2 Géométrie - Statistique - Probabilité

Objectif : Montrer que les concepts statistiques, probabilistes s'expriment géométriquement.

2.1 Statistique et Géométrie

Soient $(x_1, x_2, \cdots, x_n)$, $(y_1, y_2, \cdots, y_n)$, $\cdots$ respectivement n observations indépendantes des variables X, Y, $\cdots$ faites sur n unités statistiques différentes appelées *individus*.

La statistique a pour but de répondre à des questions telles que :

- 1 - Quels nombres résument au mieux l'information véhiculée par les observations ? Que valent ces résumés ?
- 2 - Quelle est l'intensité de la relation qui lie les variables ? Comment peut-on la mesurer à partir des observations ?
- 3 - Quelle est la fonction linéaire qui permet d'expliquer (resp. prédire) au mieux une variable en fonction d'autres variables ? Quelle est la qualité de l'explication (resp. prédiction) ?

Pour répondre à ces questions on utilise le critère des *moindres carrés* exprimé d'un point de vue géométrique.

2.1.1 Résumer des observations

D'après le critère des moindres carrés, le meilleur résumé des n observations x_i, $i = 1, \cdots, n$ est le nombre α qui minimise :

$$\sum_{i=1}^{n}(x_i - \alpha)^2 \qquad (1)$$

Soit X le vecteur de coordonnées $(x_1,\ x_2,\ \cdots,\ x_n)$, J le vecteur de coordonnées $(1, 1, \cdots, 1)$. La formule (1) s'écrit comme un produit scalaire :

$$< X - \alpha J, X - \alpha J > = \ |X - \alpha J|^2$$

Le critère des moindres carrés signifie que parmi les résumés possibles, de la forme αJ, on choisit celui qui est le plus proche de X (*i.e.*, celui situé à la "plus petite distance" quand on utilise le produit scalaire Euclidien).

Le point de J le plus proche de X est le projeté (orthogonalement) de X sur J. On peut donc déterminer la valeur de α en explicitant la projection orthogonale au moyen du produit scalaire. En effet, soit $\hat{X}_J$ une projection quelconque de X sur J, alors :

- $\hat{X}_J = \alpha J$ pour tout α
- Parmi les projections de X sur J, il en existe une (la projection orthogonale notée $\hat{X}_J$) qui satisfait aux propriétés P_1 et P_2 du § 1.3 :

$$\hat{X}_J = \alpha J \qquad (2)$$
$$< X - \hat{X}_J, J > = \ 0 \qquad (3)$$

Si on substitue (2) dans (3), alors :

$$< X - \hat{X}_J, J > = < X - \alpha J, J > = \ 0 \iff < X, J > = \ \alpha < J, J >$$

En utilisant le produit scalaire Euclidien, on déduit que :

$$\sum_{i=1}^{n} x_i = \alpha n \iff \alpha = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

On peut donc conclure que la <u>*moyenne*</u> d'échantillonnage (notée $\bar{x}$) est le meilleur nombre qui résume les données. Sur la figure 2.1 est représentée la solution suggérée par la géométrie ainsi que la décomposition de X en deux composantes orthogonales : $\hat{X}_J = \bar{x}J = (\bar{x},\ \bar{x},\ \cdots,\ \bar{x})$ et l'erreur que l'on commet quand on résume les données par $\bar{x}$.

$$E = X - \bar{x}J = (x_1 - \bar{x},\ x_2 - \bar{x},\ \cdots,\ x_n - \bar{x})$$

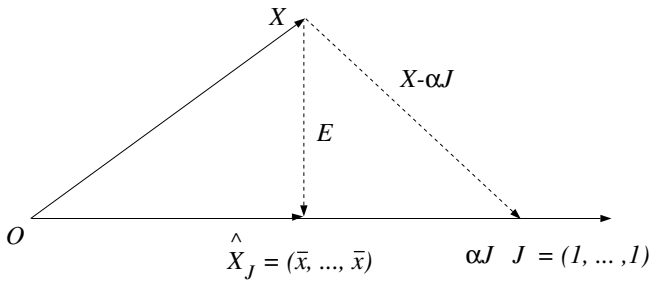

Figure 2.1 : Ecart à la moyenne.

Si toutes les observations étaient identiques, X et J seraient colinéaires, on aurait $|E|^2 = 0$ et X serait une constante . On interprète donc E comme la variation des composantes de X ($X = \bar{x}J + (X - \bar{x}J)$).

La décomposition correspondante des carrés des longueurs est obtenue en appliquant le théorème de Pythagore :

$$|X|^2 = |\bar{x}J|^2 + |X - \bar{x}J|^2 \iff \sum_{i=1}^{n} x_i^2 = n\bar{x}^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2$$

elle correspond à la décomposition classique de la somme des carrés.

$|\bar{x}J|^2$ est la *somme des carrés* expliquée par la moyenne d'échantillonnage, elle dépend du nombre n d'observations.

$|E|^2 = |X - \bar{x}J|^2$ est la *somme des carrés des erreurs* (SCE), elle dépend du nombre n d'observations.

Le vecteur $X = \bar{x}J + (X - \bar{x}J)$ est dans un espace vectoriel de dimension n. On dit que le vecteur $\bar{x}J$ est dans un *sous espace vectoriel* de dimension 1 (engendré par J) de l'espace vectoriel de dimension n car :

- J a toutes les propriétés d'un espace vectoriel (addition et multiplication par un scalaire).
- J est contenu dans l'espace de dimension n.

Le vecteur E quant à lui est dans un sous espace de dimension $n - 1$ de l'espace de dimension n car :

- E est dans le sous espace orthogonal à J (*complémentaire orthogonal*).
- La somme des dimensions de deux sous espaces complémentaires orthogonaux est égale à la dimension de l'espace vectoriel qui contient ces sous espaces.

La moyenne d'échantillonnage sera un bon résumé de l'information véhiculée par les observations si chaque composante du vecteur E est petite (proche de zéro). Pour apprécier cela, on estime la moyenne "par dimension" de la somme des carrés des erreurs en divisant SCE par $n-1$ (dimension du sous espace vectoriel qui contient E). On obtient ainsi le *carré moyen d'erreur* (CME) :

$$CME \;=\; \frac{|E|^2}{n-1} \;=\; \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

on reconnaît l'expression de la *variance* d'échantillonnage.

- Les dimensions des espaces sont appelées les *degrés de liberté*.

Illustration numérique

Soit le vecteur :

$$X \;=\; \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

des observations faites sur deux individus (*ind.1* et *ind.2*). Sur la figure 2.2 on a représenté les vecteurs X, J et $\hat{X}_J = \bar{x}J$.

