Exemples d'introduction à l'analyse des données:

De la statistique à la géométrie :

Soient X_1 et X_2 deux variables statistiques. Notons Y_1 et Y_2 les variables centrées construites à partir de X_1 et X_2 :

$$Y_1 = X_1 - \overline{X}_1 \qquad Y_2 = X_2 - \overline{X}_2$$

Convenons d'écrire les données brutes associées à chacune de ces variables, sous la forme de *n*-uplets :

$(Y_1(\omega_1),\ldots, Y_1(\omega_i),\ldots, Y_1(\omega_n))$ $(Y_2(\omega_1),\ldots, Y_2(\omega_i),\ldots, Y_2(\omega_n))$

ou, si l'on pose $y_{ij}= Y_j(\omega_i)$

$(y_{11},\ldots,y_{i1},\ldots,y_{n1})$ $(y_{12},\ldots,y_{i2},\ldots,y_{n2})$

Il est posible de considérer que ces n-uplets comme les composantes de deux vecteurs Y_1 et Y_2 éléments de IR^n (espace vectoriel de dimension n).

Exemple 1 :

Soit le tableau des données correspondant à deux variables statistiques X_1 et X_2 :

	X_1	X_2
ω_1	1	6
ω_2	3	2

On a

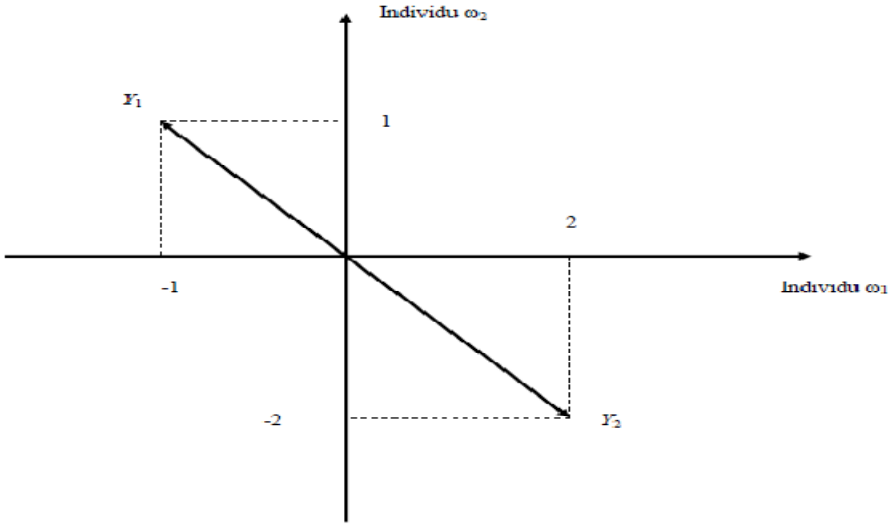
$$\overline{X}_1 = 2 \qquad \overline{X}_2 = 4$$

D'où les variables centrées Y_1 et Y_2 :

	Y_1	Y_2
ω_1	-1	2
ω_2	1	-2

On a bien $\overline{Y}_1 = 0$ $\overline{Y}_2 = 0$

Les vecteurs $Y_1(-1,1)$ et $Y_2(2,-2)$ de IR^2 ; le plan IR^2 est appelé espace des individus, car chaque axe du repère orthonormé est associé à un individu



$$\forall\ j=1,2 \qquad Var(Y_j)=\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\bar{Y}_j)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}(y_{ij})^2=\frac{1}{n}\|Y_j\|^2$$

et $\sigma(Y_j)=\dfrac{\|Y_j\|}{\sqrt{n}}$

où $\|Y_j\|=\sqrt{y_{1j}^2+\cdots+y_{ij}^2+\cdots+y_{nj}^2}$ est la norme euclidienne du vecteur Y_j, c'est-à-dire, dans un langage plus courant, la longueur du vecteur Y_j

Pour les vecteurs de l'exemple 1 ;

$$Var(Y_1)=\frac{1}{2}(1^2+1^2)=1 \qquad Var(Y_2)=\frac{1}{2}(4+4)=4$$

$$\|Y_1\|=\sqrt{1^2+1^2}=\sqrt{2} \qquad \|Y_2\|=\sqrt{4+4}=2\sqrt{2}$$

La longueur du vecteur associé à une variable statistique centrée est donc proportionnelle à l'écart-type de cette variable.

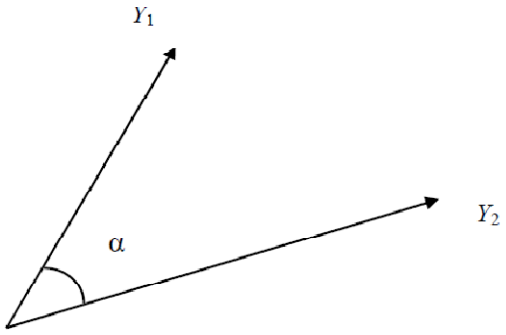
Calculons, maintenant, la covariance de (Y_1, Y_2) :

$$Cov(Y_1,Y_2)=\frac{1}{n}\sum_{i=1}^{n}(y_{i1}-\bar{Y}_1)(y_{i2}-\bar{Y}_2)$$

$$=\frac{1}{n}\sum_{i=1}^{n}y_{i1}y_{i2}=\frac{1}{n}(Y_1\cdot Y_2)$$

où (Y_1.Y_2) désigne le produit scalaire de Y_1 et de Y_2.

