait 1. On a observé $0,01.3.222 = 7$ valeurs entre 48 et 51. On a placé à côté un histogramme avec des classes de taille variable.

main à une valeur qui semble raisonnable. La figure suivante est l'estimation de densité sur les données « geyser » ; on a pris $h = sn^{-1/5} = 4,33$:

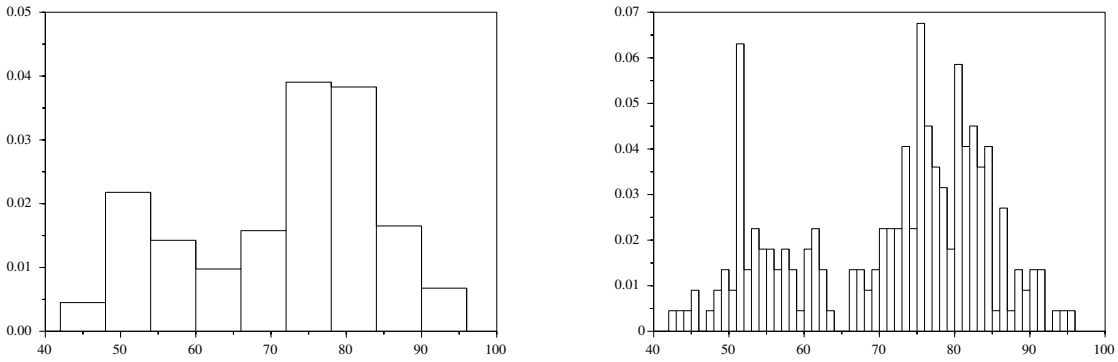
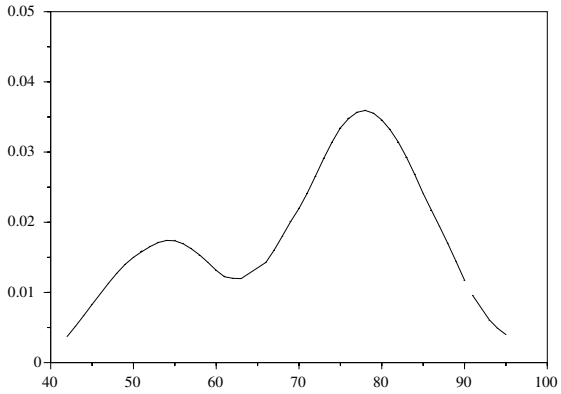


FIGURE II.6 – Histogramme des données « geyser », discrétisées cette fois par intervalles de 6 mn ; on distingue toujours deux modes. Histogramme avec une discrétisation par intervalles d'1 mn ; les blocs de hauteur 0.005 correspondent à une seule donnée.



II.3 La distribution empirique

Comme nous l'avons dit plus haut, tout ce qui suit est guidé par le modèle qui fait des données $(X_i)_{1\leqslant i\leqslant n}$ une suite d'observations indépendantes de même loi. L'hypothèse clef est donc l'**homogénéité** des données (identité des loi).

II.3.1 Distribution et moyennes empiriques

La moyenne empirique d'une fonction f des données $(X_i)_{1\leqslant i\leqslant n}$ est

$$\frac{1}{n}\sum_i f(X_i)$$

et la *probabilité empirique* d'un ensemble A est la proportion d'échantillons tombés dans A :

$$\frac{1}{n}\sum 1_A(X_i).$$

La *distribution empirique* est donc celle qui attribue à chaque valeur une probabilité égale à sa fréquence d'observation. C'est celle que l'on observe en tirant (avec remise) des échantillons au hasard dans l'ensemble des observations.

EXEMPLE : dans le cas des 88 jets de dés (table II.1) la loi empirique a les poids suivants

j	1	2	3	4	5	6
p_j	12/88	16/88	21/88	10/88	13/88	16/88

Dans le cas du tableau II.1, en normalisant simplement par le nombre d'individus, on obtient le tableau :

	H	F
fumeur	0,105	0,472
non-fumeur	0,082	0,341

La loi des grands nombres assure que si les variables $(X_i)_{i=1,2...}$ sont indépendantes de même loi, alors, pour toute fonction f continue par morceaux bornée (et même pour bien d'autres...), leur moyenne empiriques convergent vers l'espérance de f sous cette loi, $E[f(X)]$.

Sous des conditions raisonnables, cette propriété reste vraie pour des variables non-indépendantes, l'hypothèse essentielle restant l'identité des lois.

Les moyennes empiriques sont donc l'approximation la plus naturelle des espérances à partir des données.

Cas de données discrétisées en intervalles. Il arrive que les variables originales soient discrétisées sur des intervalles, par exemple si l'âge du marié est donné par tranche d'âges (tableau II.2) ; dans ce cas on ne peut pas retrouver la loi empirique des données originales. On prend alors conventionnellement comme loi empirique soit la loi *discrète* qui attribue à chaque milieu d'intervalle la probabilité de ce dernier (on fait comme si tous les mariés de l'intervalle 20-24 avaient 22 ans), soit la loi *uniforme par morceaux* dont la densité est donnée par l'histogramme correspondant (on fait comme si les mariés de l'intervalle $[20, 24]$ sont uniformément répartis). Cette dernière solution est plus naturelle car la vraie loi a une densité mais rend le calcul des espérances empiriques plus difficiles. Ces deux choix donnent des résultats différents en général sauf pour le calcul de la moyenne. Par exemple, si au lieu des données complètes, on ne dispose que du tableau II.2, on estimera l'âge moyen du marié par :

$$\frac{17 \times 213 + 22 \times 1261 + ... + 52 \times 192 + 57 \times 222}{5514} = 32 \text{ ans et } 3 \text{ mois.}$$

La valeur 57 choisie ici est arbitraire et discutable.

II.3.2 Fonction de répartition

La fonction de répartition empirique de l'échantillon est

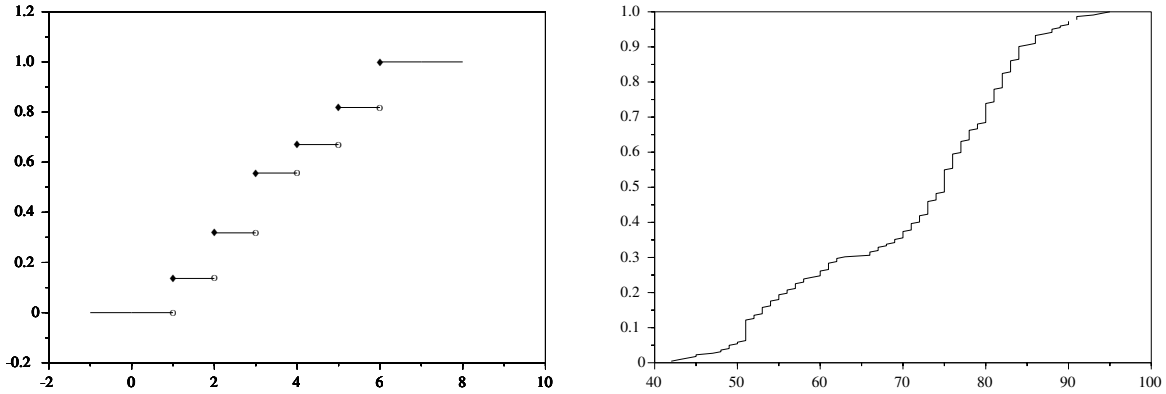
$$F_n(y) = \frac{1}{n} \sum_i 1_{x_i \leqslant y} = \frac{\text{nb de valeurs observées} \leqslant y}{\text{nb total de valeurs observées}}. \tag{II.2}$$

$F_n(y)$ est la fréquence empirique d'apparition de valeurs strictement inférieures à y. La fonction de répartition n'est donc définie que pour les variables prenant des valeurs numériques.

La loi des grands nombres implique que si les variables $(X_i)_{i=1,2...}$ sont de même loi P, alors pour tout y,

$$\lim_n F_n(y) = F(y) = P(X \leqslant y).$$

Si la fonction de répartition empirique est moins parlante que l'histogramme, son avantage est que la formule (II.2) permet de représenter directement une variable continue sans faire de discrétisation préalable en intervalles (de taille arbitraire) ; voici les fonctions de répartition correspondant à la table II.1 et des données « geyser » :



II.3.3 Quantiles

La fonction quantile empirique est la fonction Q_n approximativement inverse de F_n (elle n'est pas inversible). $Q_n(\alpha)$ **est la valeur qui sépare les données en proportion** α **et** $1-\alpha$. Elle est définie précisément par :

$$Q_n(\alpha) = \check{x}_i \quad \text{si} \quad \frac{i-1}{n} \leqslant \alpha < \frac{i}{n}.$$

où la suite $(\check{x}_i)$ est la suite des x_i réordonnés par ordre croissant. On note généralement les quartiles : $Q1 = Q_n(0,25)$, $Q2 = Q_n(0,5)$, $Q3 = Q_n(0,75)$.

Dans la figure qui suit, les données « geyser » sont représentées sur l'axe réel par leur valeur. À gauche de la valeur $Q1$ (comme à droite de $Q3$) se trouve un quart des données :

Boîtes de dispersion. Il s'agit d'un format de représentation des données par une simple boîte de largeur (ou hauteur) $Q_3 - Q_1$, où Q_2 est indiqué par une séparation (chaque compartiment contient donc un quart des données) ; cette boîte est prolongée par deux traits : l'extrémité du premier correspond à la première donnée supérieure à $Q_1 - 1,5\Delta$, où $\Delta = Q_3 - Q_1$ et l'autre à la plus grande inférieure à $Q_3 + 1,5\Delta$. Les données extérieures à l'ensemble, c.-à-d. hors de $[Q_1 - 1,5\Delta, Q_3 + 1,5\Delta]$, sont représentées individuellement (s'il y en a). Cette représentation est surtout utilisée pour comparer graphiquement différents groupes.

II.4 Indices synthétiques essentiels

II.4.1 Mesures de localisation

La moyenne empirique est la moyenne arithmétique des données :

$$\bar{x} = \frac{1}{n} \sum x_i.$$

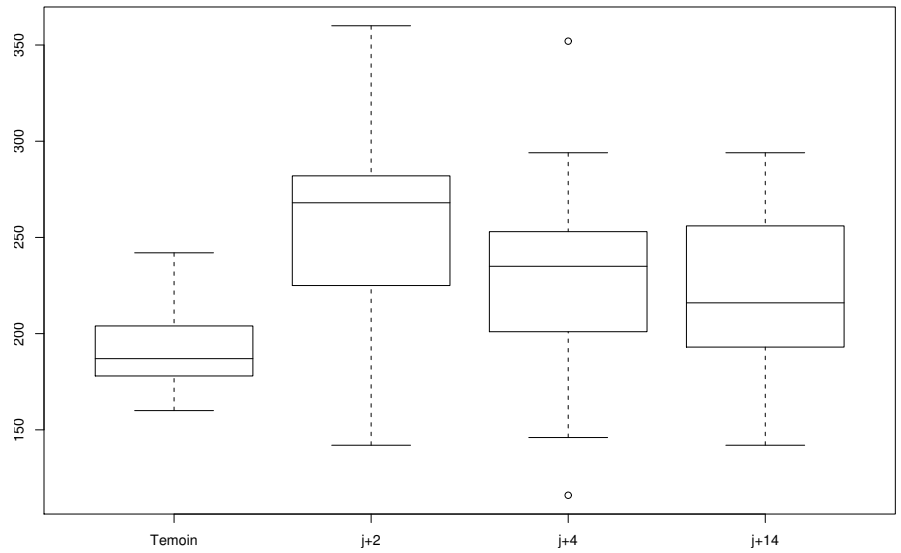


FIGURE II.7 – Boîte de dispersion. On mesure le taux de cholestérol sur un groupe témoin et sur un autre groupe de patients ayant eut une crise cardiaque, 2,4 et 14 jours après la crise. D'après OzDASL.

C'est aussi la valeur pour laquelle la somme des carrés des distances des données à cette valeur est minimale :

$$\bar{x} = \arg\min_y \sum_i (x_i - y)^2.$$

La moyenne des 222 données « geyser » est de 71mn.

Médiane. S'il y a un nombre impair de données, c'est la valeur centrale $\check{x}_{(n+1)/2}$. Sinon, c'est par convention la moyenne des deux valeurs centrales.

La médiane des 222 données « geyser » est de 75mn, ce que confirme la figure de la page 18. Sur cette figure, on trouve la première médiane à 3. Noter la dissymétrie de la distribution des données « geyser » (§ II.2.3) qui déplace la médiane (par rapport à la moyenne) vers une zone de plus grande probabilité.

INVARIANCE PAR CHANGEMENT D'ÉCHELLE MONOTONE : Si $y_i = f(x_i)$ pour une certaine fonction *monotone* f et si Q_2 est la médiane des x_i, alors $f(Q_2)$ est la médiane des y_i (à un petit décalage près si le nombre de données est pair). C'est une propriété importante qui n'est pas satisfaite par la moyenne.

II.4.2 Mesures de dispersion

On cherche ici à quantifier l'étendue des données.

La variance empirique est la variance de la distribution empirique, c.-à-d. la quantité définie par

$$Var(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 \; - \; \bar{x}^2.$$

On la note aussi s_x^2. Sa racine s_x est l'**écart-type empirique**, et sa dimension est celle des données. Sur les données « geyser », l'écart-type est $s = 12,8$.

On considère aussi des **intervalles interquantile**, par exemple $Q_3 - Q_1$ qui est l'étendue de la zone centrale contenant la moitié des données (cf § II.3.3).

II.4.3 Corrélation

Soient $x = (x_1, ...x_n)^T$ et $y = (y_1, ...y_n)^T$ deux vecteurs de $\mathbb{R}^n$, par exemple des paires (âge, revenu) pour des individus différents. On note leur moyennes respectives $\bar{x}$ et $\bar{y}$, et les vecteurs recentrés $\tilde{x}$ et $\tilde{y}$:

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{pmatrix}, \quad \bar{x} = \frac{1}{n}\sum_j x_j\,, \quad \tilde{x}_i = x_i - \bar{x}.$$

Propriété 10. *La covariance empirique de x et y est*

$$Cov(x,y) = \frac{1}{n}\sum_i \tilde{x}_i\tilde{y}_i = \overline{xy} - \bar{x}\bar{y}$$

et le coefficient de corrélation linéaire vaut

$$Cor(x,y) = \frac{Cov(x,y)}{s_x s_y} = \frac{\langle \tilde{x}, \tilde{y}\rangle}{\|\tilde{x}\|_2\ \|\tilde{y}\|_2}.$$

$Cor(x,y)$ est donc le cosinus de l'angle que forment les vecteurs des données centrées. $Cor(x,y)$ et $Cov(x,y)$ sont aussi notés parfois r_{xy} et c_{xy}.

Propriété 11. $|Cor(x,y)| \leqslant 1$. $|Cor(x,y)| = 1$ *si et seulement s'il existe* $(a,b) \in \mathbb{R}_* \times \mathbb{R}$ *tels que* $y_i = ax_i + b$, $i = 1, ..n$ *; dans ce cas* $Cor(x,y) = signe(a)$.

Une forte corrélation signifie donc que les y sont quasiment fonction linéaire des x. En pratique une faible corrélation sera souvent interprétée (abusivement) comme de l'indépendance (par analogie avec le cas gaussien).

Ainsi, l'importance de l'exposition au soleil pour le cancer de la peau a été démontrée simplement en calculant la corrélation entre les taux de cancer dans certaines régions et la latitude.

La corrélation entre l'âge de la mariée et l'âge du marié (tableau II.2) est de 0,8.

La corrélation entre la taux de CO2 et la température moyenne de le terre sur les 150 dernières années vaut 0,88. Voir la figure III.2 p. 31.

II.4.4 Corrélation partielle

Considérons l'exemple suivant : pour mesurer l'efficacité des pompiers, on calcule la corrélation entre le nombre de pompiers p envoyés sur un sinistre et le montant des dégâts d (en euros). On trouve

$$r_{pd} = 0,7.$$

Faut-il conclure de cette corrélation élevée que pour réduire les dégâts il faut diminuer les effectifs de pompiers ? Cette conclusion serait exacte si la statistique portait sur des incendies de même ampleur initiale. En d'autres termes, on voit bien que cette variable «ampleur initiale» a, étant fortement corrélée à p et d, introduit une corrélation « artificielle » entre le nombre de pompiers envoyés et les dégâts. Il faudrait calculer la corrélation sur des incendies d'ampleur initiale fixe, puis faire ensuite la moyenne sur cette variable.

La solution mathématique est de calculer la corrélation partielle entre p et d sachant a :

Définition 12. *Le coefficient de corrélation partielle entre x et y sachant z est la corrélation entre le projeté de x et de y sur l'orthogonal de z, elle est donnée par la formule*

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \tag{II.3}$$

Reprenons notre exemple. Si l'on a

$$r_{pa} = 0,9 \qquad \text{et} \qquad r_{va} = 0,9$$

alors $r_{pd|a} = -0,6$ et le signe est correct !

On interprète la nullité de $r_{xy|z}$ comme un indice d'indépendance de x et y conditionnellement à z : lorsque z est connu, y n'apporte aucune information supplémentaire sur x.

Voici deux exemples sur des données réelles où l'on voit la corrélation changer de signe (Kendall et Stuart, *Advanced Theory of Statistics*, Vol 2, Ex 27.1 & 27.2) :

1. Des statistiques portant sur 20 ans dans une région de Grande Bretagne

   $x =$ rendement à l'hectare
   $y =$ température moyenne au printemps
   $z =$ chutes de pluie

   et l'on trouve

   $$r_{yx} = -0,4, \quad r_{xz} = 0,8, \quad r_{yz} = -0,56, \quad r_{yx|z} = 0,1$$

2. Des statistiques portant sur 16 villes des Etats-Unis en 1935

   $x =$ criminalité
   $y =$ fréquentation des églises
   $z =$ nombre d'enfants par famille

   et l'on trouve

   $$r_{yx} = -0,31, \quad r_{xz} = -0,14, \quad r_{yz} = 0,85, \quad r_{yx|z} = 0,25.$$

II.5 Exercices

Exercice 1.

1. La variable gaussienne $\mathcal{N}(0,1)$ a une probabilité de 95% d'être en valeur absolue inférieure à 1,96 et 99% d'être en valeur absolue inférieure à 2,6. Quelle est la probabilité qu'elle soit inférieure à 1,96, à 2,6 ?

Une personne sème 400 graines. La probabilité de germination est $p = 0,8$. On veut calculer la probabilité P_{300} qu'au moins 300 germent.

2. Soit X_i la variable qui vaut 1 si la i-ième graine germe et 0 sinon. Quelle sont sa moyenne m et sa variance σ^2 ?
3. Quelle est approximativement la loi de $\dfrac{1}{20\sigma}\sum_i (X_i - m)$? (Utiliser le théorème-limite central).