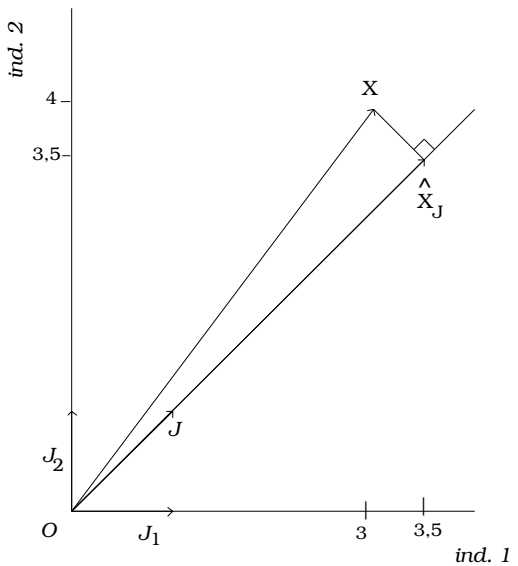

Figure 2.2 : Représentation des vecteurs observations, moyenne et erreur.

On veut vérifier que :

$$|X|^2 \;=\; |\hat{X}_J|^2 \;+\; |X \;-\; \hat{X}_J|^2$$

Pour cela, on calcule chaque élément de l'égalité précédente. Soit :

$$|X|^2 = < X, X > = 3^2 + 4^2 = 25$$

$$\hat{X}_J = \begin{bmatrix} 3,5 \\ 3,5 \end{bmatrix}$$

$$|\hat{X}_J|^2 = < \hat{X}_J, \hat{X}_J > = (3,5)^2 + (3,5)^2 = 24,5$$

$$|X - \hat{X}_J|^2 = < X - \hat{X}_J, X - \hat{X}_J > = (3-3,5)^2 + (4-3,5)^2 = 0,5$$

$$|\hat{X}_J|^2 + |X - \hat{X}_J|^2 = 24,5 + 0,5 = 25 = |X|^2$$

2.1.2 Mesurer des liaisons

Apprécier l'intensité de la relation qui lie deux variables X et Y consiste à mesurer l'influence des variations de X sur Y.

Ces variations liées de X et Y sont appréhendées à partir des composantes orthogonales à J ($X - \hat{X}_J = X - \bar{x}J$ et $Y - \hat{Y}_J = Y - \bar{y}J$), la mesure utilisée est le cosinus de l'angle θ entre ces composantes (figure 2.3). Cet angle θ est insensible à tout changement d'échelle ou translation sur X ou Y, son cosinus n'est autre que le coefficient de *corrélation simple* de X et Y.

Le produit scalaire de $X - \bar{x}J$ et $Y - \bar{y}J$ divisé par les degrés de liberté appropriés $(n-1)$ n'est autre que la *covariance* d'échantillonnage de X et Y.

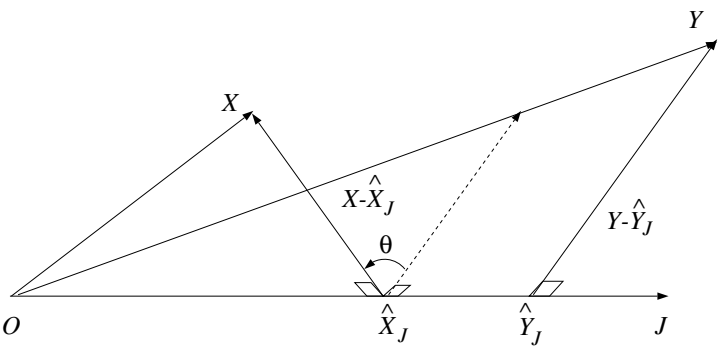

Figure 2.3 : Coefficient de corrélation simple.

D'un point de vue très général :

- les différentes mesures d'intensité des relations entre des variables X, Y, Z, $\cdots$ sont appréhendées dans le *sous espace vectoriel orthogonal* à J.
- Le sous espace vectoriel orthogonal à J est le sous espace vectoriel des *variables centrées*. Il permet de visualiser simplement les différents coefficients de corrélation (corrélations simple, multiple et partielle).

On note :

$X' = X - \bar{x}J$, $Y' = Y - \bar{y}J$ et $Z' = Z - \bar{z}J$ trois variables centrées.

$\hat{X}'_{Z'}$, $\hat{Y}'_{Z'}$ respectivement les projections de X' et Y' sur Z'.

$\hat{X}'_{Y'Z'}$ la projection de X' sur le plan déterminé par Y' et Z'.

La figure 2.4 visualise dans l'espace des variables centrées :

- Le coefficient de corrélation simple de X et Y. C'est le cosinus de l'angle que font les vecteurs X' et Y' ($cos(\theta)$).
- Le coefficient de corrélation multiple de X et Y, Z. C'est le coefficient de corrélation de X avec Y et Z considérés conjointement. C'est donc le cosinus de l'angle des vecteurs X' et $\hat{X}'_{Y'Z'}$ ($cos(\beta)$).
- Le coefficient de corrélation partielle de X et Y. C'est le coefficient de corrélation de X et Y quand l'effet de Z est éliminé. Le vecteur Z' étant orthogonal à $X' - \hat{X}'_{Z'}$ et à $\hat{X}'_{Y'Z'} - \hat{X}'_{Z'}$, le coefficient de corrélation partielle est le cosinus de l'angle que font ces deux derniers vecteurs ($cos(\alpha)$).

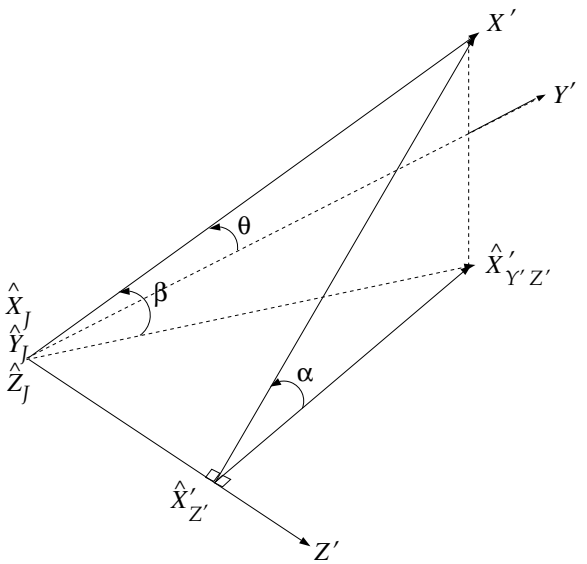

Figure 2.4 : Corrélations simple, multiple et partielle.