On rappelle que, si α est l'angle formé entre les vecteurs Y_1 et Y_2, alors :



$$\cos\alpha=\frac{Y_1\cdot Y_2}{\|Y_1\|\|Y_2\|} \qquad (1)$$

Soit encore, compte tenu de ce qui précède,

$$\cos\alpha=\frac{nCov(Y_1,Y_2)}{n\sigma(Y_1)\sigma(Y_2)}=\frac{Cov(Y_1,Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}=\rho$$

c'est le coefficient de corrélation linéaire entre Y_1 et Y_2.

Dans l'exemple 1 précédent, on peut voir clairement que les vecteurs Y_1 et Y_2 sont colinéaires et de sens contraire, l'angle de Y_1 et Y_2 est donc égal à π ; or $\cos \pi = -1$, résultat que l'on retrouve en utilisant la formule (1),

$$\cos\alpha = \frac{-2+(-2)}{\sqrt{2}\sqrt{8}} = -1$$

Lorsque les vecteurs sont linéairement dépendants (liés), il existe $\lambda \in IR_+^*$ tel que $Y_1=\lambda Y_2$, donc $\cos \alpha = \pm 1$ et réciproquement.
Quand on centre et on réduit des variables (par exemple $Z_j = \frac{Y_j - \overline{Y_j}}{\sigma(Y_j)}$), on forme des vecteurs qui ont tous la même dimension. ($Var(Z_j) = 1$)

De ce fait, la variance est la distance commune à tous les vecteurs (ils se situent sur un cercle de rayon 1) et ils se positionnent les uns par rapport aux autres par le coefficient de corrélation linéaire que l'on déduit à partir de l'angle formé par les deux vecteurs.

Exemple 2 :

Soit le tableau de données suivant :

Variables / Individus	X_1	X_2
1	4	5
2	6	7
3	8	0

$$X_1 = \begin{pmatrix} 4 \\ 6 \\ 8 \end{pmatrix} \quad \text{et} \quad X_2 = \begin{pmatrix} 5 \\ 7 \\ 0 \end{pmatrix},$$

Les moyennes : $\overline{X}_1 = \frac{8+6+4}{3} = 6$ et $\overline{X}_2 = \frac{5+7+0}{3} = 4$

Center les variables $(x_{ij} - \overline{x}_j)$:

4-6=-2 5-4=1
6-6=0 7-4=3
8-6=2 0-4=-4

Leur normes (écart-types σ_{Xi_i}) :

$$\|X_1\| = \sigma_{X1} = \sqrt{\frac{1}{3}[(-2)^2 + (2)^2]} = \sqrt{\frac{1}{3}[4+4]} = 2\sqrt{\frac{2}{3}}$$

et $$\|X_2\| = \sigma_{X2} = \sqrt{\frac{1}{3}[1^2 + 3^2 + (-4)^2]} = \sqrt{\frac{26}{3}}$$

$$Z = \begin{pmatrix} \frac{-2\sqrt{3}}{2\sqrt{2}} & \sqrt{\frac{3}{26}} \\ 0 & \frac{3\sqrt{3}}{\sqrt{26}} \\ \sqrt{\frac{3}{2}} & \frac{-4\sqrt{3}}{\sqrt{26}} \end{pmatrix} \quad \text{avec} \quad z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_{X_i}}$$

$\Rightarrow \overline{Z}_1 = \overline{C}_1^* = 0$ et $\overline{Z}_2 = \overline{C}_2^* = 0$

$\sigma_{yj} = 1; \quad j = 1,2$

De plus $r = \dfrac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$ entre deux variables *X* et *Y*,

mais si $\sigma(X) = \sigma(Y) = 1,$ alors $r = Cov(X,Y)$.

Calcul du produit matriciel $\dfrac{1}{p} Z'Z$:

$$\frac{1}{3}\begin{pmatrix} -\sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{26}} & \frac{3\sqrt{3}}{\sqrt{26}} & \frac{-4\sqrt{3}}{\sqrt{26}} \end{pmatrix}\begin{pmatrix} -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{26}} \\ 0 & \frac{3\sqrt{3}}{\sqrt{26}} \\ \sqrt{\frac{3}{2}} & \frac{-4\sqrt{3}}{\sqrt{26}} \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 3 & \frac{-15}{2\sqrt{13}} \\ \frac{-15}{2\sqrt{13}} & 3 \end{pmatrix} = \begin{pmatrix} 1 & \frac{-5}{2\sqrt{13}} \\ \frac{-5}{2\sqrt{13}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0{,}69 \\ -0{,}69 & 1 \end{pmatrix}$$

Le résultat de ce calcul est la matrice R des corrélations linéaires des variables. On l'appelle aussi la matrice d'information des variables.
R est une matrice symétrique, ayant des « 1 » sur la diagonale (les variances des variables) et tous ses éléments sont inférieurs ou égales à 1 en valeur absolue.

Calcul du produit matriciel ZZ' :

$$ZZ' = \begin{pmatrix} -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{26}} \\ 0 & \frac{3\sqrt{3}}{\sqrt{26}} \\ \sqrt{\frac{3}{2}} & \frac{-4\sqrt{3}}{\sqrt{26}} \end{pmatrix}\begin{pmatrix} -\sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{26}} & \frac{3\sqrt{3}}{\sqrt{26}} & \frac{-4\sqrt{3}}{\sqrt{26}} \end{pmatrix} = \begin{pmatrix} \frac{21}{13} & \frac{9}{26} & \frac{-51}{26} \\ \frac{9}{26} & \frac{27}{26} & \frac{-36}{26} \\ \frac{-51}{26} & \frac{-36}{26} & \frac{87}{26} \end{pmatrix} = V$$

Cette matrice V n'est pas une matrice de corrélation, mais elle porte le nom de matrice d'information des individus. Elle est symétrique aussi.