4. En déduire une valeur (approximative) de P_{300}.
5. Estimer de même P_{320} et P_{340}.

Exercice 2. Soit X de loi

$$p(x) = e^{-x} 1_{x>0} dx.$$

1. Quelle est la fonction de répartition de $U = X^2$?
2. Quelle est la densité de U ?
3. Quelle est la médiane de U ?

Exercice 3. Soient U et V deux variables aléatoires dont la loi admet la densité

$$p(u, v) = e^{-u-v^2} 1_{u>0} \frac{dudv}{\sqrt{\pi}}.$$

1. Quelle est la loi de U ? Quelle est la loi de V ?
2. Quelle est la corrélation de U et V ?

Exercice 4. Soient X et Y deux variables aléatoires indépendantes de variance 1, et $U = X+Y$, $V = 1 + X - 2Y$; quelle est la covariance et la corrélation de U et V ?

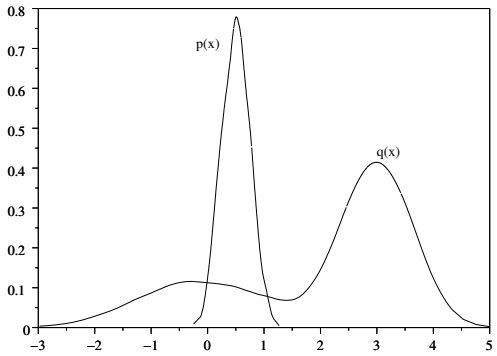
Exercice 5. Démontrer les formules II.1.

Exercice 6. Calculer les moyenne et médiane des deux lois de probabilités suivantes sur $\{1, 2, 3, 4\}$:

y_j	1	2	3	4
p_j	0,2	0,3	0,1	0,4

y_j	1	2	3	4
p_j	0,2	0,1	0,3	0,4

Exercice 7. (On justifiera brièvement les réponses) Soit les deux densités suivantes.



1. Les deux ensembles de données suivants : $A = \{0, 0.3, 0.7, 1\}$ et $B = \{-1, 2, 2.5, 3\}$ représentent chacun des v.a. tirées selon p ou q, mais on ne sait plus lequel a été tiré selon quelle loi. Doit-on attribuer (selon toute vraisemblance) A à p et B à q ou l'inverse ?
2. L'une des deux lois est gaussienne. Laquelle ?
3. Laquelle a la plus grande variance ?
4. Laquelle a la plus grande moyenne (espérance) ?

Exercice 8. Soit X_i une suite de variables indépendantes de densité p. On considère l'estimateur de densité :

$$p_n(x) = \frac{1}{nh} \sum_i K\left(\frac{x - X_i}{h}\right)$$

où $K(.)$ est par exemple le noyau d'Epanechnikov. Montrer en utilisant la loi des grand nombres (on admettra qu'elle s'applique) que pour tout h $p_n(x)$ tend vers une certaine limite $p_h(x)$. Quelle est la limite de $p_h(x)$ quand h tend vers 0 ?

Exercice 9. On considère les données suivantes sur la distribution de la population active britannique (en milliers d'habitants)

	Juin 1959	Juin 1966
Armée	565	417
Employés	23242	24974
Employeurs	1677	1673
Chômeurs	389	253

1. De combien (en %) a augmenté l'effectif des armées en Grande Bretagne ?
2. Même question pour l'effectif normalisé par le total de la population active
3. De combien diriez-vous que le chômage a augmenté en Grande Bretagne ?

Exercice 10. On dispose du tableau de notes suivant

note	0	5	7	8	9	10	11	12	14	20
effectif	2	8	11	8	9	13	14	4	0	

Pour simplifier, on suppose qu'aucune note de l'ensemble n'est entière (il y a 8 notes dans l'intervalle $]5, 7[$). Proposer un histogramme pour la distribution de ces notes. Comparer à l'histogramme par effectifs (difficilement justifiable ici).

Exercice 11. On considère les statistiques suivantes sur les taux de réussites au baccalauréat de deux lycées :

	Lycée A	Lycée B
Taux d'échec	0,03	0,02

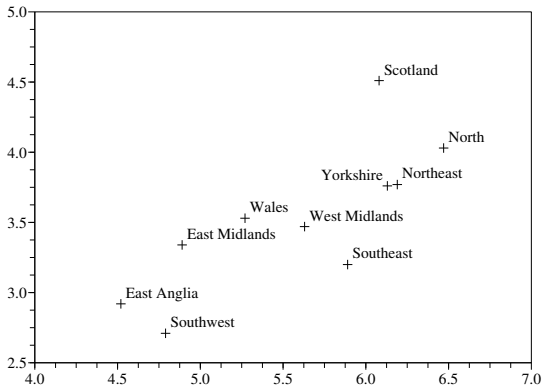
Quel lycée choisiriez-vous ? Une deuxième étude, plus fine, sépare les individus en deux groupes, ceux qui sont issus d'un milieu défavorisé et les autres :

	Lycée A		Lycée B	
	Effectif	Taux d'échec	Effectif	Taux d'échec
Défavorisés	1500	0,038	200	0,04
Favorisés	600	0,01	600	0,14
Ensemble	2100	0,03	800	0,02

Quel lycée choisiriez-vous ? Expliquer le paradoxe.

Noter que les conclusion pourraient s'inverser â nouveau s'il l'on apprenait que A a, comparé â B, beaucoup plus de sections où les taux de réussite au bac sont de manière générale plus élevés...

Exercice 12. Dépense moyenne par foyer et par semaine en alcool et en tabac (en livres), dans 10 régions de Grande Bretagne (Source : Department of Employment, 1981) :



Si x est la dépense en alcool et y la dépense en tabac, on trouve

$$\overline{\tilde{x}^2} = 0.41 \quad \overline{\tilde{y}^2} = 0.25 \quad \overline{\tilde{x}\tilde{y}} = 0.25$$

Calculer la corrélation. Qu'en dire ?

Pour l'Irlande du Nord, on trouve $x = 4.02$ et $y = 4.56$. Qu'en pensez-vous ?

Exercice 13. On considère les données de la table II.1. Calculer $\bar{x}$ et s^2. Le dé étant supposé non pipé, quelle est la valeur limite de ces quantités quand le nombre d'échantillons tend vers l'infini (utiliser la loi des grands nombres) ? Vérifier que s^2 est inférieur d'environ $0,1$ à sa limite théorique.

Exercice 14. Calculer la moyenne et la médiane de l'âge du marié de moins de 55 ans en Alaska en 1995, à partir du tableau II.2 (les données étant discrétisées par intervalles, on utilisera le mode de calcul de son choix).

Exercice 15. Soit $(x_1, ...x_n)$ une suite de données numériques. Soient $\bar{x}$ et s^2 les moyennes et variance empirique de cet échantillon.

1. Soit a un réel, que valent les moyennes et variances empiriques des suites $(x_i - a)$ et (x_i/a) ?
2. Que valent la moyenne et la variance empiriques de la suite $(x_i - \bar{x})/s$?

Exercice 16. On mesure dans 15 régions la température moyenne durant un été, la pluviosité moyenne, et la productivité en blé (en quintaux à l'hectare). On a ainsi un tableau à 3 variables (t, p, b) et 15 individus. On trouve les corrélations

$$r_{tp} = -0,5 \quad r_{tb} = -0,2 \quad r_{pb} = 0,5.$$

On observe que la corrélation entre la température et la production de blé est négative ! Expliquer.

Exercice 17. Montrer que la médiane est la valeur pour laquelle la somme des distances des données à cette valeur est minimale :

$$m_x = \arg\min_y \sum_i |x_i - y|.$$

On remarquera que la fonction $y \to \sum_i |x_i - y|$ est continue, affine par morceaux, avec une dérivée entière sur chaque morceau, allant de ? à ? par pas de ?. On pourra traiter séparément les cas « n pair » et « n impair ».

Exercice 18. Soit x un ensemble de données séparé en deux sous-ensembles y et z de taille n_y et n_z, montrer que

$$\bar{x} = p_y \bar{y} + p_z \bar{z}, \quad p_y = \frac{n_y}{n_y + n_z}, \quad p_z = \frac{n_z}{n_y + n_z}$$
$$s_x^2 = \{p_y s_y^2 + p_z s_z^2\} + \{p_y (\bar{y} - \bar{x})^2 + p_z (\bar{z} - \bar{x})^2\}.$$

Pour la seconde identité, on commencera par montrer que

$$\sum (y_i - \bar{x})^2 = \sum (y_i - \bar{y})^2 + n_y (\bar{y} - \bar{x})^2.$$

$\bar{x}$ est donc une moyenne pondérée des moyennes. s_x^2 est la somme de deux termes, le premier étant la moyenne pondérés des variances, appelée variance intra-classe ; montrer que le second, appelé variance inter-classe, peut s'interpréter comme la variance d'une certaine variable aléatoire.

III

Régression linéaire

III.1 Introduction

Supposons que l'on cherche à expliquer le « taux de fréquentation du médecin » d'une personne en général par son âge et son sexe seulement. Un des modèles les plus simples est le modèle linéaire

$$y = \beta_1 + \beta_2 a + \beta_3 s$$

où y est la variable « taux de fréquentation du médecin » (en nombre de visites par an), a est la variable « âge » et s est la variable « sexe » (à valeur 0 ou 1). β_1, β_2 et β_3 sont les paramètres. Cette relation est bien entendu approximative, en un sens que la théorie des probabilités va permettre de préciser.

On se propose alors d'estimer ces paramètres en recueillant des données (y_i, a_i, s_i) sur un certain échantillon d'individus. Ce type estimation a trois applications essentielles :

- Détermination des **facteurs significatifs** : l'âge a-t-il une influence significative sur le taux de fréquentation du médecin ? (c-à-d β_2 est-il non nul ?)
- **Prédiction/simulation** : combien de médecins faut-il pour une ville de pyramide des âges donnée ?
- **Détection de changement** : Le ticket modérateur a-t-il provoqué un changement significatif dans le comportement des patients ? (c.-à-d. β a-t-il changé ?)

Exemple : le modèle de Cobb-Douglas. On considère les variables, chacune concernant la totalité des États-Unis et i étant l'indice d'une année :

- P_i : production
- K_i : capital (valeur des usines)
- T_i : travail fourni (basé sur un calcul du nombre total de travailleurs)

On cherche à expliquer P_i à l'aide des variables (K_i, T_i). Le modèle de Cobb et Douglas [1] postule la relation suivante :

$$P = \alpha K^{\beta_2} T^{\beta_3}.$$

α et les β_j sont les paramètres du modèle. On va prendre le logarithme dans cette équation,ce

1. "A theory of production", *American Economic Review*, 18, 139-165, 1928.

qui nous ramène au cas linéaire. Si l'on pose

$$\beta_1 = \log(\alpha)$$
$$y = \log(P)$$
$$x = (1, \log(K), \log(T)),$$

il vient

$$y = \sum_{j=1}^{3} \beta_j x_j$$

qui est un modèle linéaire. Voici les données utilisées dans l'étude de Cobb et Douglas (1928) :

Année	P	K	T	Année	P	K	T	Année	P	K	T
1899	100	100	100	1907	151	176	138	1915	189	266	154
1900	101	107	105	1908	126	185	121	1916	225	298	182
1901	112	114	110	1909	155	198	140	1917	227	335	196
1902	122	122	118	1910	159	208	144	1918	223	366	200
1903	124	131	123	1911	153	216	145	1919	218	387	193
1904	122	138	116	1912	177	226	152	1920	231	407	193
1905	143	149	125	1913	184	236	154	1921	179	417	147
1906	152	163	133	1914	169	244	149	1922	240	431	161

On va voir dans ce chapitre que l'on peut estimer les β_i et l'on trouve $\beta_1 = 0.23$ et $\beta_2 = 0.81$; de plus on vérifie par une méthode d'intervalle de confiance que l'écart observé sur ces données entre $\beta_1 + \beta_2$ et 1 n'est pas significatif.

III.2 Cas unidimensionnel

III.2.1 Le modèle

Les données consistent en des variables appelées « réponses » $y_i \in \mathbb{R}$ et des « variables explicatives » (ou « régresseurs ») $x_i \in \mathbb{R}$, $i = 1, ...n$, chaque paire (y_i, x_i) représentant une expérience, un « individu ». On se donne un modèle idéal de la forme

$$y = \alpha^* + x\beta^*.$$

Ici β^* désigne la vrai paramètre, la lettre β étant utilisée comme variable muette, paramètre générique. Ce modèle idéal, qui postule que le point (x, y) appartient à une droite fixée à l'avance, n'est malheureusement pas réalisé sur des données réelles, pour des raisons diverses, comme par exemple le bruit de mesure, ou simplement les facteurs aléatoires de variation d'un individu à l'autre, voir par exemple la figure III.1 ; on représente alors les observations par l'équation

$$y_i = \alpha^* + x_i\beta^* + u_i, \quad i = 1, ...n$$

où u_i est le défaut d'adéquation de la i-ième observation au modèle. On considère généralement que la suite u_i est une suite i.i.d de variables gaussiennes centrées ; dans ce modèle, les x_i sont des quantités déterministes, tandis que y_i est une variable aléatoire gaussienne de moyenne $\alpha^* + x_i\beta^*$.

L'interprétation « prédictive » de ce modèle est que la valeur la plus vraisemblable de y_i pour quelqu'un qui n'a accès qu'à x_i est donnée par la formule

$$\widehat{y_i} = \alpha^* + x_i\beta^*.$$

On va s'intéresser essentiellement à l'estimation des valeurs de (α^*, β^*) à partir d'un ensemble de données $(x_i, y_i)_{i=1,...n}$.

III.2.2 Moindres carrés et maximum de vraisemblance gaussien

On postule le modèle

$$y_i = \alpha^* + x_i\beta^* + u_i, \quad i = 1, ..., n$$

où les u_i sont des variables aléatoires gaussiennes indépendantes de moyenne nulle et de même variance σ^2. Les x_i sont déterministes ; les y_i sont des variables aléatoires indépendantes avec

$$y_i \sim \mathcal{N}(\alpha^* + x_i\beta^*, \sigma^2).$$

Chaque y_i a une densité de probabilité

$$p(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \alpha^* - x_i\beta^*)^2}{2\sigma^2}\right).$$

Le vecteur des $y = (y_i)_i$ a donc une densité de probabilité jointe

$$p(y) = (\sigma\sqrt{2\pi})^{-n} \prod_i \exp\left(-\frac{(y_i - \alpha^* - x_i\beta^*)^2}{2\sigma^2}\right) = (\sigma\sqrt{2\pi})^{-n} \exp\left(-\frac{\sum_i (y_i - \alpha^* - x_i\beta^*)^2}{2\sigma^2}\right).$$

L'estimateur au maximum de vraisemblance de (α^*, β^*) est celui qui maximise cette probabilité, c'est-à-dire la valeur de (α^*, β^*) pour laquelle on a le plus de chance d'avoir observé les réponses. C'est donc (indépendamment de σ) :

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{(\alpha,\beta)} \sum_i (y_i - \alpha - x_i\beta)^2.$$

Cette formule exprime que l'ajustement se fait par moindre carrés : on choisit la paire (α, β) qui minimise la somme des carrés des erreurs et prédiction. Cette méthode est due à Gauss.

III.2.3 Calcul des paramètres

L'ajustement se fait donc par minimisation de

$$S(\alpha, \beta) = \sum_i (y_i - \alpha - x_i\beta)^2$$

en α et β. Cette fonction quadratique de (α, β) atteint son minimum pour $\partial S/\partial\alpha = 0$ et $\partial S/\partial\beta = 0$, soit

$$\sum_i y_i - \alpha - x_i\beta = 0 \tag{III.1}$$

$$\sum_i x_i(y_i - \alpha - x_i\beta) = 0. \tag{III.2}$$

En divisant les équations par n, il vient

$$\bar{y} - \alpha - \bar{x}\beta = 0$$

$$\overline{xy} - \alpha\bar{x} - \beta\,\overline{x^2} = 0.$$

La solution de ce système linéaire s'obtient simplement en éliminant α :

$$\widehat{\beta} = \frac{c_{xy}}{s_x^2} = \frac{s_y r_{xy}}{s_x}$$

$$\widehat{\alpha} = \bar{y} - \bar{x}\widehat{\beta}$$

où s_x et c_{xy} désignent la variance empirique des x_i et la covariance empirique de x et y. Deux exemples sont donnés sur la figure III.1.

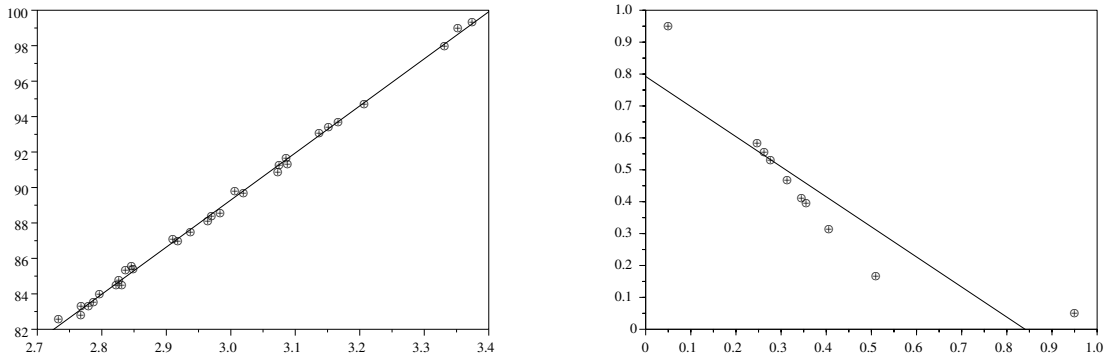


FIGURE III.1 – Points (x_i, y_i) et la droite de regression $y = \widehat{\alpha} + \widehat{\beta}x$. Dans le premier exemple, il s'agit de la température d'ébulition de l'eau mesurée en divers endroits de l'Himalaya en fonction du logarithme de la pression atmosphérique (J.D. Hooker, 1849). $r_{xy} = 0,999$. Dans le deuxième exemple soit l'hypothèse de modèle linéaire n'est pas satisfaite, soit il y a un point aberrant.

III.2.4 Propriétés des écarts résiduels. Décomposition de la variance

Définition 13. *Les écarts résiduels $\widehat{u}_i$ sont définis par*

$$\widehat{u}_i = y_i - \widehat{\alpha} - \widehat{\beta}x_i.$$

La proposition qui suit montre en particulier le lien entre la corrélation et la quantité $s^2_{\widehat{u}}/s^2_y$ que l'on peut interpréter comme l'épaisseur relative du nuage de points.