Illustration numérique

Soit la matrice :

$$X = \begin{bmatrix} 10 & 5 & 41 \\ 20 & 15 & 59 \\ 35 & 20 & 69 \end{bmatrix}$$

Les lignes représentent des individus (notés *ind.1, ind.2, ind.3*), les colonnes représentent des variables. De la première à la troisième colonne on a les variables pression en atmosphères (P), température en degrés centigrades (T_c) et température en degrés Fahrenheit (T_f). On veut vérifier graphiquement que :

- l'angle que fait T_c avec P est différent de l'angle que fait T_f avec P

-les angles que font les variables centrées T_c' et T_f' avec P' sont identiques.

Sur la figure 2.5 on a représenté les vecteurs T_c, T_f et P. On observe que T_c et T_f ne sont pas colinéaires.

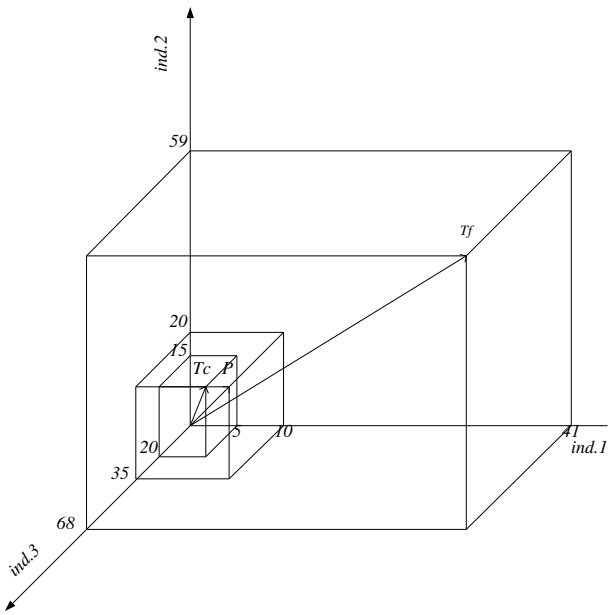

Figure 2.5 : Représentation de T_c, T_f et P.

Soient P', T_c' et T_f' les variables centrées. La matrice des observations centrées est :

$$X' = \begin{bmatrix} -11,7 & -8,4 & -15 \\ -1,7 & 1,7 & 3 \\ 13,4 & 6,7 & 12 \end{bmatrix}$$

Sur la figure 2.6 on a représenté les vecteurs des variables centr ées, on observe que T'_c et T'_f sont colinéaires.

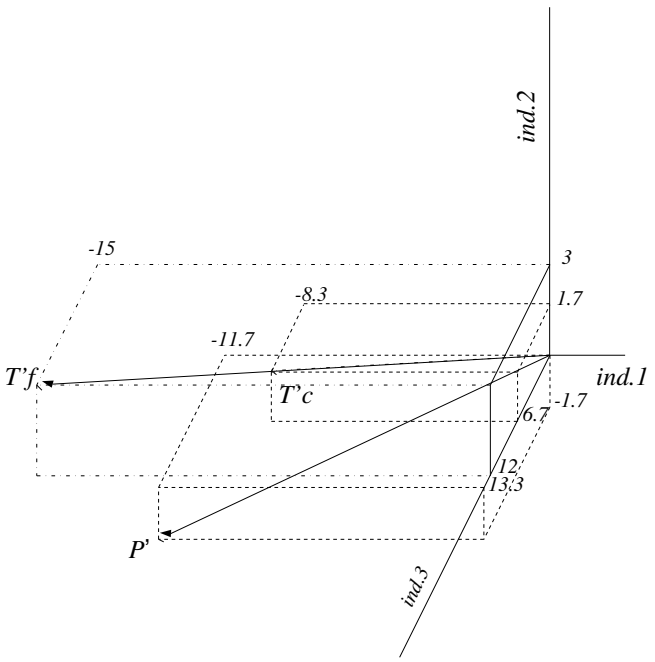

Figure 2.6 : Représentation de T'_c, T'_f et P'.

2.1.3 Expliquer ou prédire des variables

Considérons le modèle de régression $Y = \alpha + \beta X + \cdots$ dans lequel X, Y, $\cdots$ sont interprétés comme des vecteurs.

Le critère des moindres carrés consiste à choisir α, β, $\cdots$ tels que $|Y - \alpha J - \beta X - \cdots|^2$ soit minimum.

Si P est l'hyperplan déterminé par J, X, $\cdots$ alors la meilleure estimation de $\alpha J + \beta X + \cdots$ est la projection orthogonale $\hat{Y}_P$ de Y sur P. Pour mesurer la qualité de l'estimation on utilise soit $SCE = |Y - \hat{Y}_P|^2$ soit l'angle que font $Y - \hat{Y}_J$ et $\hat{Y}_P - \hat{Y}_J$.

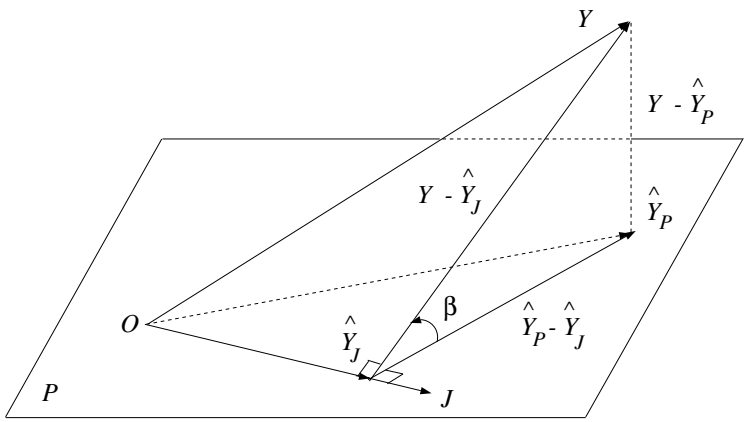

Figure 2.7 : Régression et coefficient de corrélation multiple.

Sur la figure 2.7, on a représenté l'angle β que font les vecteurs $Y - \hat{Y}_J$ et $\hat{Y}_P - \hat{Y}_J$. Ces deux vecteurs sont dans le sous espace des variables centrées car ils sont tous les deux orthogonaux à J. Une fonction de cet angle peut donc servir pour mesurer la qualité du modèle utilisé. En particulier, $cos^2(\beta)$ qui n'est autre que le rapport des carrés des longueurs des vecteurs $\hat{Y}_P - \hat{Y}_J$ et $Y - \hat{Y}_J$. $Cos^2(\beta)$ est le carré du coefficient de corrélation multiple c'est, le *coefficient de détermination* que l'on note R^2.

