PARTIE II

GEOMETRIE ET STATISTIQUE

1 REPRESENTATION DES DONNEES

- **Les espaces des individus et des variables**

2 LES RESUMES STATISTIQUES

- **Moyenne et variance**

3 LES LIAISONS STATISTIQUES

- **Exemple : la liaison pression - température**
- **Notion de corrélation**

3 LES TESTS STATISTIQUES

- **Comparaison de deux populations**

4 RESUME

1 REPRESENTATION DES DONNEES

Exemple :

Un tableau de données
les lignes $\Longleftrightarrow$ les individus
les colonnes $\Longleftrightarrow$ les variables

	var.1	*var*.2
ind.1	2	4
ind.2	4	3
ind.3	6	5

Que peut-on représenter géométriquement ?

1 REPRESENTATION DES DONNEES

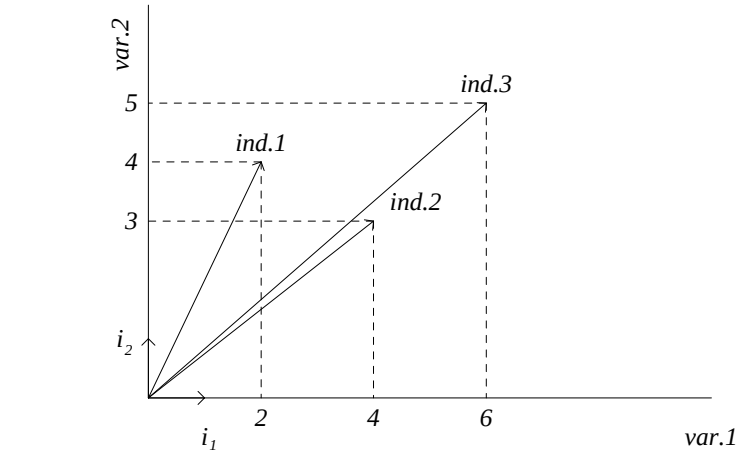
(*cf.* paragraphe 1)

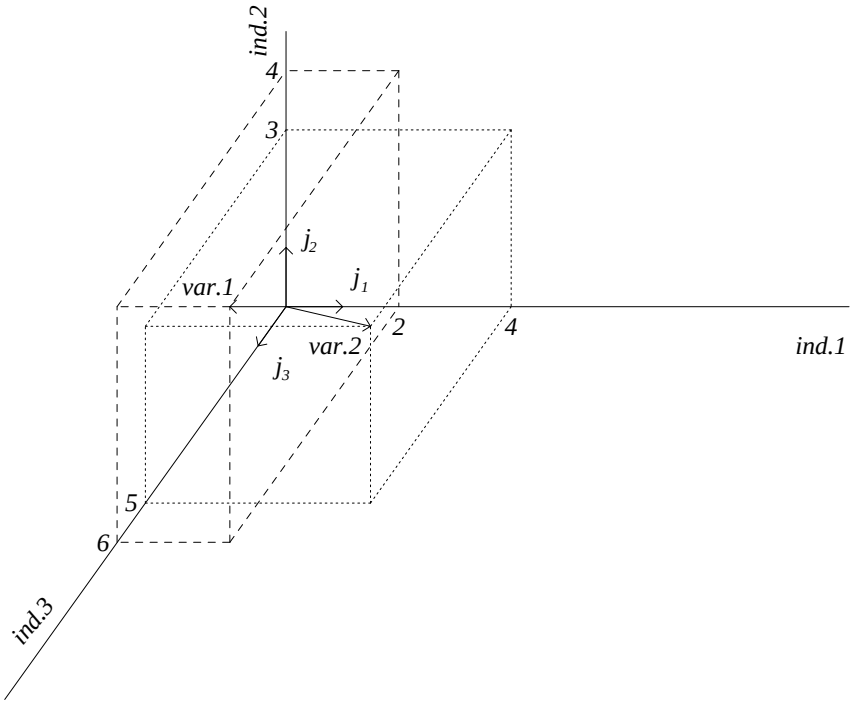
On considère le tableau de chiffres ci-contre. Il a 3 lignes (les individus) et 2 colonnes (les variables).

- Les lignes et les colonnes du tableau sont des **vecteurs** lignes et colonnes qui peuvent être représentés géométriquement.
- Pour appréhender toute l'information contenue dans le tableau, on lit le tableau ligne après ligne ou colonne après colonne. Quelque soit le mode de lecture adopté (ligne ou colonne), on lit toujours la totalité de l'information.