Proposition 14. *Les vecteurs $\widehat{u}$ et $\widehat{y}$ sont orthogonaux et*

$$\bar{\widehat{u}} = 0, \qquad s^2_{\widehat{u}} = (1 - r^2_{xy})s^2_y = s^2_y - s^2_{\widehat{y}}.$$

Démonstration. La première relation est simplement l'équation (III.1). L'orthogonalité est une conséquence immédiate de (III.1) et (III.2). Pour la seconde :

$$s^2_{\widehat{u}} = Var(y - \widehat{\beta}x) = Var(y) + \widehat{\beta}^2 Var(x) - 2\widehat{\beta}Cov(y, x) = s^2_y - r^2_{xy}s^2_y$$

ce qui est bien le résultat. La troisième vient de l'orthogonalité de $\widehat{u}$ et de $\widehat{y}$ recentré. ∎

La relation $s^2_y = s^2_{\widehat{u}} + s^2_{\widehat{y}}$ exprime que la variance des résidus (incertitude sur la réponse lorsqu'on connaît les régresseurs) est la variance de la réponse à laquelle on retranche la variance des prédictions. Elle quantifie l'apport de l'information apportée par x pour connaître y. Elle quantifie aussi la séparation effectuée par le modèle entre l'aléatoire (les u_i) et le déterministe ($x\beta$) comme annoncé §I.1.

On retrouve bien en particulier que si $|r_{xy}| = 1$, alors $\widehat{u} = 0$, c'est-à-dire que y et x sont en relation affine.

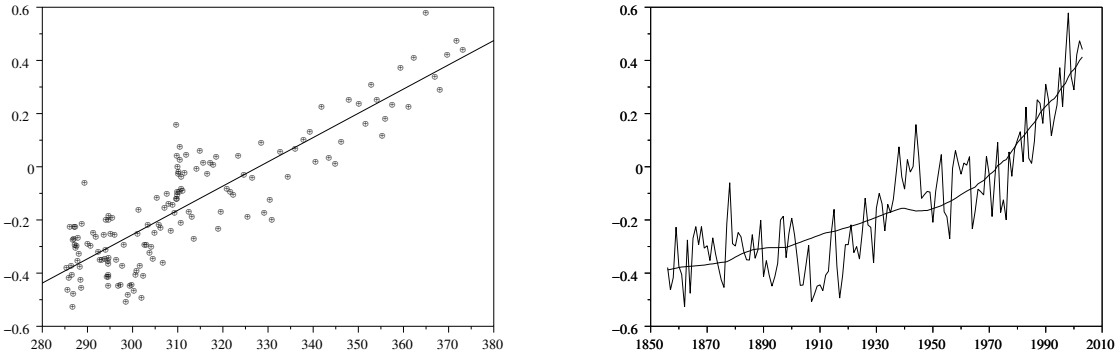


FIGURE III.2 – A gauche : Les 150 points (CO2,T) et la droite de régression : $y = 0,009\,x - 3$. On voit une nette corrélation ($r = 0.88$). A droite : Tracé de T (courbe agitée) et du CO2 modifié par un changement d'échelle linéaire (courbe régulière)

III.2.5 Exemple

On s'intéresse aux relations entre le taux de CO2 dans l'atmosphère et la température moyenne de la planète. On dispose de mesures annuelles de 1854 à 2003 [2]. Elles sont représentées sur la figure III.2 (la température est mesurée en écart à la moyenne calculée sur 1961-1990, on voit donc une augmentation d'à peu près 1 degré). Pour comparer l'évolution de ces deux variables au cours du temps la méthode consistant à les tracer telles quelles sur un même graphique n'est pas satisfaisante car elles vivent à des échelles totalement différentes. La solution est de faire un changement d'échelle linéaire sur le CO2 en utilisant la nouvelle variable issue de la régression linéaire

$$\begin{aligned} \text{Modèle :} \quad & t_i = \alpha + \beta c_i + u_i \\ \text{Nouvelle variable :} \quad & c'_i = \widehat{\alpha} + \widehat{\beta} c_i \end{aligned}$$

On obtient ainsi la figure III.2 à droite superposant correctement la température (en joignant les points $(i + 1853, T_i)$) et le taux de CO2 (points $(i + 1853, c'_i)$).

III.3 Régression multiple

III.3.1 Les données

Les données consistent en des variables observées y_i et des variables explicatives (ou régresseurs) x_i, $i = 1, ...n$, chaque paire (y_i, x_i) représentant une expérience. On les arrange dans un tableau de la façon suivante :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

x_i est donc un vecteur ligne. On notera $x_{.j}$ la j-ième colonne. On convient généralement que le premier régresseur est la constante, mais ce n'est pas obligatoire. Le modèle s'exprime cette fois

2. Source : `www.cru.uea.ac.uk/cru/data/temperature` et `www.giss.nasa.gov/data/simodel/ghgases`

cette fois sous la forme

$$y_i = x_i\beta^* + u_i = \beta_1^* + \sum_{i=2}^{p} x_{ij}\beta_j^* + u_i, \quad i = 1, ...n,$$

les u_i étant i.i.d. gaussiens ; ceci se réécrit de manière plus compacte

$$y = X\beta^* + u$$

et la prédiction linéaire de y_i par x_i est

$$\widehat{y}_i = x_i\beta^*, \quad \text{ou} \quad \widehat{y} = X\beta^*.$$

Remarquer que $X\beta^*$ n'est autre que la combinaison linéaire des colonnes de X avec les coefficients $\beta_1^*, ...\beta_p^*$.

III.3.2 Estimation de β^*

Définition 15. *Pour tout β, on définit l'erreur quadratique moyenne d'ajustement (ou erreur résiduelle) $S(\beta)$ comme*

$$S(\beta)^2 = \frac{1}{n}\|y - X\beta\|^2 = \frac{1}{n}\sum_i (y_i - x_i\beta)^2.$$

Le vecteur des coefficients de régression calculés aux moindres carrés ordinaires est

$$\widehat{\beta} = \arg\min_{\beta} S(\beta)$$

Proposition 16. *On suppose la matrice X^TX inversible. Alors $\widehat{\beta}$ est donné par $\widehat{\beta} = (X^TX)^{-1}X^Ty$.*

Démonstration. On a : $\dfrac{\partial S(\beta)^2}{\partial \beta_j} = \dfrac{1}{n}\sum_i 2x_{ij}(y_i - x_i\beta) = \dfrac{2}{n}(X^T(y - X\beta))_j$, d'où l'équation $X^T(y - X\beta) = 0$. ∎

Définition 17.

- *Vecteur des valeurs ajustées : $\widehat{y} = X\widehat{\beta}$, soit $\widehat{y}_i = x_i\widehat{\beta}$.*
- *Vecteur des résidus : $\widehat{u} = y - \widehat{y}$.*

III.3.3 Décomposition de la variance et le coefficient de corrélation multiple R

On va voir que le vecteur $\widehat{u}$ est orthogonal aux colonnes de X, ce qui a des conséquences importantes. On suppose ici que la première colonne de X est consituée de 1.

Proposition 18. *On a :*

$$\widehat{u} \perp x_{.j}, \ j = 1, ...p$$
$$s_y^2 = s_{\widehat{y}}^2 + s_{\widehat{u}}^2.$$

Démonstration. Pour l'orthogonalité, il suffit de vérifier que $X^T\widehat{u} = 0$, mais :

$$X^T\widehat{u} = X^T y - X^T X\widehat{\beta} = X^T y - X^T X(X^T X)^{-1}X^T y = 0.$$

Pour la deuxième affirmation, rappelons que

$$y = \widehat{y} + \widehat{u}.$$

Mais l'orthogonalité de $\widehat{u}$ à $\mathbf{1}$ (i.e. $x_{.1}$) implique que $\bar{\widehat{u}} = 0$, et que donc $\bar{y} = \bar{\widehat{y}} + \bar{\widehat{u}} = \bar{\widehat{y}}$. D'où

$$y - \bar{y} = (\widehat{y} - \bar{\widehat{y}}) + \widehat{u}.$$

Le fait que $\widehat{y} - \bar{\widehat{y}}$ est combinaison linéaire des colonnes de X et que $\widehat{u}$ est orthogonal à ces dernières implique que l'on peut appliquer le théorème de Pythagore, ce qui conduit à

$$\|y - \bar{y}\|^2 = \|\widehat{y} - \bar{\widehat{y}}\|^2 + \|\widehat{u}\|^2$$

qui est le résultat cherché car $\bar{\widehat{u}} = 0$. ∎

Définition 19. *Le coefficient de détermination R^2 est le rapport de la « variance expliquée » par la « variance totale »*

$$R = \frac{s_{\widehat{y}}}{s_y}, \qquad R^2 = 1 - \frac{s^2_{\widehat{u}}}{s^2_y}.$$

Plus R est proche de 1, plus le modèle est adéquat sur les données. R^2 est donc une mesure de comparaison entre la prédiction par le modèle et la prédiction par une constante, via les erreurs d'ajustement.

On montre facilement que si $p = 2$ (cas du §III.2), $R = |r_{xy}|$.

III.3.4 Distribution de $\widehat{\beta}$. Estimation de σ_u^2. Intervalle de confiance

On montre facilement, en remplaçant y par $X\beta^* + u$ dans les formules, que

$$\widehat{\beta} = \beta^* + (X^T X)^{-1}X^T u.$$

Comme les u_i sont i.i.d. gaussiens, ceci implique que $\widehat{\beta}_j$ est une variable gaussienne de moyenne β_j^* ; un calcul permet alors d'obtenir sa variance en fonction de celle commune aux u_i :

$$E[(\widehat{\beta}_j - \beta_j^*)^2] = \sigma_j^2 = \sigma_u^2((X^T X)^{-1})_{jj}.$$

La variable gaussienne centrée réduite $\sigma_j^{-1}(\widehat{\beta}_j - \beta_j^*)$ étant en valeur absolue inférieure à $1,96$ avec probabilité 0,95, on a l'intervalle de confiance pour β_j^* :

$$\beta_j^* \in [\widehat{\beta}_j - 1,96\ \sigma_j\ ,\ \widehat{\beta}_j + 1,96\ \sigma_j] \qquad \text{avec probabilité de confiance 0,95.}$$

Ce qui veut dire qu'on a 95% de chance de ne pas se tromper en affirmant que le paramètre inconnu β_j^* appartient à cet intervalle (le concept d'intervalle de confiance sera discuté plus en détail au chapitre suivant). Cette formule ne peut être utilisée telle quelle car σ_u est inconnu. L'estimateur usuel de σ_u est le suivant

$$\widehat{\sigma}_u^2 = \frac{1}{n-p}\sum_{i=1}^{n}\widehat{u}_i^2.$$

On montre que cet estimateur est non biaisé [3] : $E[\widehat{\sigma}_u^2] = \sigma_u^2$. On pourrait remplacer σ_u par $\widehat{\sigma}_u$ dans l'intervalle de confiance ci-dessus, ce qui conduit à une approximation raisonnable. Les statisticiens ont cependant découvert une propriété remarquable [4] :

$$\frac{\widehat{\beta}_j - \beta_j^*}{\widehat{\sigma}_j} \quad \text{suit une loi de Student à } n-p \text{ degrés de libertés,} \qquad \widehat{\sigma}_j^2 = \widehat{\sigma}_u^2((X^TX)^{-1})_{jj}.$$

c.-à-d. la loi d'une v.a. $\mathcal{N}(0,1)$ divisée par un un χ^2 à $n-p$ degrés de libertés, ou encore la loi de

$$\frac{X}{(Y_1^2 + ... + Y_{n-p}^2)/(n-p)}$$

où X et les Y_i sont $\mathcal{N}(0,1)$ indépendants. On voit que si n est grand cette variable est presque gaussienne. Si l'on note $t_{n-p}(x)$ la fonction quantile de cette loi (les ordinateurs la calculent pour vous), la variable de Student appartient à l'intervalle $[t_{n-p}(\alpha/2), t_{n-p}(1-\alpha/2)]$ avec probabilité $1-\alpha$ et comme par symétrie $t_{n-p}(\alpha/2) = -t_{n-p}(1-\alpha/2)$ on l'intervalle de confiance (en notation allégée)

$$\boxed{\beta_j^* = \widehat{\beta}_j \pm \widehat{\sigma}_j t_{n-p}(1-\alpha/2) \qquad \text{avec probabilité de confiance } 1-\alpha.}$$

Si 0 est en dehors de l'intervalle de probabilité de confiance $1-\alpha$ pour β_j^* on dit que $\widehat{\beta}_j$ est **significativement non-nul** au niveau α (on prend typiquement $\alpha = 5\%$ ou 1%).

III.3.5 Application : efficacité d'un médicament

Les idées présentées ici seront précisées dans le chapitre suivant.

Pour voir si un médicament a un effet, on prend deux populations de patients, la première étant traitée avec un placebo et la seconde avec le médicament à tester. La réponse y est le temps de rémission. Comme on pense que la réponse a de fortes chances d'être influencée par le poids du sujet, on utilisera également cette variable pour le modèle. On postule alors le modèle :

$$y = x\beta^* + u$$

où le régresseur vaut

$$x = (1, 1_{medic}, poids).$$

Si bien que pour un patient sous placebo on a

$$y_i = \beta_1^* + p_i\beta_3^* + u_i$$

et pour un patient sous médicament

$$y_i = \beta_1^* + \beta_2^* + p_i\beta_3^* + u_i.$$

β_2^* représente donc l'effet du médicament en terme de diminution du temps de rémission. Finalement, dire que le médicament a de l'effet revient à dire que β_2^* est significativement non-nul.

3. La division par $n-p$ au lieu de n s'explique intuitivement ainsi : un estimateur non biaisé de σ_u^2 est $\frac{1}{n}\sum_{i=1}^n u_i^2$; en remplaçant les u_i par les $\widehat{u}_i$ on fait diminuer la somme (car on fait les moindres carrés) ; il se trouve que de diviser par $n-p$ au lieu de n compense exactement

4. Il est en effet remarquable que la loi de la statistique $\widehat{\sigma}_j^{-1}(\widehat{\beta}_j - \beta_j^*)$ ne dépende pas des données, c'est une statistique « pivotale » ; l'existence de statistiques pivotales intéressantes est loin d'être une généralité, c'est ici une contrepartie de la simplicité du modèle de régression linéaire.

III.3.6 Sélection des variables

Il s'agit de choisir les variables les plus significatives, l'idée étant d'éliminer les régresseurs dont la contribution à la prédiction est trop faible, c'est-à-dire d'éliminer les $\widehat{\beta}_j$ non significativement non-nuls.

La méthode la plus simple est la méthode descendante : On part du modèle qui ajuste y_i avec tous les régresseurs

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i$$

et l'on obtient un certain vecteur de résidus $\widehat{u}$.

On se propose alors de retrancher un à un les autres régresseurs, par ordre d'importance décroissante. Pour chacun des p régresseurs présents on calcule la valeur de $\|\widehat{u}'\|$ correspondant au retrait de ce régresseur et l'on choisit celui pour lequel $\|\widehat{u}'\|$ est le plus petit. On a alors un modèle avec un régresseur de moins :

$$y_i = \sum_{j \neq j_1} x_{ij}\beta_j + u_i.$$

Noter que les $\widehat{\beta}_j$ ont changé depuis le premier modèle, car on les réestime.

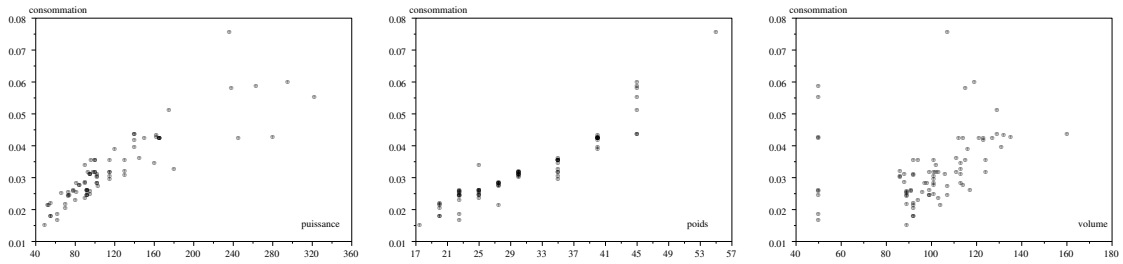
On choisit ensuite la variable à retrancher parmi les $p-1$ restantes, etc...

On s'arrête quand le coefficient $\widehat{\beta}_j$ que l'on s'apprête à retirer est jugé significativement non-nul à un niveau α spécifié à l'avance. Un calcul algébrique montre que ceci revient à

- S'arrêter si $\quad (n-k)\dfrac{\|\widehat{u}'\|^2 - \|\widehat{u}\|^2}{\|\widehat{u}\|^2} > t^2_{n-k}(1-\alpha/2).$

k est le nombre de variables en cours (p la première fois). On s'arrête donc quand l'erreur de prédiction fait un saut relatif trop grand.

Donnons par exemple le cas de la prédiction de la consommation (gallons par mile) des voitures [5] en fonction des variables « Volume » (pieds-cube), « Puissance » (chevaux) et « Poids » (livres). Voici une représentation des données



La régression donne la table d'analyse des coefficients suivante (82 individus). Attention tous les chiffres doivent être divisés par 10000 :

Facteur	Estimée $\widehat{\beta}_j$	Écart-type $\widehat{\sigma}_j$
Volume	-0.53	2
Poids	90	9.8
Puissance	5.8	1.3

5. Heavenrich, Murrell, and Hellman, "Light Duty AutomotiveTechnology and Fuel Economy Trends Through 1991, U.S.", Environmental Protection Agency, 1991 (EPA/AA/CTAB/91-02). Disponible par Internet sur DASL.

Comme $t_{78}(0.025) \simeq 2$ ($78 = 82 - 4$, ne pas oublier le régresseur constant) on trouve que le volume n'a pas d'influence significative à 5% (il n'en n'a pas non plus à 50% !). La méthode descendante élimine bien cette variable puis garde les deux autres. La régression avec les deux variables restantes conduit à un R^2 de $0,91$, et $\widehat{\sigma} = 0.0032$ pour une réponse variant entre $0,01$ et $0,08$.

III.4 Exercices

Exercice 1. Pour vérifier les relations d'allométrie entre insectes, on a retenu les deux variables

x=logarithme de la longueur de l'élytre

y=logarithme de la largeur de la tête.

Les mesures sur 50 insectes, notées (x_i, y_i) ont fourni les résultats suivants :

$$\sum_{i=1}^{50} x_i = 155 \qquad \sum_{i=1}^{50} y_i = 125 \qquad \sum_{i=1}^{50} x_i y_i = 391,1$$
$$\sum_{i=1}^{50} x_i^2 = 482,5 \qquad \sum_{i=1}^{50} y_i^2 = 320,5 \qquad \sum_{i=1}^{50} x_i^2 y_i^2 = 3468,7$$

1. Calculer
   (a) La moyenne et l'écart-type du caractère x sur l'échantillon observé
   (b) La moyenne et l'écart-type du caractère y sur l'échantillon observé
   (c) La covariance empirique et la corrélation empirique des variables x et y
   (d) L'équation de la droite de régression de y sur x obtenue par estimation sur ces données.
2. En déduire la loi d'allométrie exprimant la largeur de la tête en fonction de la longueur de l'élytre ($\log(0,046) = -3,08$).

Exercice 2. Les données suivantes représentent la taille de 161 enfants du village Égyptien de Kalama suivis entre 18 et 29 mois, moyennées par tranche d'âge [6] :

âge (mois)	18	19	20	21	22	23	24	25	26	27	28	29
taille (cm)	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

Tracer les points, la droite de régression (au jugé...) et en déduire graphiquement l'équation de la droite de régression (on cherchera deux points visiblement proches de la vraie droite de régression ce qui permettra de calculer son équation, à une légère erreur près).