Le coefficient de détermination est donc le rapport du carré de la longueur d'un côté de l'angle droit et du carré de la longueur de l'hypothénuse du triangle limité par les extrémités Y, $\hat{Y}_p$ et $\hat{Y}_J$ des vecteurs OY, $O\hat{Y}_P$ et $O\hat{Y}_J$. Dans ces conditions, il est clair, que le coefficient de détermination est, au même titre que $|Y - \hat{Y}_P|^2$, une mesure de la qualité du modèle utilisé. En effet, le théorème de pythagore permet lui aussi d'exprimer l'erreur que l'on commet quand on utilise un modèle car $|Y - \hat{Y}_P|^2$ est lié à $|Y - \hat{Y}_J|^2$ et $|\hat{Y}_P - \hat{Y}_J|^2$ par la relation $|Y - \hat{Y}_J|^2 = |Y - \hat{Y}_P|^2 + |\hat{Y}_P - \hat{Y}_J|^2$.

Notion de dualité

On illustre la notion de dualité sur l'exemple de la droite de régression (Bachacou *et al.*, 1981).

Classiquement, on estime les paramètres α et β de la droite de régression de Y en X sur un échantillon de n individus ($y_i = \alpha + \beta x_i + e_i$; $i = 1, \cdots, n$) par la méthode des moindres carrés, en se plaçant dans l'espace des individus ($I\!R^2$) . Nous allons montrer que ces estimations peuvent aussi être obtenues en raisonnant dans l'espace des variables ($I\!R^n$). Soit :

- OY le vecteur de composantes $(y_1, \cdots, y_i, \cdots, y_n)$

- OX le vecteur de composantes $(x_1, \cdots, x_i, \cdots, x_n)$
- OJ le vecteur de composantes $(1, \cdots, 1, \cdots, 1)$

Le problème consiste à trouver le vecteur $O\hat{Y} = \alpha\, OJ + \beta\, OX$ tel que $|OY - O\hat{Y}|^2$ soit minimum ($O\hat{Y}$ projection orthogonale de OY sur le plan défini par OJ et OX).

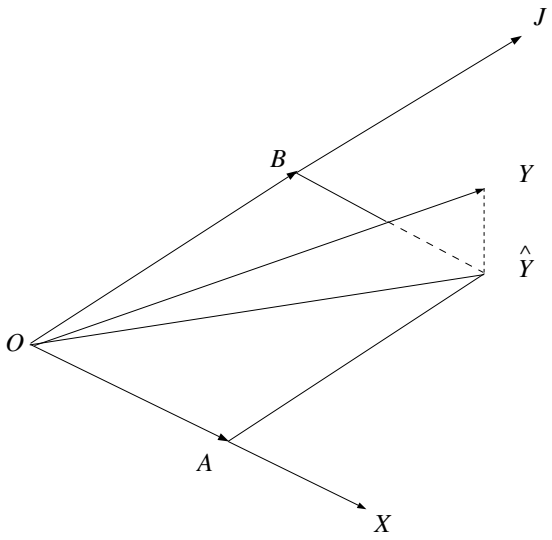

Figure 2.8 : Projection sur le plan XOJ.

On cherche α et β tels que :

$$OA = \beta OX \text{ et } OB = \alpha OJ$$

<u>*1 - Estimation de α*</u>

On projette orthogonalement OY, $O\hat{Y}$, OX et OA sur OJ (figure 2.9). D'après le théorème des trois perpendiculaires, Y et $\hat{Y}$ se projettent au même point $\bar{Y}$; X et A se projettent respectivement en $\bar{X}$ et $\bar{A}$. De $OB = A\hat{Y}$ (figure 2.8) et $A\hat{Y} = \bar{A}\bar{Y}$ (figure 2.9), on déduit que :

$$OB = O\bar{Y} + \bar{Y}B = O\bar{Y} - O\bar{A} \text{ car } \bar{Y}B = -O\bar{A}$$

$$OB = O\bar{Y} - \beta\, O\bar{X} \Longrightarrow \alpha = \bar{y} - \beta\, \bar{x}$$

avec

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(cf. § 2.1.1)}$$

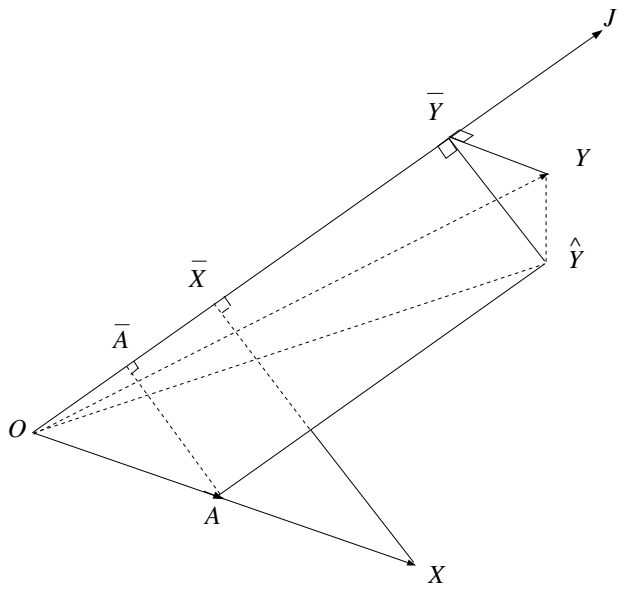

Figure 2.9 : Projections sur OJ.

2 - Estimation de β

On projette orthogonalement Y, $\hat{Y}$, X et A sur le sous espace des variables centrées orthogonal à OJ. On obtient les vecteurs OX' et OY' des variables centrées ainsi que la projection orthogonale de A et de $\hat{Y}$ en A' (figure 2.10).

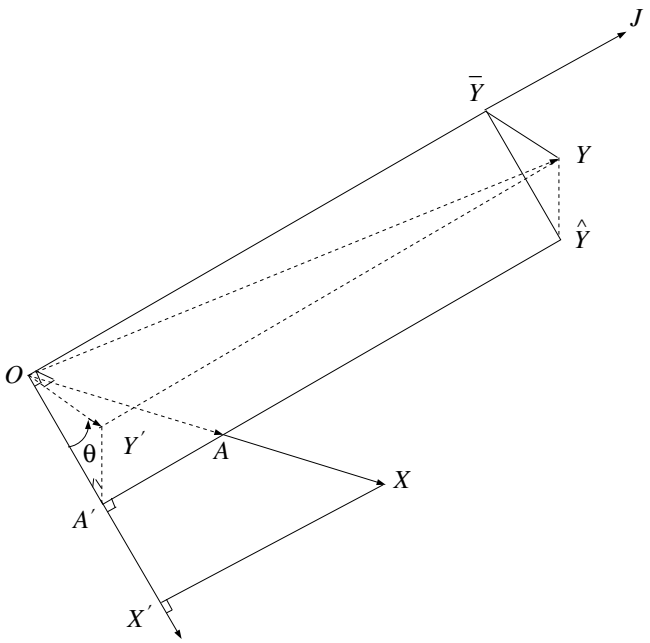

Figure 2.10 : Projections sur le sous espace des variables centrées.