Il y a donc deux espaces vectoriels différents dans lesquels les données peuvent être représentées. Ce sont **les espaces** des **individus** et des **variables.**

L'espace des individus et l'espace des variables





L'espace des individus et l'espace des variables

- Chaque ligne du tableau contient deux chiffres qui représentent les coordonnées de l'extrémité d'un **vecteur "individu"** dans l'espace défini par les deux variables. Cet espace est appelé **espace des individus**.
- Chaque colonne du tableau contient trois chiffres qui représentent les coordonnées de l'extrémité d'un **vecteur "variable"** dans l'espace défini par les trois individus. Cet espace est appelé **espace des variables**.
- Ces espaces ont des dimensions différentes (2 pour l'espace des individus et 3 pour l'espace des variables), ce sont des représentations différentes de l'information contenue dans le tableau des données. Ces deux espaces sont dits **espaces duaux**.

2 LES RESUMES STATISTIQUES

- $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ une variable observée sur n individus.
- $J = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ le vecteur à n composantes égales à 1.
- $\alpha J = \begin{bmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{bmatrix}$ un vecteur colinéaire à J.
- $\hat{X}_J$ la projection orthogonale de X sur J.
- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ la moyenne.
- $|X|^2$ le carré de la longueur du vecteur X

2 LES RESUMES STATISTIQUES

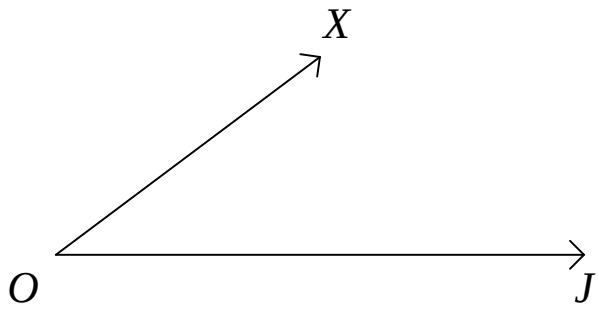
(*cf.* paragraphe 2)

Pour présenter les notions de moyenne, de variance, de coefficient de corrélation et de test on se place dans l'espace des variables. Le but poursuivi est double :

- Montrer que l'on retrouve bien les mêmes résultats que quand on raisonne sur l'espace des individus.
- Montrer que l'on peut représenter des variables statistiques telles que des sommes des carrés ou des coefficients de corrélation alors que ces visualisations ne sont pas possibles dans l'espace des individus.

On donne ici les principales notations qui seront utilisées par la suite.

Moyenne et Variance



Moyenne et Variance

(*cf.* paragraphe 2.1.1)

- On observe une variable X sur n individus.
- On représente :
  - Les n observations de la variable X par le vecteur :

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

  - Le vecteur :

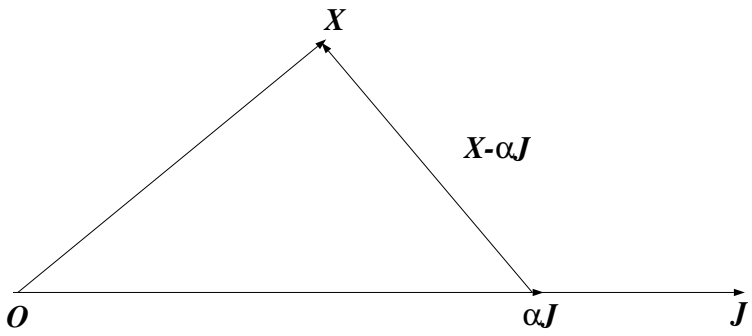
$$J = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

    dans l'espace qui est de dimension n.

- On se pose la question :

  Comment résumer l'information contenue dans le vecteur X ?

Moyenne et variance



Moyenne et variance

La réponse à la question posée est :

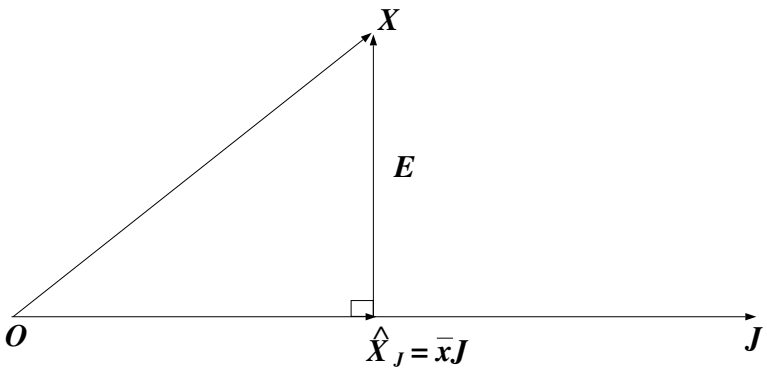
En trouvant une image de X dans un sous espace vectoriel de dimension inférieure à n.