Exercice 3. Soit le modèle

$$y_i = b^* x_i + u_i$$

où les u_i sont i.i.d $\mathcal{N}(0, \sigma^2)$. Par exemple y_i est le poids du i-ième individu et x_i le carré de sa taille.

1. Comment va-t-on définir cette fois l'erreur quadratique d'ajustement $S(b)$?
2. Donner l'expression du $\widehat{b}$ correspondant.

6. Source : D.S. Moore & G.P. McCabe (1989), *Introduction to the Practice of Statistics*, p. 118.

3. On propose un autre choix

$$\check{b} = \frac{\sum y_i}{\sum x_i}.$$

Quelle est la loi de $\widehat{b}$? Quelle est la loi de $\check{b}$? Quel est à votre avis le meilleur estimateur ?

Exercice 4. On dispose de deux qualités de papier. La première donne des feuilles de poids unitaire β_1, et la seconde des feuilles de poids unitaire β_2 ; β_1 et β_2 sont inconnus.

1. Une première ramette contient 200 feuilles de papier de qualité 1 et 300 de qualité 2. La mesure du poids de cette ramette donne p_1. Une seconde ramette contient 600 feuilles de papier de qualité 1 et 400 de qualité 2. La même mesure donne p_2. Trouver β_1 et β_2 en fonction de p_1 et p_2.
2. Une troisième ramette contient 500 feuilles de chaque qualité Quel poids attendrait-on ? Les incertitudes dans les pesées font que p_3 n'a pas la valeur attendue. Construire un modèle de régression multiple permettant de faire l'ajustement du poids d'une ramette sur le nombre de feuilles de chaque qualité dans la ramette.
3. On observe $p_1 = 610, p_2 = 1290, p_3 = 1240$; donner $(\widehat{\beta}_1, \widehat{\beta}_2)$ et $\widehat{\sigma}_u^2$. En déduire des intervalles de confiance approchés pour β_1 et β_2.

Exercice 5. Un fabricant de chocolats s'intéresse à l'influence de la teneur en sucre sur la qualité du produit. On fait donc goûter des chocolats de même composition (en dehors de la proportion de sucre) à différents testeurs. Ces derniers mettent une note comprise entre 0 et 20. On a donc les variables « proportion de sucre » et « note ».

On suppose que la note s'approxime bien par une fonction quadratique (c-à-d un polynôme du second degré) de la proportion.

1. Mettre cette expérience sous forme d'un problème de régression.
2. Pourquoi aurait-il été déraisonnable de considérer une fonction linéaire (plutôt que quadratique) de la proportion.
3. Les chocolats ont en fait une autre différence : l'origine du chocolat. Il y a donc une variable supplémentaire à valeur 0 ou 1 indiquant la provenance (« Venezuela » ou « Brésil »). Comment modifier le modèle pour prendre en compte cette nouvelle information ? Comment voir si l'origine compte effectivement ?

Exercice 6. On se donne une équation exprimant l'effet d'un traitement en fonction de la dose de produit utilisé

$$y = e^{\beta_1 + \beta_2 x} x^{\beta_3}$$

et l'on veut estimer $(\beta_1, \beta_2, \beta_3)$ à partir de données (x_i, y_i). Proposer un modèle linéaire de régression basé sur un changement de variables.

Exercice 7. On fait une régression de y sur deux variables explicatives x et z, c-à-d $X = (\mathbf{1}, x, z)$; il y a en tout n individus. On a obtenu le résultat suivant :

$$X^T X = \begin{pmatrix} 5 & 3 & 0 \\ . & 3 & 1 \\ . & . & 1 \end{pmatrix}$$

1. Reconstituer $X^T X$. Que vaut n ?

2. Que valent $\bar{x}$ et $\bar{z}$? Que vaut le coefficient de corrélation linéaire empirique entre x et z ?

La régression linéaire fournit les résultats :

$$y_i = -1 + 3x_i + 4z_i + \widehat{u}_i, \quad \|\widehat{u}\| = 2.$$

3. Que vaut la moyenne empirique $\bar{y}$?

On s'intéresse au modèle privé du régresseur z :

$$y = X_0\beta_0 + u, \quad X_0 = (\mathbf{1}_n, x).$$

4. Calculer numériquement $X_0^T y$ (commencer par calculer $X^T y$).
5. Calculer numériquement $\widehat{\beta}_0$.
6. Calculer la norme de $\widehat{u}_0$:
   (a) Remarquer que $\widehat{u}_0 = \widehat{u} + (\widehat{y} - \widehat{y}_0)$ et utiliser le théorème de Pythagore (se souvenir que $\widehat{u}$ est orthogonal aux colonnes de X).
   (b) Développer ensuite en remplaçant matriciellement les prédictions par leur valeur en fonction des régresseurs et des coefficients estimés. On utilisera également que pour deux vecteurs colonnes v et w, on a $\langle v, w\rangle = v^T w$ et $\|v\|^2 = v^T v$.
7. Calculer le coefficient de corrélation empirique r_{xy} entre x et y. On utilisera la proposition 14.

Exercice 8. Calcul du R^2.

1. En utilisant que $\widehat{u} \perp x_{.j}$ avec $j = 1$, montrer que $\sum_{i=1}^n \widehat{y}_i = n\bar{y}$
2. Montrer alors simplement avec le théorème de Pythagore que : $\displaystyle\sum_i \widehat{y}_i^2 = \sum_i (\widehat{y}_i - \bar{y})^2 + n\bar{y}^2$.
3. Exprimer R^2 en fonction de $\bar{y}$, $\|\widehat{y}\|$ et $\|\widehat{u}\|$.

Exercice 9. On est dans le cadre de la régression linéaire multiple. Montrer que l'on a $\widehat{y} = Py$ où P est une matrice de projection orthogonale ne dépendant que de X (c-à-d que $P^T = P$ et $P^2 = P$).

Exercice 10. On est dans le cadre de la régression linéaire multiple.

1. Exprimer matriciellement $\widehat{\beta}$ en fonction de β^*, X et u.
2. Soit v un vecteur colonne, quelle est la composante (i, j) de la matrice vv^T ?
3. Que vaut $E[uu^T]$?
4. Calculer matriciellement $E[(\widehat{\beta} - \beta^*)(\widehat{\beta} - \beta^*)^T]$ en exploitant la linéarité de l'espérance.
5. En déduire la covariance de $\widehat{\beta}_j$ et $\widehat{\beta}_k$.

Exercice 11. On dispose des données suivantes :

x	-1	0	1	2
y	1	0	2	2

pour lesquelles on postule le modèle linéaire habituel : $y_i = \beta_1 + \beta_2 x_i + u_i$.

1. Exprimer la matrice X et calculer $\widehat{\beta}$. Donner les résidus correspondants.
2. Quelqu'un vous dit d'utiliser le régresseur $|x_i|$ plutôt que x_i. Calculer les résidus correspondants. Cette personne a-t-elle raison ?
3. Tracer sur un même graphique les points (x_i, y_i), la première droite de régression, et la seconde fonction de régression.

Exercice 12. Démontrer que si $p = 1$, $R = |r_{xy}|$.

IV

Estimation. Tests. Exemples

IV.1 Introduction

Le but de l'estimation est de calculer une certaine quantité dépendant de la distribution d'une variable aléatoire Y. Cette quantité peut être un moment :

$$\theta^* = E[Y^4]$$

un quartile

$$P(Y < \theta^*) = 0,25$$

ou autre....

Dans un problème d'estimation, on ignore la distribution de Y, mais on a à sa disposition une suite $Y_1, ...Y_n$ de v.a. indépendantes de loi commune identique à celle de Y.

La majeure partie des estimateurs connus d'une certaine quantité θ^* sont obtenus, à quelques modifications près, par un principe d'empirisme :

(1) exprimer θ^* comme une fonction de la distribution de Y
(2) définir l'estimateur $\widehat{\theta}_n$ comme la valeur de cette fonction sur la distribution empirique.

Pour le premier exemple la fonction est « espérance de la puissance 4 » et l'on obtient

$$\widehat{\theta}_n = \frac{1}{n}\sum Y_k^4$$

et dans le second l'estimée sera le premier quartile des données, c'est-à-dire la valeur qui sépare les 25% plus petites données de 75% plus grandes.

La convergence de $\widehat{\theta}_n$ vers θ^* sera généralement une conséquence, plus ou moins directe, de la loi des grands nombres.

Dans un cadre plus général de données dépendantes, on a plutôt recours à des estimateurs du type « maximum de vraisemblance » que l'on ne considérera pas ici, mais qui souvent peuvent s'interpréter comme plus haut. Leur convergence est encore basée sur des versions de la loi des grands nombres.

IV.2 Quelques estimateurs. La loi des grands nombres

Moyenne. Soit une suite de variables aléatoires $Y_1, ...Y_n$ de même loi. On voudrait calculer leur espérance commune m et leur variance v. L'estimateur empirique est :

$$\widehat{m}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

$\widehat{m}$ est l'espérance de la variable aléatoire obtenue par tirage uniforme dans l'ensemble $\{Y_1, ...Y_n\}$. La loi des grands nombres assure que ces quantités convergent vers m :

$$\lim_n \widehat{m}_n = E[Y] = m.$$

Variance. De même, un estimateur de la variance est la variance empirique

$$\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \widehat{m}_n^2.$$

La loi des grands nombres assure que $\widehat{v}_n$ est convergent :

$$\lim_n \widehat{v}_n = \lim_n \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \lim_n \widehat{m}_n^2 = E[Y^2] - E[Y]^2 = v.$$

L'estimateur de l'écart-type est

$$\widehat{\sigma}_n = \sqrt{\widehat{v}_n}$$

Corrélation. Soit une suite de variables aléatoires $(Y_1, Y_1'), ...(Y_n, Y_n')$ de même loi. Comme précédemment, des estimateurs convergents de leur corrélation et de leur covariance sont

$$\widehat{c}_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_n)(Y_i' - \widehat{m}_n') = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i' - \widehat{m}_n\widehat{m}_n'$$

$$\widehat{r}_n = \frac{\widehat{c}_n}{\sqrt{\widehat{v}_n\widehat{v}_n'}}$$

où $\widehat{m}_n$, $\widehat{m}_n'$, $\widehat{v}_n$ et $\widehat{v}_n'$ sont les estimateurs définis précédemment de la moyenne et de la variance pour les deux lois. La convergence se montre de la même façon.

Fonction de répartition. De même, si l'on veut estimer la fonction de répartition en un point x, c'est-à-dire la probabilité que la variable Y soit inférieure à x, on fera

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \leqslant x} = \frac{1}{n}\ \#\{i :\ Y_i \leqslant x\}.$$

et

$$\lim_n \widehat{F}_n(x) = E[1_{Y \leqslant x}] = P(Y \leqslant x) = F(x).$$

Quantile. On cherche la plus petite valeur A qui n'est dépassée par y qu'avec probabilité disons 5%, c'est-à-dire la solution de

$$P(Y > A) = 5\%.$$

On suppose pour simplifier que la fonction de répartition de Y est continue. Comme $A = F^{-1}(0,95)$ un estimateur naturel est

$$\widehat{A}_n = F_n^{-1}(0,95).$$

C'est-à-dire que $\widehat{A}_n$ est simplement la valeur telle que 5% des données lui sont supérieures. Cet estimateur découle également de l'utilisation de la loi des grands nombres, mais cette fois-ci de manière indirecte.

Probabilité d'un Bernoulli. Si chaque Y_i est $\mathcal{B}(p,1)$, c'est-à-dire vaut 1 avec probabilité p et 0 avec probabilité $1-p$, alors p est l'espérance de Y et une estimée est

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Définition 20. *Soit $Y_1, Y_2, ...Y_n$ une suite i.i.d. et θ^* une quantité dépendant de la distribution commune aux Y_i (une fonction des moments, un quantile,...). Soit $\widehat{\theta} = \widehat{\theta}(Y_1, ...Y_n)$ une fonction de $Y_1, ..Y_n$. On dit que $\widehat{\theta}$ est un estimateur non-biaisé de θ^* si*

$$E[\widehat{\theta}] = \theta^*.$$

On dit que la suite $\widehat{\theta}_n$ est un estimateur fortement convergent (ou consistant) de θ^* si avec probabilité 1

$$\lim_{n\to\infty} \widehat{\theta}_n = \theta^*.$$

Les estimateurs qu'on a vu jusqu'à présent sont tous convergents. L'estimateur $\widehat{m}_n$ est sans biais. En revanche, on verra dans l'exercice 1 que $\widehat{v}_n$ est un estimateur biaisé de v.

IV.3 Loi asymptotique des estimateurs

Les estimateurs sont des variables aléatoires, puisque ce sont des fonctions des observations. Il sont construits de sorte à avoir une faible variance autour d'une valeur à trouver. On peut étudier cela de plus près.

IV.3.1 Normalité asymptotique

On va voir que les estimateurs étant généralement basés sur des moyennes empiriques (parfois de manière indirecte), leur comportement asymptotique est essentiellement gouverné par le théorème-limite central et leur distribution asymptotique, après normalisation sera gaussienne. On aura donc très souvent une vitesse de convergence en $n^{-1/2}$ avec la limite en loi

$$\sqrt{n}\,(\widehat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, \sigma^2),$$

pour un σ à calculer.

Estimateur de la moyenne. En vertu du théorème-limite central,

- $\sqrt{n}\,(\widehat{m}_n - m) \longrightarrow \mathcal{N}(0, v)$,

est asymptotiquement gaussien de variance v (variance de chaque Y_i).

Estimateur de la variance. La distribution asymptotique de $\widehat{v}_n$ est plus difficile à obtenir puisque l'expression donnée plus haut pour $\widehat{v}_n$ ne se présente pas comme une moyenne de v.a. indépendantes. Comme $\widehat{v}_n$ est aussi la variance empirique de la suite $(Y_i - m)$ on a

$$\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 - (\widehat{m}_n - m)^2.$$

Il s'ensuit que

$$\sqrt{n}(\widehat{v}_n - v) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i - m)^2 - v] - \sqrt{n}(\widehat{m}_n - m)^2.$$

En raison de ce qui précède, le deuxième terme tend vers 0 et le premier converge en loi vers une variable gaussienne de variance :

$$v_e = E[((Y_i - m)^2 - v)^2] = E[(Y_i - m)^4] - v^2.$$

et donc

- $\sqrt{n}\ (\widehat{v}_n - v) \longrightarrow \mathcal{N}(0, v_e)$.

Estimateur de la covariance et de la corrélation. Par une méthode analogue on montre que **si les variables Y et Y' sont indépendantes** on a

- $\sqrt{n}\ \widehat{c}_n \longrightarrow \mathcal{N}(0, \sigma^2\sigma'^2)$, et $\sqrt{n}\ \widehat{r}_n \longrightarrow \mathcal{N}(0, 1)$.

S'il y a dépendance, on trouve des formules plus compliquées.

Bernoulli. Le théorème-limite central implique que

- $\sqrt{n}(\widehat{p}_n - p) \longrightarrow \mathcal{N}(0, p(1-p))$.

Fonction de répartition. De la même façon, puisque $1_{Y_i<x}$ est un Bernoulli

- $\sqrt{n}(F_n(x) - F(x)) \longrightarrow \mathcal{N}(0, F(x)(1 - F(x)))$.

Médiane. Soit m la médiane de la loi commune aux Y_i, et $\widehat{m}_n \simeq \check{Y}_{n/2}$ la médiane de l'échantillon ; on peut montrer que

- $\sqrt{n}(\widehat{m}_n - m) \longrightarrow \mathcal{N}(0, 1/(4f(m)^2))$.

où f est la densité de Y.

IV.3.2 Théorème de Kolmogorov

On a vu que la construction des estimateurs résultait souvent du remplacement de F (fonction de répartition de Y) par la fonction de répartition empirique F_n. Il existe deux théorèmes qui quantifient la proximité de ces deux fonctions. On posera

$$d_n = \sup_x |F_n(x) - F(x)|.$$

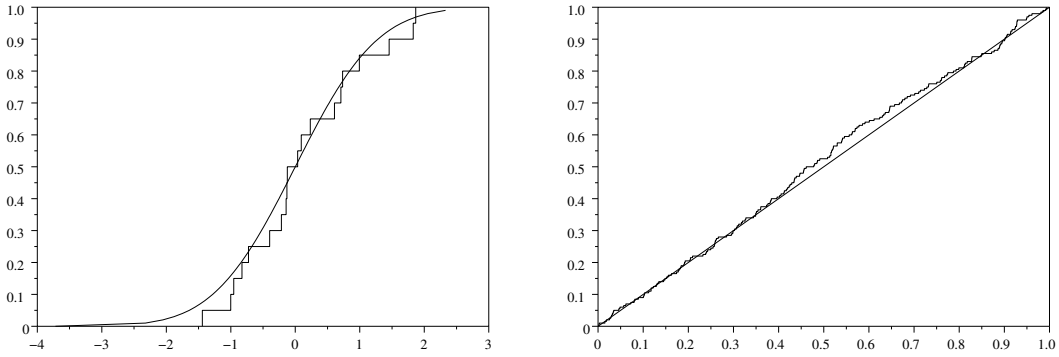


FIGURE IV.1 – Exemple de fonction de répartition empirique d'un échantillon de 20 valeurs gausiennes centrées réduites et fonction de répartition de la gaussienne. Même expérience avec 200 v.a. uniformes.

Théorème 21. *(Glivenko-Cantelli) Avec probabilité 1, d_n converge vers 0.*

La figure IV.1 illustre ce phénomène. Le deuxième est une convergence en loi

Théorème 22. *(Kolmogorov) Si $F(x)$ est continue, la loi de d_n est indépendante de F et la suite $\sqrt{n}d_n$ converge en loi :*

$$\lim_{n\to\infty} P(\sqrt{n}d_n < y) = 1 + 2\sum_{k=1}^{+\infty}(-1)^k e^{-2k^2y^2}.$$

Le fait que la loi de d_n est indépendante de la loi des Y_k est un fait simple à vérifier si F est strictement croissante :

$$\begin{aligned} d_n &= \sup_y |F_n(F^{-1}(y)) - y|, \quad \text{on a posé } F(x) = y \\ &= \sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{Y_k \leqslant F^{-1}(y)} - y| \\ &= \sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{F(Y_k) \leqslant y} - y| \end{aligned}$$

comme les $F(Y_k)$ sont i.i.d $\mathcal{U}([0,1])$ la loi de d_n est celle de

$$\sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{U_k \leqslant y} - y|$$

où les U_k sont i.i.d $\mathcal{U}([0,1])$.