La régression sur les variables centrées se résout en projetant orthogonalement OY' sur OX'. Comme OA' est le projeté de OY' sur OX', on écrit que le produit scalaire des vecteurs OA' et $OY' - OA'$ est nul :

$$< OA', OY' - OA' > = < \beta OX', OY' - \beta OX' >$$

$$< \beta OX', OY' - \beta OX' > = \beta < OX', OY' > - \beta^2 < OX', OX' > = 0$$

$$\beta = \frac{< OX', OY' >}{< OX', OX' >} = \frac{< OX', OY' >}{< OX', OX' >^{1/2} < OY', OY' >^{1/2}} \times \frac{< OY', OY' >^{1/2}}{< OX', OX' >^{1/2}}$$

$$\beta = \rho \frac{S_y}{S_x}$$

ρ est l'estimation du coefficient de corrélation, S_y et S_x les estimations des écarts types des y et des x. En effet :

$$< OX', OY' > = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$< OX', OX' > = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$< OY', OY' > = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Le sous espace des variables centrées est de dimension $n - 1$ *(degrés de liberté).*

Les résultats ainsi obtenus sont identiques à ceux que l'on obtient quand on raisonne sur l'espace des individus.

2.1.4 Comparer des populations

On veut comparer deux populations P_1 et P_2. Soit Y une variable aléatoire observée sur P_1 et P_2.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

On suppose Y_1 et Y_2 normalement distribuées :

$$Y_1 \sim N(\mu_1, \sigma_1^2) \text{ et } Y_2 \sim N(\mu_2, \sigma_2^2)$$

On note $y_{11}, \cdots, y_{1n_1}$ et $y_{21}, \cdots, y_{2n_2}$ les n_1 et n_2 observations de Y_1 et Y_2 respectivement.

Pour comparer les moyennes des populations P_1 et P_2 :

1 - On estime les moyennes (μ_1 et μ_2) et la variance (σ^2). Soient $y_{1.}$ et $y_{2.}$ les estimateurs de μ_1 et μ_2 ; S_1^2, S_2^2 les estimateurs des variances des populations P_1, P_2 et S^2 l'estimateur de la variance commune aux deux populations :

Moyennes :

$$y_{1.} = \sum_{j=1}^{n_1} \frac{y_{1j}}{n_1} \qquad y_{2.} = \sum_{j=1}^{n_2} \frac{y_{2j}}{n_2}$$

Variances :

$$S_1^2 = \sum_{j=1}^{n_1} \frac{(y_{1j} - y_{1.})^2}{(n_1 - 1)} \qquad S_2^2 = \sum_{j=1}^{n_2} \frac{(y_{2j} - y_{2.})^2}{(n_2 - 1)}$$

Variance commune :

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - y_{1.})^2 + \sum_{j=1}^{n_2}(y_{2j} - y_{2.})^2}{n_1 + n_2 - 2}$$

2 -On utilise la statistique de test :

$$T = \frac{y_{1.} - y_{2.}}{S(\frac{1}{n_1} + \frac{1}{n_2})^{\frac{1}{2}}} \Longrightarrow T^2 = \frac{n_1 \times n_2}{n_1 + n_2}(y_{1.} - y_{2.})S^{-2}(y_{1.} - y_{2.})$$

T est la statistique de Student à $n_1 + n_2 - 2$ degrés de liberté.

Quelle est l'interprétation géométrique de la statistique de Student ?

Considérons le terme $(y_{1.} - y_{2.})S^{-2}(y_{1.} - y_{2.})$ de T^2, il contient :

1 - Le terme $(y_{1.} - y_{2.})(y_{1.} - y_{2.})$ qui n'est autre que le carré de la distance entre les points moyens des deux populations P_1 et P_2.

2 - Le terme S^{-2}.

Donc dans T^2, le carré de la distance entre les points moyens des deux populations est pondéré par l'inverse de la variance commune aux deux populations.

Faisons le changement de variable $y'_{1.} = y_{1.}/S$ et $y'_{2.} = y_{2.}/S$ alors :

$$(y_{1.} - y_{2.})S^{-2}(y_{1.} - y_{2.}) = (y'_{1.} - y'_{2.})(y'_{1.} - y'_{2.})$$

Le deuxième membre de cette expression est le carré d'une distance au sens où nous l'avons défini jusqu'à présent. On en déduit que le premier membre est aussi le carré d'une distance. Comme une distance dérive d'un produit scalaire, on peut définir le produit scalaire de deux vecteurs U et V par $< U, V > = U'MV$ avec U' le vecteur U *transposé* et M une matrice *symétrique définie positive* désignée sous le nom de *métrique* (ces notions seront définies ultérieurement).

Dans le cas qui nous intéresse, $U = U' = V = y_{1.} - y_{2.}$ et $M = S^{-2}$. Ainsi, le carré de la statistique de Student n'est autre (au coefficient $n_1 n_2/(n_1 + n_2)$ près)

qu'une distance au carré qui prend en compte la variabilité des observations par le truchement d'une métrique dite *intra-population*. Cette notion de métrique est essentielle en analyse multidimensionnelle (analyses factorielles, classifications). En analyse de variance multidimensionnelle, on retrouve un équivalent de la distance mise en évidence dans la statistique de Student, c'est la distance dite de Mahalanobis.

Comment poser ce problème de test d'un point de vue géométrique ?

1 - On représente les variables par des vecteurs. Soit :

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} \quad \hat{Y}_P = \begin{bmatrix} y_{1.} \\ \vdots \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{2.} \end{bmatrix} \quad \hat{Y}_J = \begin{bmatrix} y_{..} \\ \vdots \\ y_{..} \\ y_{..} \\ \vdots \\ y_{..} \end{bmatrix} \quad J = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad X_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad X_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Ces vecteurs sont dans des espaces vectoriels :

- $Y \in I\!R^n$ (espace de dimension $n = n_1 + n_2$).
- $\hat{Y}_P \in P$ (sous espace de $I\!R^n$ de dimension 2 engendré par X_1 et X_2).
- $\hat{Y}_J \in$ sous espace de P, de dimension 1, engendré par J.
- $Y - \hat{Y}_P \in P^\perp$ (sous espace orthogonal à P, de dimension $n-2$).
- $\hat{Y}_P - \hat{Y}_J \in$ sous espace de P, orthogonal à J, de dimension 1.