- Le résumé le plus simple est obtenu en **projetant** le vecteur X sur un espace à **une** dimension (une droite).

  On projette donc X sur J. Soit αJ un projeté **quelconque** de X sur J.

- **Parmi tous les vecteurs αJ en existe-t-il un qui soit meilleur que tous les autres ?**

Moyenne et variance



Moyenne et variance

Le meilleur résumé est celui pour lequel αJ est le plus proche de X.

- Soit $\hat{X}_J$ le projeté orthogonal de X sur J. C'est le meilleur résumé car c'est pour $\hat{X}_J$ que la distance $|X - \alpha J|$ est minimale. On a :

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

$$< X, J >= \alpha < J, J > \Longleftrightarrow \sum_{i=1}^{n} x_i = \alpha n$$

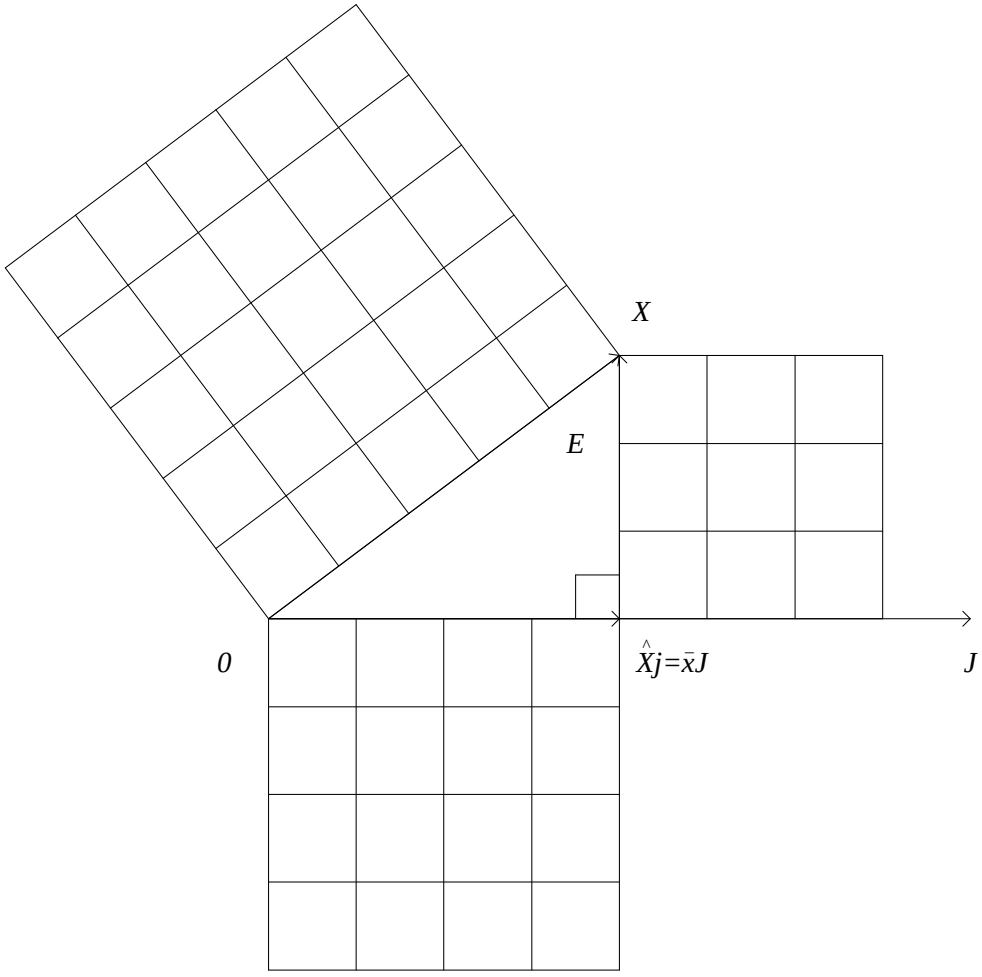
$$\alpha = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

  $\hat{X}_J$ est donc le vecteur dont toutes les composantes α sont égales à la moyenne d'échantillonnage.