IV.4 Intervalles de confiance

IV.4.1 Introduction. Définition

Plaçons nous dans le cas où l'on a la convergence en loi de l'estimateur $\widehat{\theta}$:

$$\frac{\sqrt{n}}{\sigma}(\widehat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0,1), \quad \text{en loi.} \tag{IV.1}$$

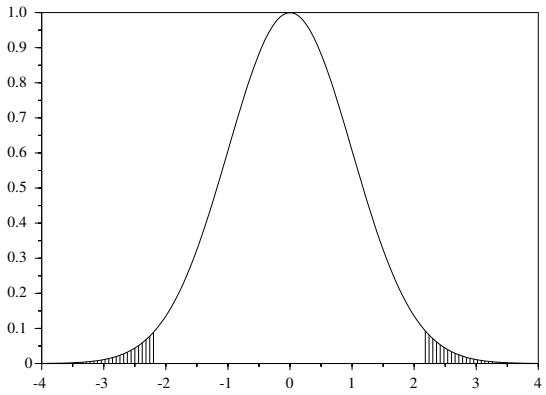


FIGURE IV.2 – Densité de la gaussienne. Chaque région hachurée est d'intégrale $\alpha/2$. Les abscisses correspondantes $(\pm 2,2)$ sont $\pm M_\alpha$. La probabilité de tomber dans la région non-hachurée est $1-\alpha$.

Notons M_α la valeur telle que la variable gaussienne tombe dans l'intervalle $[-M_\alpha; M_\alpha]$ avec probabilité $1-\alpha$ (cf figure IV.2) :

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{-M_\alpha} e^{-x^2/2} dx + \frac{1}{\sqrt{2\pi}} \int_{M_\alpha}^{+\infty} e^{-x^2/2} dx = 2 \frac{1}{\sqrt{2\pi}} \int_{M_\alpha}^{+\infty} e^{-x^2/2} dx.$$

Si l'on considère que $\sigma^{-1}\sqrt{n}(\widehat{\theta}_n - \theta^*)$ a une loi gaussienne, alors cette variable est comprise entre $-M_\alpha$ et M_α avec probabilité α. On a donc la relation

$$-M_\alpha \leqslant \frac{\sqrt{n}}{\sigma} (\widehat{\theta}_n - \theta^*) \leqslant M_\alpha \quad \text{avec probabilité } 1-\alpha$$

qui se réécrit

$$\theta^* \in \left[\widehat{\theta}_n - \frac{\sigma M_\alpha}{\sqrt{n}}, \widehat{\theta}_n + \frac{\sigma M_\alpha}{\sqrt{n}} \right], \quad \text{avec probabilité } 1-\alpha.$$

Intervalle à 99%. Pour un seuil $\alpha = 1\%$, on a $M_\alpha = 2,6$ $(P(|\mathcal{N}(0,1)| > 2,6) \simeq 0,01)$

$$\theta^* \in [\widehat{\theta}_n - 2,6\sigma/\sqrt{n}, \widehat{\theta}_n + 2,6\sigma/\sqrt{n}] \quad \text{avec probabilité } 0,99.$$

Cet intervalle de confiance est asymptotique, c'est-à-dire valide en théorie pour n grand seulement. On peut noter plus simplement

$$\boxed{\theta^* = \widehat{\theta}_n \pm 2,6\ \sigma/\sqrt{n} \quad \text{avec probabilité de confiance } 99\%.}$$

Intervalle à 95%. De la même façon, on a $P(|\mathcal{N}(0,1)| > 1,96) \simeq 0,05)$ et

$$\boxed{\theta^* = \widehat{\theta}_n \pm 1,96\ \sigma/\sqrt{n} \quad \text{avec probabilité de confiance } 95\%.}$$

On a la définition plus générale :

Définition 23. *Un intervalle de confiance pour θ^* de probabilité de confiance $1-\alpha$ est un intervalle aléatoire tel que*

$$\theta^* \in I \quad \textit{avec probabilité } 1-\alpha.$$

α est appelé le niveau.

Un peu de philosophie. Noter que l'équation (IV.1) conduit plus naturellement à

$$\widehat{\theta}_n = \theta^* \pm 1,96\ \sigma/\sqrt{n} \quad \text{avec probabilité } 95\%$$

qui est un intervalle déterministe pour $\widehat{\theta}_n$. C'est le point de vue des probabilités. Le renversement de cette équation correspond au passage des probabilités aux statistiques expliqué au chapitre 1.

IV.4.2 Intervalles exacts et intervalles approchés.

Les intervalles de confiances présentés au début du paragraphe précédent ne sont pas exacts car $\widehat{\theta}_n - \theta^*$ ne suit pas rigoureusement une loi normale ; ils ont donc un niveau α_n différent de α ; on peut cependant dire que α_n tend vers α.

Malheureusement, on ne connaît pas la valeur de n pour laquelle on peut considérer ces approximations comme raisonnablement valide. C'est pourquoi il est déraisonnable de considérer des intervalles asymptotiques à probabilité de confiance très élevée (disons 99%) pour des n petits (disons 10).

IV.4.3 Un exemple d'intervalle exact

Soient $Y_1, ...Y_n$ des v.a. gaussiennes de variance connue σ^2 et de moyenne inconnue μ^*. On sait que leur moyenne empirique $\widehat{\mu}_n$ a pour loi $\mathcal{N}(\mu^*, \sigma^2/n)$. La variable $\sqrt{n}(\mu_n - \mu^*)/\sigma$ suit donc une loi $\mathcal{N}(0,1)$. Donc

$$-2,6 \leqslant \frac{\sqrt{n}}{\sigma}\ (\widehat{\mu}_n - \mu^*) \leqslant 2,6 \quad \text{avec probabilité } 99\%$$

ce qui se réécrit

$$\mu^* = \widehat{\mu}_n \pm 2,6\ \frac{\sigma}{\sqrt{n}} \quad \text{avec probabilité de confiance } 99\%.$$

IV.4.4 Exemples d'intervalles approchés

Estimation de la moyenne. On observe 10 malades traités par un nouveau médicament. Pour chacun des malades le temps de guérison a été en jours ;

T :	12	16	21	10	13	16	25	8	13	15

On voudrait savoir le temps moyen de guérison. La moyenne empirique est $\bar{T} = 14,9$ et la variance empirique 23, soit un écart type d'environ 4,8. La variance de l'estimateur de la moyenne étant égale à la variance de la variable elle-même divisée par le nombre de points, on a

$$E[T] = 14,9 \pm 1,96\,.\,\sigma/\sqrt{10} \quad \text{avec probabilité de confiance } 95\%$$

ce qui donne en remplaçant σ par son estimée $4,8$:

$$E[T] = 14,9 \pm 3 \quad \text{avec probabilité de confiance } 95\%.$$

Estimation d'une proportion. On fait un sondage pour savoir qui de A ou B va gagner les élections. On obtient 1154 pour A et 1301 pour B et l'on suppose que l'échantillon est représentatif. La proportion p d'électeurs allant voter pour A s'estime à $\widehat{p} = 1154/2455 = 0,47$.

La variable $\sqrt{n}(\widehat{p}-p)$ est approximativement $\mathcal{N}(0,\sigma^2)$, avec $\sigma^2 = p(1-p) \simeq \widehat{p}(1-\widehat{p})$, d'où $\sigma/\sqrt{n} \simeq 0,01$. On a donc l'intervalle de confiance à 95% :

$$p = \widehat{p} \pm 1,96\,.\,\sqrt{\widehat{p}(1-\widehat{p})}/\sqrt{n} = 0,47 \pm 1,96\,.\,0,01 = 0,47 \pm 0,02 \quad \text{à } 95\%.$$

La victoire de B est quasi certaine (si tant est que l'échantillon est représentatif).

Le remplacement de σ par $\widehat{\sigma}$ est valide car il n'introduit qu'une erreur du deuxième ordre (c-à-d petite devant la largeur de l'intervalle).

IV.5 Tests de significativité

IV.5.1 Introduction

Commençons par un exemple volontairement simpliste. On veut tester si l'état de santé de certains malades s'améliore significativement à la suite d'un certain traitement. Pour cela on dispose de mesures de l'état de santé de 10 malades avant et après traitement

A	12	16	21	10	13	16	25	8	13	15
B	14	17	22	13	14	18	24	7	12	17

Il s'agit de tester l'hypothèse H_0 :« la variable $B-A$ a une moyenne nulle » (pas d'effet) contre son contraire H_1 qui assure d'un effet significatif du médicament. Il est clair qu'une priorité du test est de ne pas conclure H_1 si H_0 est vraie, ce qui entraînerait la mise sur le marché d'un médicament inefficace. Notons la dissysmétrie : la décision H_1 doit être convaincante car elle a des conséquences importantes.

Le test sera ici simplement une fonction des 20 variables aléatoires observées.

Définition 24. *Soient $Y_1, ...Y_n$ une suite de v.a. Un test est une statistique $\varphi(Y_1, ...Y_n)$ dont la valeur, 0 ou 1, décide entre deux hypothèses H_0 et H_1 portant sur la distribution de l'échantillon.*

En toute généralité, H_0 et H_1 correspondent donc à deux ensembles de distributions de probabilités disjoints ; par exemple (Y_i sont supposées i.i.d)

— $H_0 : Y_i \sim \mathcal{N}(0,1), \quad H_1 : Y_i \sim \mathcal{N}(2,1)$
— $H_0 : Y_i \sim \mathcal{N}(0,1), \quad H_1 : Y_i \sim \mathcal{N}(\mu,1), \quad \mu > 0$

Dans ces deux exemples, l'hypothèse H_0 est dite **simple** car elle détermine complètement la loi de Y, contrairement à l'hypothèse H_1 du deuxième exemple, qui est dite **composite**. Dans la suite, on s'intéressera essentiellement au cas où H_0 est simple.

On appelle **probabilité d'erreur de première espèce** ou **niveau du test**, la probabilité α de décider H_1 si H_0 est vraie, c'est-à-dire la valeur maximum de $E[\varphi]$ sous H_0 (ou la valeur tout court si H_0 est simple). Noter que le test qui décide toujours H_0 a un niveau faible mais ne présente aucun intérêt : sa probabilité d'erreur de seconde espèce (probabilité de décider H_0 sous H_1) est égale à 1.

Le but des tests de significativité est de déterminer si un ensemble de données permet d'invalider une hypothèse H_0. Dans l'exemple précédent, le laboratoire pharmaceutique cherchera à prouver qu'on a observé un effet *significatif* du nouveau traitement sur les malades, au sens où il est statistiquement très peu probable que H_0 soit valide.

Pour être fiable, un tel test devra avoir une très faible probabilité de décider H_1 si H_0 est vraie. On veut donc un petit niveau. Pour être intéressant, il devra être construit de sorte à décider H_1 le plus souvent possible quand H_1 est vraie (ceci est lié à la probabilité d'erreur de seconde espèce ; on dit que le test doit être puissant).

Fin de l'exemple. Notons m l'espérance de $B - A$, on a l'estimateur

$$\widehat{m} = \frac{1}{10} \sum_{i=1}^{10} B_i - A_i = 0,9.$$

Pour φ, on prend la v.a. qui vaut 0 si 0 est dans l'intervalle de confiance à 95% pour m basé sur $\widehat{m}$ et 1 sinon. Le niveau de ce test est de 5% (par définition de l'intervalle de confiance).

La variance empirique de $B - A$ est 1,89, on obtient donc un intervalle de confiance à 95%

$$m = 0,9 \pm 1,96 \sqrt{1,89/10} = 0,9 \pm 0,85$$

On peut donc décider d'un effet significatif sur cet ensemble de 10 patients, pour un niveau de 5%. On vérifie qu'il n'est toutefois pas significatif à 1%.

IV.5.2 Tests basés sur un estimateur et un intervalle de confiance

Soit θ^* un certain paramètre de la loi de Y. On veut tester $H_0 : \theta^* = \theta_0$ contre son contraire, c'est-à-dire voir si les données permettent d'affirmer si θ^* est sensiblement différent de θ_0.

Soit un estimateur $\widehat{\theta}$ de θ^* et $[\widehat{\theta} - \delta, \widehat{\theta} + \delta]$ un intervalle de confiance de probabilité de confiance $1 - \alpha$ pour θ^* :

$$\theta^* \in [\widehat{\theta} - \delta, \widehat{\theta} + \delta] \quad \text{avec probabilité } 1 - \alpha.$$

Considérons le test : Refuser H_0 si $\theta_0 \notin [\widehat{\theta} - \delta, \widehat{\theta} + \delta]$.

On sait que si H_0 est vraie $\theta^* = \theta_0$, le test décidera par erreur H_1 avec une probabilité ne dépassant pas α. Ce test a donc un niveau de α.

Noter que ce test est en fait

Refuser $\theta^* = \theta_0$ si $\lvert\widehat{\theta} - \theta_0\rvert > \delta$.

IV.5.3 Approche générale basée sur une statistique

On base en général les tests sur une statistique S que l'on juge pertinente pour distinguer au mieux les deux hypothèses (p. ex. $S = |\widehat{\theta} - \theta_0|$) ; par exemple S est plutôt petite sur H_0 et grande sous H_1 ; puis on se donne un seuil λ définissant le test

$$\varphi = 1_{S > \lambda}.$$

ou encore

Refuser H_0 si $S > \lambda$.

On tente de régler le seuil assez grand de sorte que la probabilité d'erreur de première espèce α ne dépasse pas une valeur prescrite :

$$P(S > \lambda) = \alpha, \quad \text{sous } H_0.$$

Toute valeur λ telle que $P(S > \lambda) \leqslant \alpha$ conviendrait, mais on refuserait H_0 moins souvent, ce qui est contraire à l'esprit du test.

Il faut interpréter la conclusion du test avec précaution :

— Si le test refuse H_0 (décide H_1), on peut dire que cette conclusion est fausse avec probabilité au plus α, par définition du niveau.

— Si le test décide H_0, on ne peut en général rien dire. Pour éviter de décider H_0 quand H_1 est vraie, il faut choisir une bonne statistique et avoir suffisamment d'échantillons.

Calcul du seuil. Si H_0 est simple, on connaît, au moins théoriquement, la valeur de λ telle que $P(S > \lambda) = \alpha$ sous H_0 ; cette valeur est la fonction quantile en $1-\alpha$. Si le calcul est trop difficile, on peut très bien l'estimer par simulation, en tirant sous H_0, par exemple 100000 réalisations indépendantes de S, et en l'estimant par la 95000^{e} valeur obtenue (par ordre croissant).

Cas où H_1 est non-H_0. Par exemple si H_0 est « $\theta^* = 0$ » et H_1 est « $\theta^* \neq 0$ » on conclura de la façon suivante :

— si le test refuse H_0 (décide H_1) : « θ^* est significativement (au niveau α) différent de 0 »

— si le test décide H_0 : « les données ne permettent pas de conclure que θ^* est significativement différent de 0 » Ceci peut arriver simplement parce que l'on a trop peu de données, ou parce que $\theta^* = 0$.

IV.5.4 Test de nullité d'une moyenne. Test de Student

On reprend l'exemple du paragraphe IV.5.1. Il s'agit de tester si une suite d'observations $Y_1, ...Y_n$ est issue d'une distribution de moyenne nulle Soit $\widehat{m}$ la moyenne empirique de l'échantillon :

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Supposons que Y_i a pour moyenne m et variance σ. On peut alors assimiler $\widehat{m}$ à une variable aléatoire gaussienne de moyenne m et de variance σ^2/n. Ceci implique l'intervalle de confiance (asymptotique)

$$-1,96 \leqslant \frac{\widehat{m} - m}{\sigma/\sqrt{n}} \leqslant 1,96 \quad \text{avec probabilité de confiance 95\%.}$$

Comme la variance est inconnue, on la remplace par son estimée empirique, et l'on obtient l'intervalle de confiance asymptotique :

$$m = \widehat{m} \pm 1,96\frac{\widehat{\sigma}}{\sqrt{n}} \quad \text{avec probabilité de confiance 95\%.}$$

D'où le test (de probabilité de confiance 95%) :

► Refuser la nullité de la moyenne si $\frac{\sqrt{n}|\widehat{m}|}{\widehat{\sigma}} > 1,96$.

On retrouve bien la forme annoncée au § IV.5.3.

Si les Y_i sont gaussiens $\mathcal{N}(0, \sigma^2)$ (hypothèse H_0 un peut particulière), la loi de la statistique $\frac{\sqrt{n}|\widehat{m}|}{\widehat{\sigma}}$ est bien entendu indépendante de σ ; il se trouve que c'est une loi de Student à $n-1$ degrés de libertés (elle dépend de n) ; on préfère souvent utiliser dans le test le quantile correspondant de cette loi plutôt que celui de la gaussienne (l'écart est souvent faible, par exemple si $n = 30$ 1,96 devient 2,04).

IV.5.5 Test d'identité de deux moyennes

On se donne deux échantillons de population d'origine différente et l'on voudrait décider si leur espérance de vie est différente. Soient $\widehat{m}_1$ et $\widehat{m}_2$ l'espérance de vie moyenne (empirique) dans chaque population :

$$\widehat{m}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i$$
$$\widehat{m}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Z_i$$

où Y_i et Z_i sont les durées de vie des individus sélectionnés dans chaque population. On suppose que Y_i et Z_i ont pour moyenne m_1 et m_2 et pour variance σ_1 et σ_2. On peut alors assimiler $\widehat{m}_1$ et $\widehat{m}_2$ à deux variables aléatoires gaussiennes indépendantes de moyenne m_1 et m_2 et de variance σ_1^2/n_1 et σ_2^2/n_2. Sous cette approximation, la variable $\widehat{m}_1 - \widehat{m}_2$ a pour moyenne $m_1 - m_2$ et pour variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Ceci implique l'intervalle de confiance (asymptotique)

$$-1,96 \leqslant \frac{\widehat{m}_1 - \widehat{m}_2 - m_1 + m_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \leqslant 1,96 \quad \text{avec probabilité de confiance 95\%.}$$

Comme les variances sont inconnues, on les remplace par leurs estimées empiriques, et l'on obtient l'intervalle de confiance asymptotique :

$$m_1 - m_2 = \widehat{m}_1 - \widehat{m}_2 \pm 1,96\sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}} \quad \text{avec probabilité de confiance 95\%.}$$

D'où le test (de probabilité de confiance 95%) qui décide de la différence des moyennes si zéro sort de l'intervalle de confiance :

► Refuser l'égalité des moyennes si $\dfrac{|\widehat{m}_1 - \widehat{m}_2|}{\sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}} > 1,96.$

On retrouve encore la forme annoncée au § IV.5.3. Dans le cas $n_1 = n_2 = n$ et $\sigma_1 = \sigma_2 = \sigma$, la statistique de test devient $\sqrt{n/2}|\widehat{m}_1 - \widehat{m}_2|/\widehat{\sigma}$; son interprétation est simple puisque c'est l'écart des moyennes empiriques normalisé par l'écart-type empirique, et par le facteur $\sqrt{n}$ du théorème-limite central.