La représentation géométrique de ces vecteurs est la suivante (figure 2.11) :

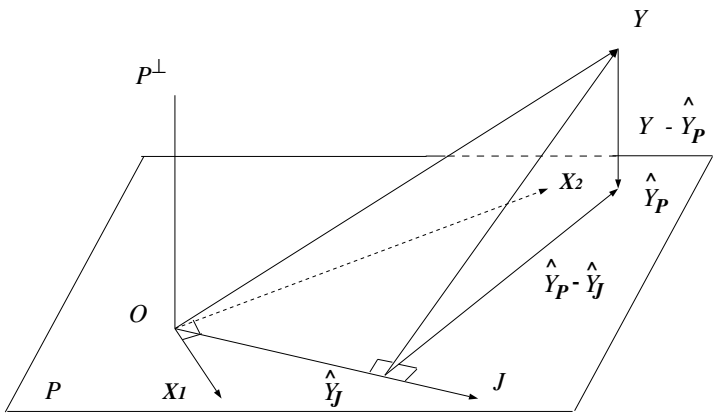

Figure 2.11 : Comparaison de deux populations P_1 et P_2.

2 - Pour chaque observation, on écrit que la valeur observée est égale à la moyenne générale (m) plus un écart à la moyenne générale spécifique à chaque population (m_1 pour les observations de P_1 et m_2 pour les observations de P_2) plus un terme d'erreur. Soit :

$$\begin{cases} y_{11} & = & m & + & m_1 & + & e_{11} \\ \cdots & & \cdots & & \cdots & & \cdots \\ y_{1n_1} & = & m & + & m_1 & + & e_{1n_1} \\ y_{21} & = & m & + & m_2 & + & e_{21} \\ \cdots & & \cdots & & \cdots & & \cdots \\ y_{2n_2} & = & m & + & m_2 & + & e_{2n_2} \end{cases}$$

L'écriture *matricielle* de ce système est :

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} m \\ m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ \vdots \\ e_{2n_2} \end{bmatrix}$$

On l'écrit symboliquement : $Y = X \times \Theta + E$

Pour estimer les paramètres (m, m_1, m_2) et la variance (S^2), on minimise :

$$|E|^2 = |Y - \hat{Y}_P|^2 = < Y - X\Theta, Y - X\Theta >$$

3 - Pour éprouver l'hypothèse que les deux populations P_1 et P_2 sont identiques ($\mu_1 = \mu_2$) contre l'hypothèse alternative, elles sont différentes ($\mu_1 \neq \mu_2$), on évalue $|\hat{Y}_P - \hat{Y}_J|^2$ et $|Y - \hat{Y}_P|^2$ (si P_1 et P_2 sont identiques, les écarts (m_1 et m_2) à la moyenne m sont nuls et Y se projette sur J). Soit :

$$|\hat{Y}_P - \hat{Y}_J|^2 = < \hat{Y}_P - \hat{Y}_J, \hat{Y}_P - \hat{Y}_J > = \sum_{i=1}^{2} (y_{i.} - y_{..})^2 = \frac{n_1 \times n_2}{n_1 + n_2} (y_{1.} - y_{2.})^2$$

avec $y_{..} = \dfrac{1}{n_1 + n_2}(n_1 y_{1.} + n_2 y_{2.})$

$$|Y - \hat{Y}_P|^2 = < Y - \hat{Y}_P, Y - \hat{Y}_P > = \sum_{j=1}^{n_1} (y_{1j} - y_{1.})^2 + \sum_{j=1}^{n_2} (y_{2j} - y_{2.})^2$$

On utilise la statistique F de Fisher à 1 et $n_1 + n_2 - 2$ degrés de liberté :

$$F = \frac{\dfrac{|\hat{Y}_P - \hat{Y}_J|^2}{1}}{\dfrac{|Y - \hat{Y}_P|^2}{n_1 + n_2 - 2}} = \frac{(y_{1.} - y_{2.})^2}{S^2(\dfrac{1}{n_1} + \dfrac{1}{n_2})}$$

Conclusion

1 - Les sommes des carrés sont des longueurs de vecteurs au carré.

2 - Les degrés de liberté sont les dimensions des espaces qui contiennent les vecteurs.

3 - L'hypothèse testée est :

$H_0 \ : \ \mu_1 \ = \ \mu_2 \ = \ \mu$

$\bar{H_0} \ : \ \mu_1 \ \neq = \ \mu_2$

Sous H_0, F suit une loi de Fisher $F_{(1,n_1+n_2-2)}$ degrés de liberté.

Si F est supérieure à une certaine valeur (lue dans une table) $\Longleftrightarrow$ Si $|\hat{Y}_P - \hat{Y}_J|$ est grand, on rejette H_0.

4 - La statistique de Fisher à 1 et $n_1 + n_2 - 2$ degrés de liberté est identique au carré de la statistique de Student à $n_1 + n_2 - 2$ degrés de liberté.

2.2 Probabilité et Géométrie

A condition de bien identifier les notions de vecteur et de produit scalaire, on peut raisonner géométriquement avec des variables aléatoires (cf. § 1.2). Supposons que nous avons des variables aléatoires X, Y, Z, $\cdots$ conjointement distribuées $F(x, y, z, \cdots)$.

2.2.1 Espérance mathématique

L'espérance d'une variable aléatoire est calculée par :

$$E(X) \ = \ \int x dF(x)$$

avec $F()$ la fonction de répartition des variables impliquées.

Supposons que les variables aléatoires X, Y, ... ont pour moyenne $E(X) \ = \ \mu_X$, $E(Y) \ = \ \mu_Y$, $\cdots$. Ces variables aléatoires peuvent être représentées par des vecteurs qu'on peut additionner et multiplier par des scalaires. La quantité :

$$< X, Y > = \ E(XY) \ = \ \int xy dF(x, y)$$

définit le produit scalaire de X et Y.

Ici, l'équivalent du vecteur J (vecteur dont toutes les composantes sont égales à 1) est la constante $\mathbf{1}$. Par exemple, la projection d'une variable aléatoire X sur $\mathbf{1}$ doit être telle que $\hat{X}_1 = \alpha \mathbf{1}$ quelque soit α (cf. propriété P_1 du § 1.3) et telle que

$< X - \hat{X}_1, \mathbf{1} > = 0$ (cf. propriété P_2 du § 1.3). On en déduit que $E(X) = \alpha$ de telle sorte que $\mu_X \mathbf{1} = \mu_X$ est la projection de X sur $\mathbf{1}$.

Le vecteur $X - \mu_X$ est une variable *aléatoire centrée*.

2.2.2 Variance

Pour toute variable aléatoire X, $< X, X > = E(X^2)$ est la longueur au carré du vecteur X. Le carré de la longueur de la variable aléatoire $X - \mu_X$ est :

$$|X - \mu_X|^2 = < X - \mu_X, X - \mu_X > = E(X - \mu_X)^2 = \sigma_X^2$$

c'est la *variance* de X.