- On a par ailleurs $X = \bar{x}J + (X - \bar{x}J)$ avec :
  - $\bar{x}J$ dans un sous espace de dimension 1.
  - $E = X - \bar{x}J$ dans un sous espace de dimension $(n-1)$ appelé **complémentaire orthogonal** à J.

Comment évaluer la qualité du résumé ?

Moyenne et variance



Moyenne et variance

- D'après le théorème de Pythagore :

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

  - $|\bar{x}J|^2$ est la somme des carrés expliquée par la moyenne d'échantillonnage.
  - $|E|^2$ est la somme des carrés d'erreur.

  Si toutes les observations étaient identiques, X serait sur $J \Rightarrow E = 0$. Donc, E représente la variation des composantes de X.

- La moyenne d'échantillonnage est un bon résumé si chaque composante de E est petite. On calcule donc la moyenne "par dimension" de la somme des carrés d'erreur. Soit :

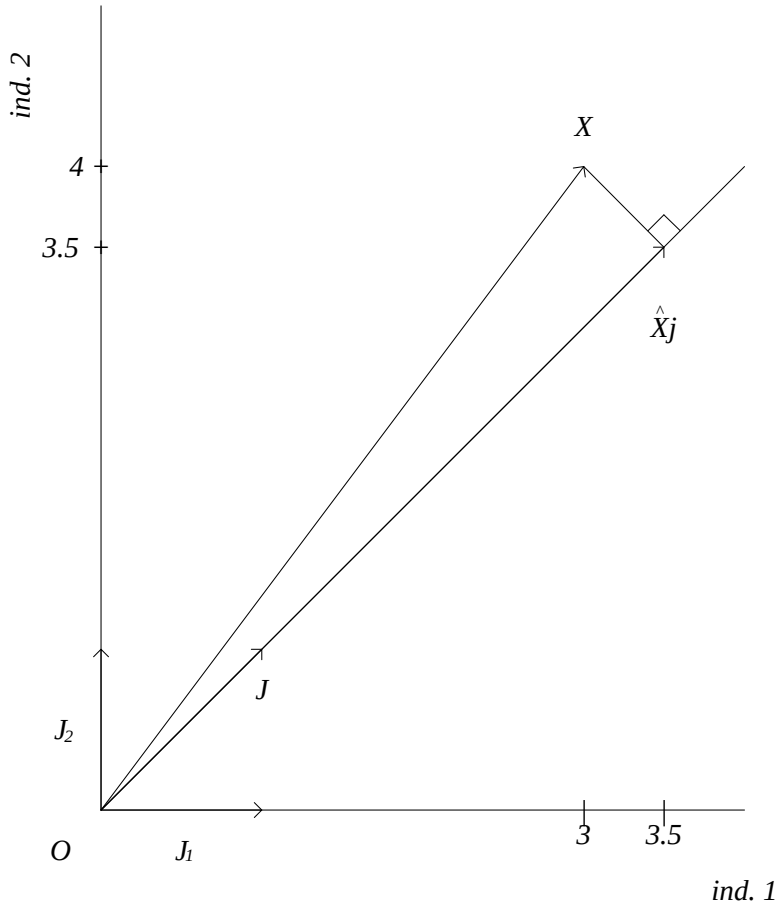
$$\frac{|E|^2}{n-1} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  On reconnaît la **variance d'échantillonnage**.

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

- La division par les degrés de liberté permet d'évaluer la qualité d'un résumé par rapport à l'erreur que l'on commet en résumant l'information.

Moyenne et variance
Exemple : $X' = [3\ 4]$. Moyenne ?



Moyenne et variance

Cet exemple a pour objectif d'illustrer numériquement les résultats précédents.

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

$$\hat{X}_J = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}$$

$$|\hat{X}_J|^2 =< \hat{X}_J, \hat{X}_J >= (3.5)^2 + (3.5)^2 = 24.5$$

$$|X - \hat{X}_J|^2 =< X - \hat{X}_J, X - \hat{X}_J >$$

$$|X - \hat{X}_J|^2 = (3 - 3.5)^2 + (4 - 3.5)^2 = 0.5$$

$$|\hat{X}_J|^2 + |X - \hat{X}_J|^2 = 24.5 + 0.5 = 25 = |X|^2$$

3 LES LIAISONS STATISTIQUES

- $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$ trois variables
- $\hat{X}_J$, $\hat{Y}_J$ et $\hat{Z}_J$ les projections orthogonales de X, Y et Z sur J.
- X', Y' et Z' les variables X, Y et Z centrées.