IV.5.6 Test de comparaison de proportions

On veut tester l'efficacité d'un vaccin. Pour cela on se propose d'estimer si la probabilité d'attraper la maladie est inférieure si l'on a pris le vaccin. On considère deux populations, une non-vaccinée et une vaccinée. On est dans la situation du paragraphe précédent sauf que cette fois-ci Y_i est la variable aléatoire qui vaut 0 si l'on a pas été atteint et 1 sinon ; Z_i est la variable analogue observée sur la population vaccinée ; $m_i = p_i$ est la probabilité d'attraper la maladie et $\sigma_i^2 = p_i(1-p_i)$; on a donc le test à 95%

► Refuser l'identité des lois si $\dfrac{|\widehat{p}_1 - \widehat{p}_2|}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}} > 1,96.$ **Exemple : passagers du Titanic.**

On compare la probabilité d'être survivant entre les trois classes à partir des données suivantes :

	1-ière	2-ième	3-ième	Total
Survivants	193	119	138	450
Morts	129	161	573	863
Total	322	280	711	1313

On trouve les probabilités empiriques de survie $\widehat{p}_1 = 0,6$, $\widehat{p}_2 = 0,425$ et $\widehat{p}_3 = 0,2$. Les trois tests d'identités de loi ont pour statistique : $S_{12} = 4,35$, $S_{13} = 12,8$, et $S_{23} = 6,8$. Il y a bien une différence très significative.

Le risque relatif (RR) est le rapport p_1/p_2 des risques dans les deux populations. Le théorème limite central permet d'obtenir l'intervalle de confiance asymptotique à 95% suivant pour son logarithme

$$\log \frac{p_1}{p_2} = \log \frac{\widehat{p}_1}{\widehat{p}_2} \pm 1.96 \sqrt{\frac{1-\widehat{p}_1}{\widehat{p}_1 n_1} + \frac{1-\widehat{p}_2}{\widehat{p}_2 n_2}}$$

L'odds ratio (OR) est le rapport $OR = p_1(1-p_2)/(1-p_1)p_2$. Le théorème limite central permet d'obtenir l'intervalle de confiance asymptotique à 95% suivant pour son logarithme

$$\log(OR) = \log\left(\frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)}\right) \pm 1.96 \sqrt{\frac{1}{\widehat{p}_1 n_1} + \frac{1}{(1-\widehat{p}_1)n_1} + \frac{1}{\widehat{p}_2 n_2} + \frac{1}{(1-\widehat{p}_2)n_2}}$$

Parenthèse : OR et RR en biostatistiques. De manière générale l'OR est souvent préféré pour les raisons suivantes :

- Si l'on remplace l'évènement « survie » par l'évènement « décès » pour le calcul du RR, on obtient $\frac{1-p_1}{1-p_2}$ qui n'est pas fonction du RR de départ, tandis que l'OR est simplement remplacé par son inverse. Il y a donc en fait deux RR mais un seul OR.
- Lors des « études de cas témoins » (« case-control studies ») on tire d'abord au hasard un nombre équivalent de personnes guéries (ayant survécu...) et d'autres malades (décédées...) afin d'avoir suffisament d'individus dans les deux situations et ensuite on sépare chaque groupe en deux (traitement/non-traitement, classe1/classe2...). L'exemple suivant [1] concerne les accidents veineux thrombo-emboliques en Europe selon l'utilisation ou non de contraceptifs oraux où l'on a tiré au hasard 433 personnes ayant eu un accident veineux et 1044 n'en ayant pas eu

	Contraceptifs	Pas de contraceptifs	Total
Cas d'accident	265	168	433
Contrôles	356	688	1044
Total	621	856	1477

Cette proportion de 433/1044 ne reflète ici aucune la réalité ; on ne peut pas estimer p_1, qui n'a rien à voir avec 265/621, et pas davantage RR. En revanche 265/433 est bien une estimation de la probabilité d'utiliser un contraceptif sachant que l'on a eu un accident veineux, et de même pour les trois autres rapports analogues ; par conséquent si l'on remarque que par la formule de Bayes (A=accident, C=contraceptif, $\bar{A}$=non-A)

$$OR = \frac{P(A|C)P(\bar{A}|\bar{C})}{P(\bar{A}|C)P(A|\bar{C})} = \frac{P(A,C)P(\bar{A},\bar{C})}{P(\bar{A},C)P(A,\bar{C})} = \frac{P(C|A)P(\bar{C}|\bar{A})}{P(C|\bar{A})P(\bar{C}|A)}$$

l'OR est correctement estimé par $265 \times 688/(356 \times 168) \simeq 3$.

1. Table 3 de l'article : "Venous thromboembolic disease and combined oral contraceptives", *The Lancet*, pp. 1575-1582, 1995

IV.5.7 Test de corrélations

La loi asymptotique montrée plus haut dans le cas (hypothèse H_0) où les deux variables sont indépendantes conduit au test à 95%

- Refuser l'indépendance si $\sqrt{n}\ |\widehat{r}_n| > 1,96$

Application : Test de dépendance de deux Bernoullis. On voudrait savoir s'il existe chez les couples une corrélation entre le fait de posséder un animal domestique et ne pas avoir d'enfant. On mesure les deux variables U_i, qui vaut 1 si le couple numéro i a un animal domestique, et V_i qui vaut 1 si le couple numéro i a au moins un enfant. On a ici (cf la formule II.1)

$$\widehat{r}_n = \frac{\widehat{p}_{ae} - \widehat{p}_a\widehat{p}_e}{\sqrt{(1-\widehat{p}_a)\widehat{p}_a(1-\widehat{p}_e)\widehat{p}_e}}$$

où $\widehat{p}_a$ (resp. $\widehat{p}_e$, $\widehat{p}_{ae}$) est la proportion de couples ayant un animal (resp. un enfant, un animal et un enfant).

IV.5.8 Un exemple

Voici le début du commentaire du docteur Serge Hercberg publié dans le Quotidien du Médecin (22 juin 2003) concernant l'étude Suvimax effectuée sur un échantillon de 13017 personnes (7876 femmes et 5141 hommes) visant à évaluer l'importance de la consommation d'antioxydants (fruits et légumes) sur les risques de cancer. Cette étude a duré 8 ans pendant lesquels 6481 ont reçu des vitamines et minéraux antioxydant tandis que 6536 ont reçu un placebo.

« Les résultats sont très significatifs. Ils montrent nettement que l'apport de vitamines et de minéraux antioxydants à doses nutritionnelles réduit le risque de cancers ainsi que la mortalité globale chez les hommes. Cette baisse du taux de cancers de 31 % est très importante puisque près d'un cancer sur trois est évité en moins de huit ans (124 dans le groupe placebo contre 88 dans le groupe antioxydants ; RR=0,69, IC 95%=0,53-0,91 ; p<0,008). La différence entre les deux groupes est retrouvée pour la plupart des localisations de cancers, principalement digestifs, ORL, respiratoires et cutanés. La randomisation permet d'affirmer que la réduction observée a bien été causée par les antioxydants. Le nombre de décès chez les hommes était moindre dans le groupe antioxydants (40) que dans le groupe placebo (63) (p<0,02). En revanche, cet effet n'a pas été retrouvé chez les femmes (171 cancers dans le groupe placebo et 179 dans le groupe antioxydants ; 35 décès dans le groupe placebo, 36 dans le groupe antioxydants).

Comment expliquer l'absence d'effet chez la femme ?

Très probablement par un meilleur état du statut nutritionnel en antioxydants des femmes (bêta-carotène et vitamine C). En effet, les hommes avaient au départ de l'étude des taux sanguins de bêta-carotène plus bas que les femmes. Celles-ci consomment davantage de fruits et légumes. Or les niveaux sanguins de bêta-carotène sont corrélés positivement avec la consommation de fruits et légumes (r = 0,20 ; p<0,001). Autrement dit, les petits consommateurs de fruits et légumes ont les niveaux sanguins les plus faibles et réciproquement. Les femmes de l'étude Suvimax, n'étant pas carencées au départ, n'ont pas eu de bénéfice à être supplémentées.

A-t-on trouvé un bénéfice sur le plan cardio-vasculaire ?

Non. Nous avons comptabilisé 134 cardiopathies ischémiques dans le groupe antioxydants et 137 dans le groupe placebo. Il n'y avait donc pas de différence entre les deux groupes. [...] »

La différence de risque entre les deux populations (placebo et non-placebo) est jugée significative. La mention p<0,008 signifie qu'il faudrait faire un test de niveau inférieur à 0,8% pour ne plus être significatif. RR est le risque relatif. On est donc certain à 95% d'une diminution du risque comprise en 9% et 47%.

De même r=0,2 est l'estimée de la corrélation, et p<0,001 signifie qu'un test à 0,1% refuse H_0 :« r=0 » (on trouve en fait un niveau limite bien plus petit).

Bien que certains sujets aient abandonnés ou aient été perdus de vue ce qui modifie les chiffres, on retrouve bien en gros les valeurs numériques annoncées pour les niveaux limites.

IV.5.9 Test du χ^2 d'indépendances de caractères

On peut s'intéresser à tester l'indépendance de deux variables prenant plus de deux modalités, p.ex. le tableau de contingence suivante associant la profession du père et le type d'études suivies par le fils :

	Droit	Sciences	Médecine	IUT	Total
Expl. Agric.	80	99	65	58	302
Patron	168	137	208	62	575
Cadre Sup.	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
Total	1029	962	1411	382	3784

La statistique de Pearson est

$$S = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(\widehat{p}_{ij} - \widehat{p}_i \widehat{q}_j)^2}{\widehat{p}_i \widehat{q}_j} \tag{IV.2}$$

où I (resp. J) est le nombre de modalités de la première (resp. deuxième) variable, et les termes chapeautés les estimées de probabilités correspondantes. Sous l'hypothèse H_0 d'indépendance, la loi de S est (asymptotiquement pour n grand) celle d'un $\chi^2_{(I-1)(J-1)}$. Dans notre exemple elle vaut 320 ce qui serait excessivement grand pour un χ^2_{12}.

Noter que si $I = J = 2$ on retrouve bien le test de corélation car les 4 termes ont même numérateur et

$$\frac{1}{pq} + \frac{1}{p(1-q)} + \frac{1}{(1-p)q} + \frac{1}{(1-p)(1-q)} = \frac{1}{pq(1-p)(1-q)}.$$

Mentionnons le V de Cramer qui n'est pas un mesure de significativité mais une *quantité descriptive* qui mesure d'écart à l'indépendance (la différence est essentielle : *sur un grand échantillon, l'écart peut être significatif mais faible*) :

$$V = \sqrt{\frac{S}{n \min(I-1, J-1)}}$$

Ce coefficient s'interprète comme une corrélation en valeur absolue (cf. le cas $I = J = 2$).

IV.5.10 Test du χ^2 : adéquation à une distribution discrète, comparaison

On veut savoir si un dé est pipé ou non. Il s'agit de voir si une variable discrète suit bien une distribution prescrite, ici la distribution uniforme $p_1 = p_2 = ... = p_6 = 1/6$. On fait un certain

nombre n de tirages et l'on construit la statistique

$$S = n \sum_{i=1}^{k} \frac{(\widehat{p}_i - p_i)^2}{p_i} \tag{IV.3}$$

où $\widehat{p}_i$ et la fréquence d'apparition de la i-ième modalité et k le nombre de modalités (6 dans notre exemple), et p_i sa probabilité d'apparition sous H_0.

On peut montrer que sous H_0 :« les p_i correspondent bien à la distribution des données », la loi de S est (asymptotiquement pour n grand) celle d'un χ^2_{k-1}, c'est-à-dire la somme des carrés de $k-1$ variables normales standard indépendantes. On a donc le test (asymptotique) de niveau α

► Refuser la loi (p_i) si $S > Q_{k-1}(1-\alpha)$

où $Q_k(\alpha)$ est le quantile d'ordre α, c-à-d le nombre tel que $P(\chi^2_k > Q_k(1-\alpha)) = \alpha$. Ces quantités sont facilement disponibles sur ordinateur, et l'on a par exemple pour des seuil à 1% et 5% :

$\alpha \backslash k$	1	2	3	4	5
5%	3,8	6	7,8	9,5	11
1%	6,6	9,2	11,3	13,3	15

Valeur de $Q_k(\alpha)$ pour deux α et k=1,...5.

On peut très bien sinon estimer le seuil par simulation, comme expliqué au §IV.5.3, ce qui permet d'avoir un test exact (non-asymptotique) :

1. Tirer 100 000 (par ex.) échantillons (de taille n) sous H_0 (i.e. sous la loi $(p_1, ...p_k)$)
2. En déduire les 100 000 valeurs de S correspondantes et les ordonner
3. $Q(\alpha)$ est la p-ième valeur, avec $p = 100\,000(1-\alpha)$.

Un exemple un peu plus compliqué : le test de Hardy-Weinberg. Un allèle est une version d'un gène. Dans une cellule diploïde, il y a deux allèles pour chaque gène : un allèle transmis par chaque parent.

Soit A et a, deux allèles de fréquence respectivement p et $q = 1-p$ dans la population. La loi de Hardy-Weinberg prévoit les fréquences suivantes pour les trois différents génotype :

— $p_0 = p^2$: la fréquence d'un génotype homozygote AA
— $p_1 = 2pq$: la fréquence d'un génotype hétérozygote Aa
— $p_2 = q^2$: la fréquence d'un génotype homozygote aa

Cette loi se base sur le modèle le plus simple de reproduction dans la population et est expérimentalement vérifiée. Pour voir si un gène intervient dans une certaine maladie, et plus précisément la présence d'un allèle particulier, une méthode consiste à sélectionner un certain nombre de malades et à regarder si la distribution du génotype satisfait la loi (hypothèse H_0) ou non (hypothèse H_1). Comme p n'est pas connu (à moins de l'estimer par une expérience antérieure), le test du χ^2 ne peut être appliqué tel quel.

Si $\widehat{p}_0, \widehat{p}_1, \widehat{p}_2$ sont les proportions d'individus observés dans la population malade pour les génotypes AA, Aa, aa, on peut estimer p sous H_0 par

$$\widehat{p} = \widehat{p}_0 + \widehat{p}_1/2, \quad \widehat{q} = 1 - \widehat{p}$$

et la statistique de test de Hardy-Weinberg est définie par

$$S_{HW} = n \left(\frac{(\widehat{p}_0 - \widehat{p}^2)^2}{\widehat{p}^2} + \frac{(\widehat{p}_1 - 2\widehat{p}\widehat{q})^2}{2\widehat{p}\widehat{q}} + \frac{(\widehat{p}_2 - \widehat{q}^2)^2}{\widehat{q}^2} \right).$$

On voit qu'elle s'inspire de la statistique du χ^2 (IV.3). On peut montrer que sous H_0 cette statistique suit asymptotiquement un χ^2_1 ce qui permet de réaliser des tests.

Comparaison de deux échantillons. Pour décider si deux échantillons ont même loi (p.ex. taux de réussite au bac dans deux lycées différents, p_i^j étant la probabilité pour un lycéen du lycée j d'être reçu avec mention j), on peut utiliser la statistique suivante

$$S' = S(n_1, \widehat{p}^1, \widehat{p}^{12}) + S(n_2, \widehat{p}^2, \widehat{p}^{12})$$

où $S(n, \widehat{p}, p)$ désigne la statistique (IV.3), $\widehat{p}^1$ (resp. $\widehat{p}^2$, $\widehat{p}^{12}$) est le vecteur de probabilité estimée sur la base du premier échantillon (resp. du deuxième, des deux) et n_i la taille de l'échantillon i. On comparera cette statistique à un χ^2_{k-1} ([1] p.387) :

► Refuser l'identité des lois si $S' > Q_{k-1}(1-\alpha)$

IV.5.11 Tests de Kolmogorov et Smirnov : adéquation à une distribution continue, comparaison

Test d'une loi. On veut tester si les échantillons sont tirés selon une loi de fonction de répartition continue donnée $F(x)$ (hypothèse H_0). La statistique de test est

$$S_n = \sqrt{n} \sup_x |F_n(x) - F(x)|.$$

Au vu des résultats du § IV.3.2, la loi de S_n sous H_0 est indépendante de F. Définissons $q_n(\alpha)$ comme le quantile de cette loi connue :

$$P(S_n > q_n(1-\alpha)) = \alpha \quad \text{(sous } H_0\text{)}.$$

On a par exemple $q_{20}(5\%) = 1,57$ pour $n = 20$ (ces quantités sont tabulées). On a alors, par exemple, le test de probabilité de confiance 95% pour $n = 20$:

► La loi est significativement différente de F si $S_{20} > 1,57$.

Le niveau de ce test est la probabilité d'observer, sous H_0, que $\sqrt{n}\sup_x |F_n(x) - F(x)| > 1,57$; c'est 5% en raison du choix du seuil.

Comparaison de deux échantillons. On utilise le même principe pour comparer deux échantillons $Y_1, ...Y_{n_1}$ et $Z_1, ...Z_{n_2}$ de fonction de répartition F et G continues. On s'intéresse à l'hypothèse H_0 :« $F = G$ ». Soient F_{n_1} et G_{n_2} les fonctions de répartition empirique des deux échantillons. La statistique est cette fois :

$$S = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_x |F_{n_1}(x) - G_{n_2}(x)|$$

où n_1 et n_2 sont les longueurs respectives des échantillons. En raisonnant sur le même type de principe, on a le test, ici à 95% :

► Les lois sont significativement différentes si $S > 1,36$

(1,36 est ici le seuil asymptotique pour n grand). Noter qu'on peut montrer comme au § IV.3.2 que la loi de S sous H_0 est également celle de

$$\sup_y |\frac{1}{n_1}\sum_{k=1}^{n_1} 1_{U_k \leqslant y} - \frac{1}{n_2}\sum_{k=1}^{n_2} 1_{V_k \leqslant y}|$$

où les U_k et les V_k sont toues i.i.d $\mathcal{U}([0,1])$. On peut aisément estimer le quantile d'ordre α de cette variable par simulation.

IV.6 Exercices

Exercice 1. Soient $X_1, ...X_n$ des variables aléatoires indépendantes et de même loi, admettant une espérance m et une variance σ^2. On pose

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad V = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

1. Quelles sont les limites de $\bar{X}$ et V quand n tend vers l'infini ?
2. Donner l'espérance et la variance de $\bar{X}$.
3. Calculer l'espérance de V (on calculera l'espérance de $\bar{X}^2$ en utlisant qu'on a déjà calculé la variance de $\bar{X}$).
4. Proposer un estimateur non-biaisé de σ^2.

Exercice 2. On lance trois fois une pièce en l'air. On note p la probabilité pour que cette pièce tombe sur pile. Donner un estimateur de p, et déterminer sa loi ainsi que sa variance.

Exercice 3. Un orage gronde. Cinq personnes calculent la durée qu'ils ont perçue entre l'éclair et le coup de tonnerre. Sachant que la vitesse du son est de 330 m/s (on suppose la vitesse de la lumière infinie) et que les observations sont

T (sec) : 2 1,5 2 2 2,5

estimer un intervalle de confiance à 95% (asymptotiquement) pour la distance à laquelle se trouve l'orage.

Exercice 4. On reprend l'exemple de l'estimation d'une proportion :

$$p = \widehat{p}_n \pm \frac{1,96}{\sqrt{n}}\sqrt{p(1-p)}.$$

La largeur de cet intervalle à 95% dépendant de p, on a vu la solution consistant à remplacer p par $\widehat{p}_n$ dans le membre de droite. Proposer une méthode évitant cette approximation (on commencera par réécrire l'équation ci-dessus en faisant intervenir $(p-\widehat{p}_n)^2$).