2.2.3 Corrélation

L'angle θ que font les variables centrées $X - \mu_X$ et $Y - \mu_Y$ joue un rôle important car :

$$cos(\theta) = \frac{< X - \mu_X, Y - \mu_Y >}{|X - \mu_X||Y - \mu_Y|}$$

représente la *corrélation* entre X et Y.

Quand en théorie des probabilités on écrit :

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

on ne fait qu'écrire la loi du cosinus.

Dire que deux variables aléatoires ne sont pas corrélées est équivalent à dire que les vecteurs correspondants sont orthogonaux. Les interprétations géométriques sont analogues à celles qui ont été faites précédemment, la seule différence est qu'ici, on interprète un modèle théorique alors qu'en statistique, on interprète les données.

2.2.4 Espérance conditionnelle

D'un point de vue plus général, l'espérance conditionnelle de X sachant Y, Z, $\cdots$, notée $E(X|Y, Z, \cdots)$, est identique à la projection de X sur un plan P déterminé par Y, Z, $\cdots$. En fait, P est l'espace des fonctions de Y, Z, $\cdots$ de carré intégrable, il est fermé pour les opérations linéaires. C'est pour cette raison qu'il est interprété comme un sous espace linéaire.

Considérons les trois propriétés des espérances conditionnelles :

$$E(X|Y, Z, \cdots) \text{ est une fonction de } Y,\ Z,\ \cdots \qquad (4)$$

$$E[E(X|Y,Z,\cdots)] \;=\; E(X) \qquad (5)$$

Si g est une fonction de Y, Z, $\cdots$, alors :

$$E(Xg|Y,Z,\cdots) \;=\; gE(X|Y,Z,\cdots) \qquad (6)$$

De (5) et (6), on déduit que si g est une fonction de Y, Z, $\cdots$ alors :

$$< g, X - E(X|Y,Z,\cdots) > \;=\; 0 \qquad (7)$$

Donc, (4) dit que $E(X|Y,Z,\cdots)$ est dans le plan P, tandis que (7) précise que $X - E(X|Y,Z,\cdots)$ est orthogonal à tout vecteur g de P donc en particulier à $E(X|Y,Z,\cdots)$.

L'espérance conditionnelle est donc l'analogue d'une projection (figure 2.12).

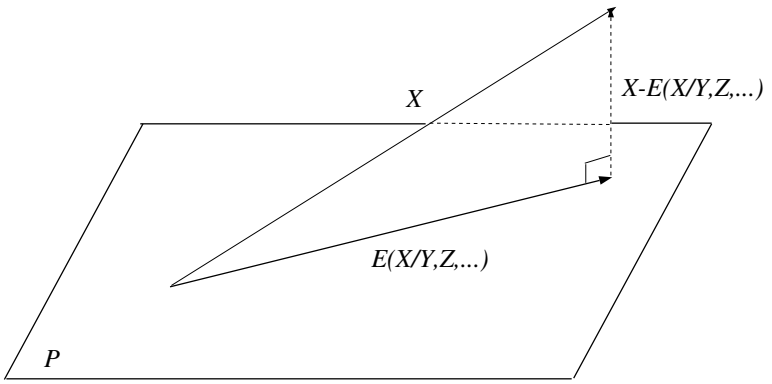

Figure 2.12 : Espérance conditionnelle.

Les concepts géométriques permettent d'interpréter la projection $\hat{X}_P$ de X sur P comme le point de P le plus proche de X. Ainsi, si Y, Z, $\cdots$ représentent la *connaissance* et X une valeur à prédire, alors la projection est la fonction de la connaissance qui prédit le mieux X car elle donne le plus petit carré moyen d'erreur.

2.2.5 Variance et variance conditionnelle

Un autre théorème de la théorie des probabilités dit que :

$$var(X) \;=\; var[E(X|Y,Z,\cdots)] + E[var(X|Y,Z,\cdots)] \qquad (8)$$

D'après (5), $E[X - E(X|Y,Z,\cdots)] \;=\; 0$ alors :

$$var[X - E(X|Y,Z,\cdots)] \;=\; E[X - E(X|Y,Z,\cdots)]^2$$

Puisque la variance conditionnelle est la variance relative à la distribution conditionnelle, (5) implique que c'est aussi égal à $E[var(X|Y,Z,\cdots)]$. Donc (8) s'écrit :

$$var(X) \;=\; var[E(X|Y,Z,\cdots)] + var[X - E(X|Y,Z,\cdots)]$$

qui n'est autre qu'une extension du théorème de Pythagore.

2.2.6 Exemple

Il y a différentes façons de démontrer qu'une combinaison linéaire de deux variables aléatoires normales, indépendantes suit une loi normale (Anderson, 1958 pp. 19-24 ; Mood, Graybill et Boes, 1974 pp. 185-186, 193-194). Cette démonstration proposée par Hombas (1989) a pour objectif d'appliquer les notions de géométrie analytique qui ont été présentées, elle a le mérite d'être plus simple et plus imagée que les démonstrations faites classiquement.

- Soient X_1 et X_2 deux variables aléatoires distribuées indépendamment suivant une loi normale centrée réduite $N(0,1)$. Soit $X \;=\; (X_1, X_2)$.

  La densité conjointe de X_1 et X_2 est :

$$f_X(x_1,x_2) \;=\; (2\pi)^{-1} exp(-(x_1^2 + x_2^2)/2)$$

$$-\infty < x_1 < \infty,\; -\infty < x_2 < \infty$$

- Soient X_1' et X_2' deux autres variables aléatoires, indépendantes et distribuées normalement $N(0,1)$. Soit $X' \;=\; (X_1', X_2')$.

  La densité conjointe de X_1' et X_2' est :

$$f_{X'}(x_1',x_2') \;=\; (2\pi)^{-1} exp(-(x_1'^2 + x_2'^2)/2) \qquad (9)$$

$$-\infty < x_1' < \infty,\; -\infty < x_2' < \infty$$

- Dans les deux cas, les densités conjointes ont pour contours des cercles centrés sur 0. Pour les espaces échantillons (X_1, X_2) et (X_1', X_2') qui sont coplanaires, les surfaces définies par les densités de probabilité coïncident si $x_1'^2 + x_2'^2 \;=\; x_1^2 + x_2^2$. Une transformation $(X_1, X_2) \longrightarrow (X_1', X_2')$ qui satisfait cette condition est la rotation orthogonale montrée sur la figure 2.13 :

$$\left\{ \begin{array}{lcl} X_1' &=& X_1 cos\theta + X_2 sin\theta \\ X_2' &=& -X_1 sin\theta + X_2 cos\theta \end{array} \right. \qquad (10)$$

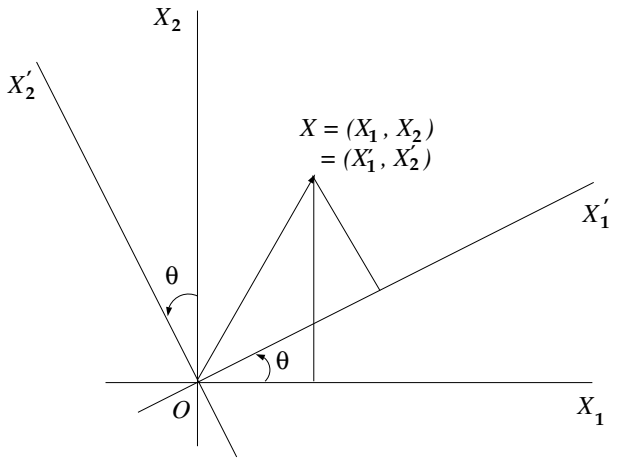

Figure 2.13 : Rotation orthogonale dans le plan.