Exercice 5. On dispose de mesures suivantes de l'efficacité du vaccin contre la polio (essais du vaccin Salk, 1954) [2]

	Polio	Sain	Total
Vaccinés	33	200712	200745
Non-vaccinés	115	201114	201229
Total	148	401826	401974

Étudier l'efficacité du vaccin.

Exercice 6. On dispose du sondage suivant sur l'opinion des hommes et des femmes sur l'avortement (250 sondés) :

2. "The biggest public health experiment ever : The 1954 Field Trial of the Salk Poliomyelitis Vaccine", Paul Meier, in *Statistic : a guide to the unknown, Pacific Grove*, J.M. Tanur & al. edt. 1988. Disponible à l'adresse `http://www.math.luc.edu/ mgb/courses/s103h/MeierPolio.htm`

	Pour	Contre
Hommes	52	48
Femmes	95	55

Est-ce que les femmes sont significativement plus favorables que les hommes à l'avortement ?

Exercice 7. Cinq observations d'un phénomène conduisent aux mesures suivantes :

236,2 262,4 238 230,6 228,7

On assimile ces données à la réalisation d'un échantillon de cinq v.a. normales indépendantes de moyenne inconnue m et d'écart-type 10.

1. Donner la moyenne empirique de cet échantillon, ainsi qu'un intervalle de confiance pour m de probabilité de confiance 95%.
2. En déduire le résultat du test de l'hypothèse H_0 :« $m = 250$ » au niveau 5%
3. Que faire si l'on ne connaît pas l'écart-type ?

Exercice 8. On observe des échantillons $X_1, X_2, ...$ de la loi uniforme sur $[0, \theta^*]$, θ^* étant un paramètre positif inconnu. Les X_i sont indépendants. Soit la variable aléatoire $Z_n = \sup_{i \leqslant n} X_i$.

1. Rappeler la fonction de répartition de Z_n ainsi que sa densité.
2. Montrer que Z_n converge en probabilité vers θ^* quand n tend vers l'infini, c'est-à-dire que : $P(|Z_n - \theta^*| < \varepsilon) \to 1$ pour tout $\varepsilon > 0$. En déduire un estimateur $\widehat{\theta}_n$ de θ^*.
3. Soit α étant un réel compris entre 0 et 1 donné, trouver ε pour que $P(Z_n \in [\theta^* - \varepsilon, \theta^*]) = 1 - \alpha$.
4. En déduire intervalle de confiance de probabilité de confiance $1 - \alpha$ pour θ^*.

Exercice 9. On dispose d'une étrange pièce de monnaie dont la probabilité, inconnue, de tomber sur pile lorsqu'on la lance est p. L'expérience consiste à lancer la pièce n fois et l'on note $k = k(\omega)$ le nombre de pile observés.

On veut discriminer les deux hypothèses H_0 :« $p = 0,49$ » et H_1 :« $p = 0,51$ » avec une probabilité de se tromper inférieure à $2,5\%$, quelle que soit la vérité (H_0 ou H_1). Le but de l'exercice est de déterminer le nombre minimum de lancers de la pièce pour y parvenir.

1. Quel test simple proposeriez-vous pour décider entre H_0 et H_1, à partir des résultats des lancers, sachant que le but est d'avoir le moins de chances de se tromper, quelle que soit la vérité (penser à la symétrie du problème) ? Il reste à déterminer le niveau de confiance.
2. Donner $\widehat{p}$ et l'approximation gaussienne pour cette variable ; que peut-on dire des variances asymptotiques sous H_0 et sous H_1 ?
3. En déduire la valeur, pour ce test, de la probabilité de décider H_1 si H_0 est vrai. On notera F la fonction de répartition de la variable normale centrée réduite.
4. Que vaut, pour ce test la probabilité de décider H_0 si H_1 est vrai ?
5. Que valent $P(\mathcal{N}(0,1) > 1,96)$ et $P(\mathcal{N}(0,1) < -1,96)$?
6. À quelle condition sur n la probabilité de se tromper est-elle inférieure à $2,5\%$?

Exercice 10. Calibration d'une diode Zener. Le modèle physique liant la tension x et l'intensité du courant y en sortie d'une diode Zener est

$$y = \theta_1(e^{\theta_2 x} - 1)$$

où θ_1 et θ_2 sont les paramètres de la diode.

On observe par des expériences des paires $(x_i, y_i)_{1 \leqslant i \leqslant n}$ où x_i est la tension et y_i l'intensité du courant, en sortie d'une diode. En raison des incertitudes de mesure, ces paires sont considérées comme des variables aléatoires et l'on définit θ^* comme le paramètre qui réalise au mieux l'adéquation au modèle physique de la diode :

$$\theta^* = \arg\min_{\theta} E\left[(y - \theta_1(e^{\theta_2 x} - 1))^2\right].$$

Donner les deux équations satisfaites par $\widehat{\theta_1}$ et $\widehat{\theta_2}$ obtenus par le principe décrit au § IV.1.

V

Analyse en composantes principales

V.1 Introduction

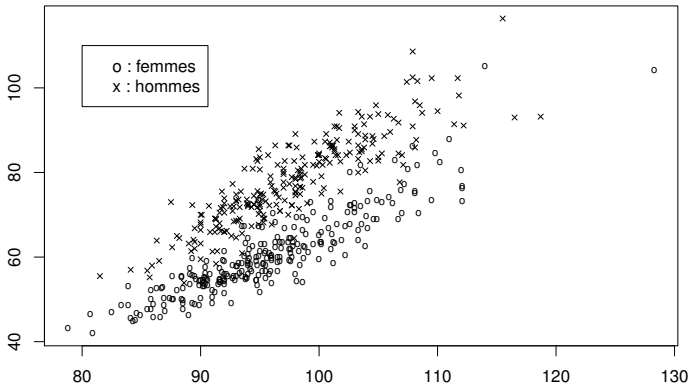
Soit un tableau de n individus et p variables, $n > p$

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}.$$

n est typiquement grand devant p. Chaque ligne x_i est un individu et chaque colonne x^k représente une variable. Chaque individu est un point de l'espace $\mathbb{R}^p$.

Premier exemple. x_k^1 est le tour de taille du sujet k et x_k^2 son poids [1]

tour de taille (cm)	poids (kg)
93.5	65.6
94.8	71.8
⋮	⋮
95.0	80.7



Dans tous les cas où il n'y a que deux variables seulement, la représentation des individus ne pose pas de problème : on les pose dans le plan, chaque axe représentant une variable ; on peut voir apparaître ainsi des liens entre les variables, ou bien des groupes d'individus. Dans cet exemple on a distingué les hommes et les femmes dans le tracé (il y a donc une troisième variable : le sexe) et l'on voit d'une part un lien entre ces deux variables (alignement approximatif des points) et également deux groupes distincts où se trouvent les femmes et les hommes respectivement.

1. G. Heinz, L.J. Peterson, R.W. Johnson et C.J. Kerk, «Exploring Relationships in Body Dimensions»' *Journal of Statistics Education Volume 11, Number 2 (2003).* www.amstat.org/publications/jse/v11n2/datasets.heinz.html

AUTRE EXEMPLE. Considérons comme exemple des données consistant en la composition de 45 poteries trouvées en Grande Bretagne datant de l'époque romaine [2]. La composition est la teneur en 9 composants (fer, manganèse....). Elles proviennent de 5 fours différents. La matrice X est la matrice 45×9 des compositions.

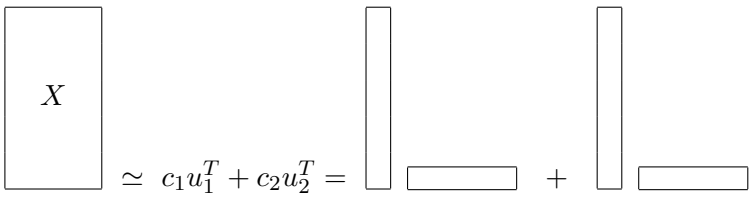
PAl_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	MnO	BaO	Four
18.8	9.52	2	0.79	0. 4	3.2	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0. 4	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0. 4	3.07	0.98	0.087	0.014	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015	5
19.1	1.64	0.6	0.1	0.03	1.75	1.04	0.007	0.018	5

Notons qu'en archéologie, l'analyse de la composition de matériaux est un outil important pour l'étude des échanges dans les économies antiques. Des objets d'origines distinctes ont généralement des signatures chimiques différentes qui permettent d'identifier leur origine. Pour identifier ces signatures il faut être capable de regrouper entre eux des objets de composition similaire.

Il y a maintenant bien trop de variables pour pouvoir représenter les individus aisément. On peut à la rigueur comparer deux individus, voir s'ils sont proches ou éloignés d'un individu donné en inspectant les valeurs numériques, mais il est impossible de faire apparaitre directement des groupements d'individus ou quelque structure particulière que ce soit.

L'objectif de l'ACP est de réaliser la démarche du premier exemple dans le cas de plus de deux variables, $p > 2$. L'idée est la suivante : Supposons dans un premier temps que les individus soient en fait concentrés dans un plan de $\mathbb{R}^p$, le bon sens veut alors que l'on représente directement les individus dans ce plan. Mathématiquement cela signifie que l'on fait un changement de base où les deux premiers axes sont dans le plan et les autres leurs sont orthogonaux ; les coordonnées (nouvelles variables) 3 à p seront donc nulles pour tous les individus.

Supposons maintenant que les individus soient tous proches d'un plan, l'idée est alors de trouver le plan le meilleur (au sens où la somme des distances des points au plan est la plus petite possible) et de représenter les données par le projection sur ce plan. Notons que la qualité recherchée pour cette représentation dans le plan est que les distances entre individus apparaissant avec cette représentation reflètent au mieux les distances entre individus dans $\mathbb{R}^p$. Le principe sera d'approcher X par une matrice de rang 2



où l'on voit que chaque individu se retrouve désormais dans le plan engendré par les *axes principaux* u_1 et u_2, qui seront normés à 1.

Si ce « meilleur plan » n'est en réalité pas très bon, on peut alors chercher le sous-espace de dimension trois le meilleur, mais la représentation des données sera plus difficile : en pratique on représentera plusieurs projections en dimension deux, mais l'interprétation commence à être délicate.

Soulignons dès à présent qu'une des difficultés majeures est l'incertitude liée à la normalisation

2. D'apres J.Holland Jones et I.G.Robertson, `www.stanford.edu/class/anthsci192`

des variables. En effet si par exemple $p = 3$ et que le nuage de point est en gros une sphère, on ne pourra pas considérer que les individus sont presque sur un plan ; en revanche une homothétie sur le troisième axe pourra les faire se concentrer sur cet axe ou au contraire sur le plan contenant les deux premiers.

L'ACP a un autre objectif que la représentation de données dans le plan, c'est la compression afin de pouvoir travailler sur des données de taille réduite (les comparer...), par exemple passer de $p = 10000$ (une image) à $p = 20$.

Les variables. L'ACP permet aussi de comprendre les liens entre variables. En effet, l'identité $X \simeq c_1 u_1^T + c_2 u_2^T$ exprime que chaque variable, colonne de X est approximativement combinaison linéaire de c_1 et c_2.

V.2 Approximation du nuage et décomposition en valeurs singulières

V.2.1 Rappels d'algèbre matricielle

Dans ce qui suit, on identifie toute application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^p$ à sa repésentation matricielle $M \in \mathbb{R}^{n\times p}$. De même un vecteur $x \in \mathbb{R}^p$ sera identifé à une matrice $p \times 1$, et éventuellement $1 \times p$ (ligne d'une matrice). On pourra écrire par exemple $\langle Mx, Ny\rangle = x^T M^T N y = Tr(Nyx^T M^T)$.

Une application P est une projection orthogonale si pour tout $x \in \mathbb{R}^n$, $P^2 x = Px$ et $\langle Px, (I - P)x\rangle = 0$. On vérifie facilement que ceci signifie que $P^2 = P = P^T$ (application idempotente et symétrique).

Toute matrice symétrique est diagonalisable dans une base orthogonale.

Le rang d'une application est la dimension de son espace image (espace engendré par les colonnes de la matrice). Si $M \in \mathbb{R}^{p\times n}$, son rang est inférieur à $n \wedge p$, et si son rang vaut cette valeur elle est dite de rang plein. Si P est une matrice de projection orthogonale, ses valeurs propres sont égales à 0 ou 1 et son rang vaut sa trace (une diagonalisation ne change ni le rang ni la trace).

Si $M \in \mathbb{R}^{n\times p}$, $n \geqslant p$, est de rang p, la matrice de projection orthogonale sur l'espace colonnes de M est $M(M^T M)^{-1} M^T \in \mathbb{R}^{n\times n}$.

Si l'on note c_j la j-ième colonne de M et l_i sa i-ième ligne, alors

$$MM^T = \sum_j c_j c_j^T, \qquad M^T M = \sum_i l_i^T l_i.$$

V.3 La décomposition en valeurs singulières

Le problème est donc de trouver l'espace affine de dimension k qui approche au mieux les individus, au sens où la projection orthogonale sur cet espace les déplace le moins possible. Pour avoir une analyse indépendante des effets de translation sur les variables, on travaillera sur les individus recentrés :

$$\tilde{x}_i = x_i - \bar{x}, \quad \tilde{X} = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}.$$

La moyenne empirique $\bar{x}$ est aussi appelée centre de gravité. On les standardise également souvent pour être invariant aux changements d'échelle, mais nous reviendrons sur cette question plus bas.

On suppose désormais le recentrage fait et l'on note X pour $\tilde{X}$. Ceci fait, on peut se restreindre désormais à une projection vectorielle.

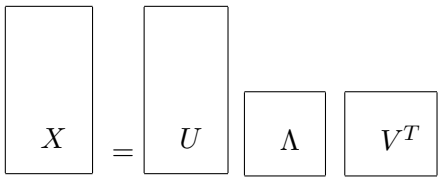
Il s'agit donc de trouver la matrice P_k de projection orthogonale sur un espace F_k de dimension k qui minimise $\sum_i \|x_i - P_k x_i\|^2$, soit

$$P_k = \arg \min_{\text{Rang}(P) \leqslant k} \|X - XP\|_F^2, \qquad F_k = \text{Im}(P_k). \tag{V.1}$$

où $\|.\|_F$ désigne la norme de Frobenius :

$$\|M\|_F^2 = \sum_{ij} M_{ij}^2 = Tr(M^T M).$$

La décomposition en valeurs singulières sera l'instrument central pour résoudre ce problème. Elle décompose une matrice X sous la forme



où U est à colonnes orthonormées ($U^T U = I$), Λ est diagonale, et V est orthogonale. La transposition de l'identité informelle ci-dessus donne la forme dans le cas où X est horizontalement allongée. Il existe des algorithmes numériquement très efficaces pour réaliser cette décomposition, c'est pourquoi son usage est recommandé en pratique.

Théorème 25. *Soit $X \in \mathbb{R}^{n \times p}$ une matrice, avec $n \geqslant p$. Il existe deux matrices à colonnes orthonormées U et V (i.e. $U^T U = V^T V = Id$) et une matrice diagonale $\Lambda \in \mathbb{R}^{p \times p}$ à entrées positives telles que*

$$X = U \Lambda V^T$$

Par conséquent, en appelant u_i et v_i les vecteurs colonne de U et V (vecteurs singuliers à droite et à gauche), on a la décomposition en somme de matrices de rang 1 :

$$X = \sum_{i=1}^{p} \lambda_i u_i v_i^T, \quad \lambda_i = \Lambda_{ii}.$$

La matrice Λ contient nécessairement les racines carrées des valeurs propres de $X^T X$, appelées valeurs singulières, et si ces dernières sont distinctes et rangées par ordre décroissant, cette décomposition est unique.

Démonstration. On ne traite que le cas où $X^T X$ a toutes ses valeurs propores non nulles. On peut diagonaliser $X^T X$:

$$X^T X = V D V^T.$$

On vérifie alors que $\Lambda = \sqrt{D}$ et $U = XV\Lambda^{-1}$ convient.

Pour l'unicité, noter que u_i (resp. de v_i) est le vecteur propre de XX^T (resp. $X^T X$) associé à λ_i^2. ∎

On peut maintenant donner la solution :

Théorème 26. *On suppose les valeurs singulières distinctes et ordonnées : $\lambda_1 \geqslant \lambda_2 \geqslant \dots \lambda_n$. Soit P_k*

$$P_k = \sum_{i \leqslant k} v_i v_i^T$$

la projection orthogonale sur l'espace engendré par les k plus premiers vecteurs singuliers à droite. Pour toute matrice de projection P orthogonale sur un espace de dimension $\leqslant k$, on a

$$\|X - XP_k\|_F \leqslant \|X - XP\|_F.$$

Remarque. Il y a unicité si et seulement si $\lambda_k = \lambda_{k+1}$. Nous ne détailons pas ce point.

Démonstration. En raison du théorème de Pythagore $\|X\|_F^2 = \|XP\|_F^2 + \|X(I-P)\|_F^2$; il suffit donc de montrer que $\|XP_k\|_F \geqslant \|XP\|_F$. En reprenant la relation $X^TX = VDV^T$ et

$$\|XP\|_F^2 = Tr(PVDV^TP) = Tr(DQ)$$

où $Q = V^TPV$ est un projecteur orthogonal de rang $\leqslant k$. Comme $Tr(Q^2) = Tr(Q) = k$, la somme de ses coefficients diagonaux est $\leqslant k$. Par ailleurs, comme $Q^TQ = Id$, le carré de la norme de la colonne $Q_{.i}$ vaut 1, en particulier $|Q_{ii}| \leqslant 1$. Donc

$$\|XP\|_F^2 = \sum_i D_i Q_{ii} \leqslant \max\Big\{ \sum_i D_i x_i : |x_i| \leqslant 1, \sum x_i \leqslant k \Big\} = \sum_{i=1}^k D_i.$$

C'est la valeur atteinte pour $P = P_k$. ∎

On a même plus généralement :

Théorème 27. *Pour toute matrice A de rang k on a*

$$\|X - XP_k\|_F \leqslant \|X - A\|_F.$$

Démonstration. En effet la matrice A s'écrit AP où P est le projecteur orthogonal sur l'orthogonal du noyau et donc, en utilisant le théorème de Pythagore (noter que $\|X\|_F^2 = Tr(X^TX) = Tr(XX^T)$) :

$$\|X - A\|_F^2 = \|X(Id-P) + (X-A)P\|_F^2 = \|X(Id-P)\|_F^2 + \|(X-A)P\|_F^2 \geqslant \|X - XP_k\|_F.$$ ∎

Corollaire 28. *L'espace F_k de dimension k solution de (V.1) est engendré par les k premiers vecteurs singuliers à droite $(v_1, \dots v_k)$; si des valeurs propres sont égales il n'y a pas nécessairement unicité.*

Exercice. Vérifier que $Xv_i = \lambda_i u_i$ et $X^Tu_i = \lambda_i v_i$. On suppose maintenant que les valeurs singulières sont distinctes. Vérifier que toute paire (u, v) telle que $Xv = \lambda u$ et $X^Tu = \lambda v$ pour un certain $\lambda > 0$ apparaît dans la décomposition (à un facteur près).