- La factorisation de (9) donne :

$$f_{X'}(x_1', x_2') \;=\; (2\pi)^{-1/2} exp(-x_1'^2/2) \times (2\pi)^{-1/2} exp(-x_2'^2/2)$$

  Puisque $X_1' \sim N(0,1)$ il en est de même pour $X_1 cos\theta + X_2 sin\theta \sim N(0,1)$. De plus, X_1' et X_2' sont indépendantes.

- Donc les deux variables indépendantes, normalement distribuées $N(0,1)$, obéissent à la propriété d'invariance rotationnelle suivante : la rotation orthogonale d'un système de coordonnées déterminé par deux variables aléatoires indépendantes $N(0,1)$ préserve l'indépendance et la normalité.

Cette discussion prouve aussi que la variable aléatoire $X' = \alpha_1 X_1 + \alpha_2 X_2$ avec $\alpha_1^2 + \alpha_2^2 \;=\; 1$, qui représente l'une des composantes de (10), est elle-même normalement distribuée $N(0,1)$.

Considérons maintenant les scalaires β_1 et β_2 et soient α_1 et α_2 tels que :

$$\alpha_1 \;=\; \frac{\beta_1}{(\beta_1^2 + \beta_2^2)^{1/2}} \quad \text{et} \quad \alpha_2 \;=\; \frac{\beta_2}{(\beta_1^2 + \beta_2^2)^{1/2}}$$

Ils satisfont à $\alpha_1^2 + \alpha_2^2 \;=\; 1$.

Ainsi, si X_1 et X_2 sont $N(0,1)$, on déduit que :

$$\frac{(\beta_1 X_1 + \beta_2 X_2)}{(\beta_1^2 + \beta_2^2)^{1/2}} \sim \; N(0,1)$$

Donc :

$$\beta_1 X_1 + \beta_2 X_2 \;\sim\; N(0, \alpha_1^2 + \alpha_2^2)$$

Soient X et Y deux variables aléatoires normales indépendantes $N(\mu_i, \sigma_i^2)$, $i = 1, 2$. Alors, la variable $Z = \beta_1 X + \beta_2 Y$ peut être écrite :

$$Z = \beta_1\sigma_1(X - \mu_1)/\sigma_1 + \beta_2\sigma_2(Y - \mu_2)/\sigma_2 + \beta_1\mu_1 + \beta_2\mu_2 \qquad (11)$$

Les variables aléatoires $(X - \mu_1)/\sigma_1$ et $(Y - \mu_2)/\sigma_2$ sont $N(0, 1)$. En appliquant à (11) les résultats précédents, on déduit que :

$$Z \sim N(\beta_1\mu_1 + \beta_2\mu_2, \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2)$$

La généralisation à plus de deux variables aléatoires est évidente.

2.3 Compléments bibliographiques

La présentation géométrique des concepts utilisés en statistique a fait l'objet de quelques articles. Elle a été initiée par Fisher (1915) avec son article sur la distribution du coefficient de corrélation, elle s'est poursuivie jusqu'à la parution de l'article de Kruskal (1975) sur la géométrie des inverses généralisées. L'impopularité relative de cette approche a pour origine l'idée fausse que géométrie est synonyme d'abstraction alors que, au contraire, elle permet de mieux comprendre la signification des objets manipulés par la statistique. De ce dernier point de vue, les articles de Herr (1980) et de Margolis (1979) présentent de l'intérêt.

2.4 Tableau résumé

Espace vectoriel	Expression géométrique Représentation	Forme analytique	Expression Statistique	Probabiliste
Vecteur U	Segment orienté de 0 vers U	$U = (u_1, u_2, \cdots, u_n)$	Données $u_1, u_2, \cdots, u_n$	V. aléatoire U
Vecteur $U - V$	Segment orienté de V vers U	$U - V = (u_1 - v_1,$ $, \cdots, u_n - v_n)$	Différences $u_1 - v_1, \cdots$	V. aléatoire $U - V$
$< U, V >$	Mesure de non orthogonalité	$u_1 v_1 + \cdots + u_n v_n$	Somme des produits	$E(UV)$
$< U, U >$	Longueur au carré	$u_1^2 + \cdots + u_n^2$	Somme des carrés	$E(U^2)$
$\|U - V\|^2$	Distance au carré	$(u_1 - v_1)^2 + \cdots$	Somme carrés différences	$E(U - V)^2$
$< U, V >= 0$	Orthogonalité : U, V perpendiculaires			
$\alpha U + \beta V + \cdots$	Hyperplans contenant O, engendrés par U, V, $\cdots$		Modèles $\alpha U + \beta V + \cdots$	Fonctions de carré intégrable
$< U - \hat{U}, V >= 0$ $\forall$ V dans P	Projection de U sur P		Meilleur modèle $\alpha U + \beta V + \cdots$	$E(U\|V, W, \ldots)$
Vecteur J	Segment orienté incliné à 45°	$J = (1, 1, \ldots, 1)$	Constante $(1, 1, \cdots, 1)$	Constante **1**
$\|U - \hat{U}_J\|^2$		$\sum(u_i - \bar{u})^2$	$(n-1) \times$ variance d'échantillonnage de U	Variance de U
$\frac{<U-\hat{U}_J, V-\hat{V}_J>}{\|U-\hat{U}_J\|.\|V-\hat{V}_J\|}$	Cosinus de l'angle entre $U - \hat{U}_J$ et $V - \hat{V}_J$		Corrélation d'échantillonnage de U et V	Corrélation de U et V
$< U - \hat{U}_J, V - \hat{V}_J >= 0$	$U - \hat{U}_J$ et $V - \hat{V}_J$ perpendiculaires		U et V non corrélées	U et V non corrélées

Tableau 2.1 : Correspondance entre Espace vectoriel, Géométrie, Statistique et Probabilité.