V.4 Inertie des espaces

Présentons un autre point de vue qui plutôt que de minimiser l'erreur $x_i - Px_i$ cherche à maximiser la dispersion des individus Px_i. On va voir que ceci revient au même.

On définit l'inertie des individus par la quantité

$$I = \frac{1}{n}\sum_i \|\tilde{x}_i\|^2.$$

L'inertie est donc également la somme des variances empiriques des p variables ; c'est encore n^{-1} fois le carré de la norme de Frobenius de $\tilde{X}$ (somme des carrées des coefficients). Cette quantité réelle mesure la **dispersion** des individus dans l'espace à p dimensions. Soit la matrice de covariance empirique des individus

$$R = \frac{1}{n}\sum_i \tilde{x}_i^T\tilde{x}_i = \frac{1}{n}\tilde{X}^T\tilde{X}, \quad R_{jk} = \frac{1}{n}\sum_i \tilde{x}_i^j\tilde{x}_i^k.$$

L'inertie est simplement la trace de R.

Soit E un sous-espace de $\mathbb{R}^p$ et P_E le projecteur orthogonal sur E ; on note I_E l'inertie des individus projetés, appellée l'inertie de E. :

$$I_E = \frac{1}{n}\sum_i \|P_E(\tilde{x}_i)\|^2 = \frac{1}{n}\sum_i \|\tilde{x}_i\|^2 - \frac{1}{n}\sum_i \|\tilde{x}_i - P_E(\tilde{x}_i)\|^2$$

(noter que les individus projetés puis recentrés sont aussi les projetés des individus recentrés). *L'inertie de E est donc également une mesure de la proximité entre les individus et E.*

Le problème (V.1) est donc équivalent à trouver l'espace F_k de dimension k d'inertie maximale. Comme on vient de le voir, cet espace est engendré par les k vecteurs propres associés aux k plus grandes valeurs propres de R. Ces espaces sont emboîtés, ce qui n'était pas évident au départ.

V.5 Propriétés fondamentales de l'ACP

Calcul pratique de l'ACP. (La matrice X est supposée déjà recentrée) Faire une SVD de X : $X = U\Lambda V^T$. V contient dans ses colonnes les axes principaux et $C = U\Lambda$ les composantes principales, soit pour la k-ième colonne $c_k = \lambda_k u_k$.

On a

$$X = CV^T = \sum_i c_i v_i^T$$

Comme $C = XV$, la i-ème composante principale est la combinaison linéaire des variables avec les poids contenus dans la i-ième colonne de V.

Définition 29. *Les v_k sont les facteurs principaux, ou axes pricipaux.*

Le vecteur $c_k = Xv_k$ est la k-ième composante principale. $\|c_k\| = \lambda_k$.

$\frac{\lambda_1^2+\ldots+\lambda_k^2}{\lambda_1^2+\ldots+\lambda_p^2} = \frac{I_{F_k}}{I}$ *est la fraction d'inertie expliquée par F_k.*

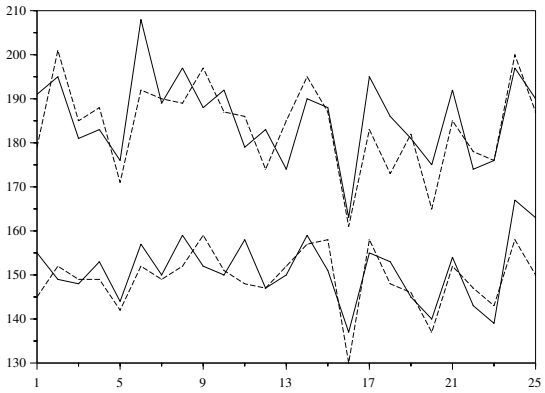


FIGURE V.1 – Longueur et largeur de la tête sur 25 couples de frères (données `frets{boot}` de R). Les données correspondant au plus vieux sont en traits plein.

Il résulte de ce qui précède que la matrice de covariance des c_i est la diagonale des λ_k.

Plus la fraction d'inertie expliquée par F_k est proche de 1, plus la projection des variables x_j sur F_k est proche de x_j, c'est-à-dire que les $c_1, ...c_k$ permettent de bien représenter les individus.

Un exemple. On a mesuré la largeur et la longueur de la tête chez des frères. Les variables sont donc au nombre de quatre dans l'ordre suivant : longueur et largeur pour l'un, longueur et largeur pour l'autre, notées $L1$, $l1$, $L2$, $l2$. Ce tableau de 25 individus est représenté sur la figure V.1. On obtient sur cet exemple

$$U = \begin{pmatrix} 0,57 & -0,69 & -0,44 & -0,01 \\ 0,41 & -0,22 & 0,87 & -0,17 \\ 0,60 & 0,63 & -0,21 & -0,44 \\ 0,39 & 0,27 & 0,06 & 0,88 \end{pmatrix}, \quad D = \Lambda^2 \begin{pmatrix} 228 & 0 & 0 & 0 \\ 0 & 29,4 & 0 & 0 \\ 0 & 0 & 17 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}$$

En première approximation, on peut dire que les composantes principales correspondent dans l'ordre à

- somme des périmètres des têtes
- différence entre les deux frères
- allongement de la tête du premier
- allongement de celle du second.

Le fait que la première valeur propre sorte du lot (80% d'inertie expliquée) indique que les individus se distinguent surtout par la somme des périmètres des crânes.

*** Une caractérisation.** La proposition suivante permet de caractériser c_k à une constante près comme vecteur qui maximise la covariance cumulée avec les variables :

Proposition 30. *La k-ième composante principale normalisée $u_k = c_k/\lambda_k$ est solution du problème de maximisation*

$$\max_u \sum_{j=1}^{p} \langle u, x^j \rangle^2, \qquad \textit{sous} \quad u \perp c_1, ...c_{k-1}, \quad \textit{et} \quad \|u\| = 1.$$

La valeur du maximum est λ_k^2.

REMARQUE. Les données étant centrées, ces produits scalaires sont donc, à facteurs constant près, des covariances empiriques. Si les x^j sont normés, ce qui est généralement le cas, ce sont même des corrélations.

Démonstration. Soit u la solution de ce problème, alors u est combinaison linéaire des x^j (en effet, si P est la projection orthogonale sur l'espace des x^j et $\|u\| = 1$, un remplacement de u par $Pu/\|Pu\|$ augmente les produits scalaires). u a donc la forme $u = Uz$ et la première contrainte signifie que z a ses $k-1$ premières composantes nulles. Le critère s'écrit (rappelons que $XX^T = UDU^T$ et $U^TU = I$)

$$\sum_{j=1}^{p} \langle u, x^j \rangle^2 = \sum_j u^T x^j x^{jT} u = u^T X X^T u = z^T D z.$$

La deuxième contrainte sur u est que $\|u_0\|^2 = 1$; les $k-1$ premières composantes de z étant nulles, on a nécessairement $z^T D z \leqslant \lambda_k^2 \|z\|^2 = \lambda_k^2$, et par ailleurs cette valeur est atteinte pour z égal au k-ième vecteur de la base canonique. Ce qui correspond bien à $u = u_k$. ∎

V.6 ACP sur données réduites

L'ACP n'est pas invariante par changement d'échelle sur les variables. Il est donc important, si les colonnes de X contiennent des données non comparables (i.e. des mètres et des kilogrammes) de les normaliser, afin d'avoir un résultat **indépendant des unités utilisées**. Si en revanche les colonnes sont comparables (comme celles de l'exemple des têtes du paragraphe précédent) on peut préférer de ne pas faire de normalisation : on considère ainsi que l'information de niveau relatif entre les différentes variables est importante, et les variables de faible amplitude seront pénalisées au sens où elles interviendront moins dans les premières composantes principales (comparer l'exemple numérique de ce paragraphe et celui du paragraphe précédent).

Sur les données du paragraphe précédent, on obtient un résultat analogue :

$$U = \begin{pmatrix} 0,49 & -0,48 & -0,72 & 0,07 \\ 0,49 & -0,54 & 0,68 & -0,08 \\ 0,51 & 0,50 & -0,05 & -0,70 \\ 0,51 & 0,48 & 0,09 & 0,71 \end{pmatrix}, \quad D = \Lambda^2 = \begin{pmatrix} 3,2 & 0 & 0 & 0 \\ 0 & 0,38 & 0 & 0 \\ 0 & 0 & 0,27 & 0 \\ 0 & 0 & 0 & 0,16 \end{pmatrix}.$$

U est cette fois très proche de la matrice

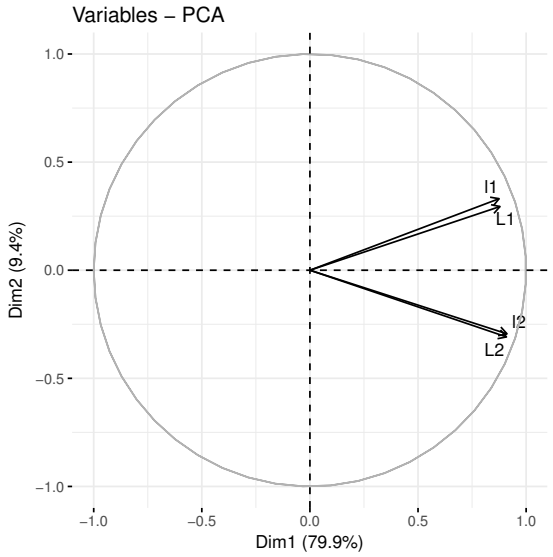
$$U = \frac{1}{2} \begin{pmatrix} 1 & -1 & -\sqrt{2} & 0 \\ 1 & -1 & \sqrt{2} & 0 \\ 1 & 1 & 0 & -\sqrt{2} \\ 1 & 1 & 0 & \sqrt{2} \end{pmatrix}.$$

Interprétation de Λ^2. Chaque terme correspond à l'inertie du sous-espace de dimension 1 correspondant, et leur somme fait l'inertie totale (théorème de Pythagore). Sa valeur mesure la contribution de l'axe aux données. Ici, le premier axe contribue à 80% de l'inertie car $3,2/(3,2 + 0,38 + 0,27 + 0,16) = 80\%$. Λ^2 contient les valeurs propres de la matrice de corrélation des données.

Le cercle des corrélations. Plutôt que de regarder les deux premières colonnes de la matrice V, il existe une représentation permettant d'étudier les liens entre variables d'origine et les nouvelles variables que sont les composantes principales : c'est la projection des variables normalisées sur le plan engendré par (c_1, c_2). Il se trouve que chacun de ces p points a pour coordonnées les corrélations avec les deux composantes. Ces paires de trouvent dans les deux premières colonnes de ΛV.

En effet, Ce plan a pour base orthonormée (u_1, u_2). La corrélation entre une des variables normalisée x et c_k, $k = 1, 2$, vaut $\langle x, \frac{c_k}{\|c_k\|} \rangle = \langle x, u_k \rangle$, qui n'est autre que la k-ième coordonnée du vecteur normé projeté sur ce plan. Ces produits scalaires se trouvent dans les deux premières colonnes de $U^T X = \Lambda V$.

On obtient pour les têtes :



Si l'ACP n'est pas normalisée, on trace quand même les corrélations. Ce tracé permet

- de voir les variables qui sont bien représentées par les deux premières composantes principales : le vecteur est de norme proche de 1
- d'interpréter chaque composante en observant quelles variables sont positivement ou négativement corrélées avec elle.

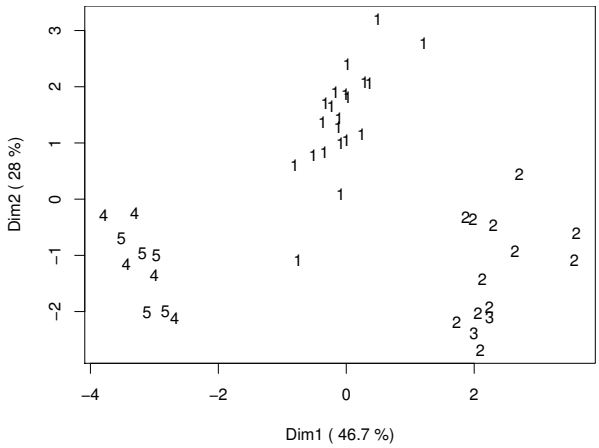
On voit ici une bonne représentation des variables, ce qui se confirme par une inertie totale de 89,3%. On voit également que la première composante mesure l'amplitude cumulée de toutes les variables et que la seconde oppose les paires (L1,l1) et (L2,l2), c'est-à-dire les deux frères.

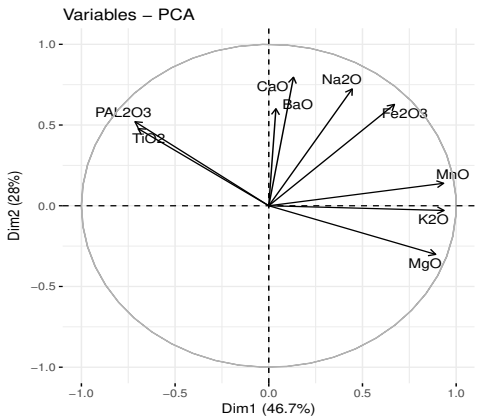
De plus, les deux proximités de vecteurs montrent que longueur et largeur sont très corrélés (Cette conclusion ne tient que du fait que les vecteurs sont de norme proche de 1 car sinon une corrélation des projection n'entraîne par une corrélation des vecteurs), mais il est plus simple de vérifier cela précisément par le calcul.

V.7 Représentations dans les plans principaux

Le tracé des individus dans le plan (v_1, v_2) peut faire apparaître des individus marginaux ou des classes bien séparées. Il s'agit simplement de représenter les point $(c_1(i), c_2(i))$. Cette représentation est d'autant meilleure que la fraction d'inertie expliquée par F_2 est grande. Si cette inertie est trop faible, on peut compléter par des repésentations dans le plan (v_1, v_3) ou (v_2, v_3).

Reprenons les données de poterie présentées dans l'introduction. On a tracé les individus dans le plan (v_1, v_2) en marquant le numéro du four. On voit nettement des regroupements qui permettent de classifier différents types de poterie (l'ACP est ici normalisée mais une ACP non normalisée donne des résultats similaires) :

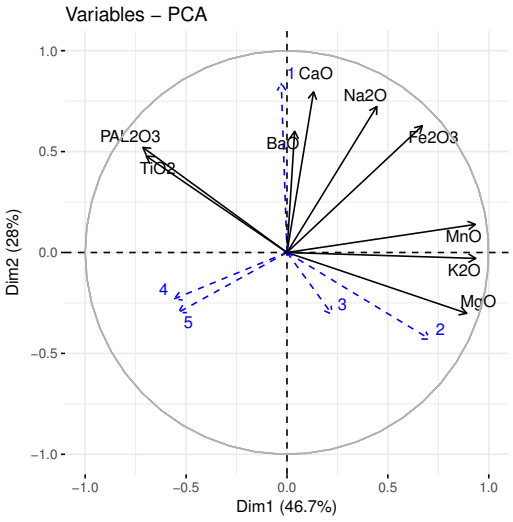




Les composantes ne séparent pas les fours 4 et 5, et un graphique non présenté ici montre que l'utilisation du troisième facteur n'améliore pas ce point. Noter que le four 3 n'a que deux représentants.

On voit sur le second graphique, projection des variables explicatives sur le plan des deux premières composantes principales, que la première composante oppose PaL_2O_3 et TiO_2 à MnO, K_2O, MgO, et Fe_2O_3. L'interprétation de la nature des facteurs principaux est ici affaire de chimistes.

On peut aussi représenter une variable $x \in \mathbb{R}^n$ n'ayant pas servi à l'analyse, dite « variable supplémentaire », toujours avec le calcul des corrélations $Cor(x, c_1)$ et $Cor(x, c_2)$. Ici nous avons placé les cinq variables indicatrices de four, ce qui illustre l'opposition 1, (2,3) et (4,5) :



V.8 Bilan

L'analyse en composantes principales propose donc de nouvelles variables explicatives $c_i = Xu_i$ (composantes principales). Ces variables sont choisies en sorte que la structure géométrique (distances entre individus) des individus restreints aux premières composantes soit la plus fidèle possible à la géométrie des individus complets, ou en d'autres termes en sorte que les premières variables concentrent l'information. Cette fidélité se mesure aux poids relatif des valeurs propres associées aux composantes conservées, les inerties. Ceci justifie la représentation des individus dans les premiers plans factoriels. Il est généralement préférable de travailler sur données réduites pour ne pas défavoriser a priori certaines variables. Le premier axe fait souvent apparaître un

effet « taille ».

Ces nouvelles variables sont décorrélées

La représentation des variables dans les plans factoriels peut s'interpréter comme le positionnement, pour chaque variable, d'un individu type, pour lequel les autres variables auraient leur valeur moyenne et la variable considérée serait amplifiée. Les logiciels permettent aussi d'y poser des variables n'ayant pas servi à l'analyse, dites « variables supplémentaires ».

La k-ième composante principale minimise la quantité $\sum_j \|x_j - c_k\|^2$ sous la contrainte d'être orthogonale aux précédentes. Les composantes principales c_i sont vecteurs propres de XX^T, les facteurs v_i sont vecteurs propres normés de X^TX. Dans la SVD de X : $X = U\Lambda V^T$, la matrice V contient les v_i en colonne, et $C = U\Lambda$ contient des c_i en colonne.

V.9 Exercices

Exercice 1. Montrer la relation

$$\frac{1}{n}\sum_i \|x_i\|^2 = I + \|\bar{x}\|^2.$$

Soit a un vecteur. En appliquant cette équation à $x_i - a$ au lieu de x_i retrouver la relation d'Huyghens, exprimant l'inertie par rapport à a en fonction de l'inertie et de la distance au centre de gravité.

Exercice 2. Montrer que I est la demi-moyenne des carrés des écarts entre individus :

$$I = \frac{1}{2n^2}\sum_{i,k} \|x_i - x_k\|^2.$$

Exercice 3. Utiliser la formule $I_E = \mathrm{Tr}(P_E R P_E)$ pour montrer que si E et F sont orthogonaux $I_{E\oplus F} = I_E + I_F$.

Bibliographie

[1] J.-J. Droesbeke, *Eléments statistique*, Ellipses, 1992.

[2] A. Morineau et A. Piron L. Lebart, *Statistique exploratoire multidimensionnelle*, Dunod, 1997.

[3] S. Morgenthaler, *Introduction à la statistique*, Enseignement des Mathématiques., Presses Polytechniques et Universitaires Romandes, 2007.

[4] G. Saporta, *Probabilité, analyse des données et statistiques*, Technip, 1990.

Cette bibliographie propose des livres dont l'objectif est proche de celui de ce cours. Il en existe de nombreux, ayant tous un contenu très analogue. Les livres choisis ici ont simplement l'avantage d'être disponibles à la bibliothèque universitaire de Rennes.