Exemple : Soit le tableau suivant des pressions et des températures (degrés centigrade et Fahrenheit).

	P	T_c	T_f
$ind.1$	10	5	41
$ind.2$	20	15	59
$ind.3$	35	20	69

Quel angle exprime la liaison entre la pression et la température ?

3 LES LIAISONS STATISTIQUES

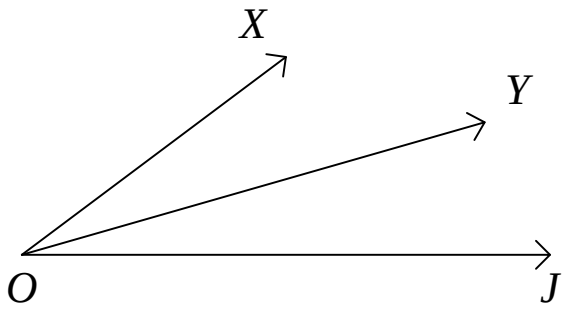
(*cf.* paragraphe 2.1.2)

- On précise les notations qui vont être utilisées. On rappelle ce qu'est une variable centrée :

$$X' = X - \hat{X}_J,\ Y' = Y - \hat{Y}_J \text{ et } Z' = Z - \hat{Z}_J$$

- On présente les données de l'exemple des pressions et des températures.

3 LES LIAISONS STATISTIQUES

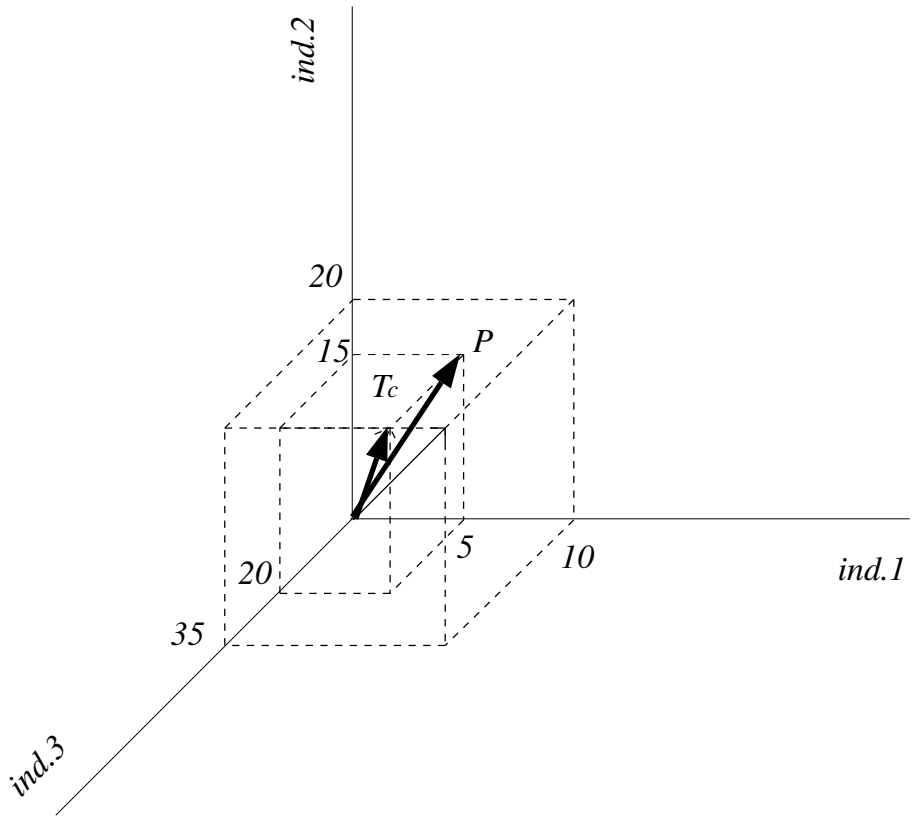


3 LES LIAISONS STATISTIQUES

- On observe les variables X et Y sur n individus.
- **Comment peut-on mesurer l'intensité de la relation entre X et Y ?**
- Intuitivement, on peut penser à l'angle entre les vecteurs X et Y.

  Cet angle ne convient pas, on va le montrer sur l'exemple des pressions et des températures.

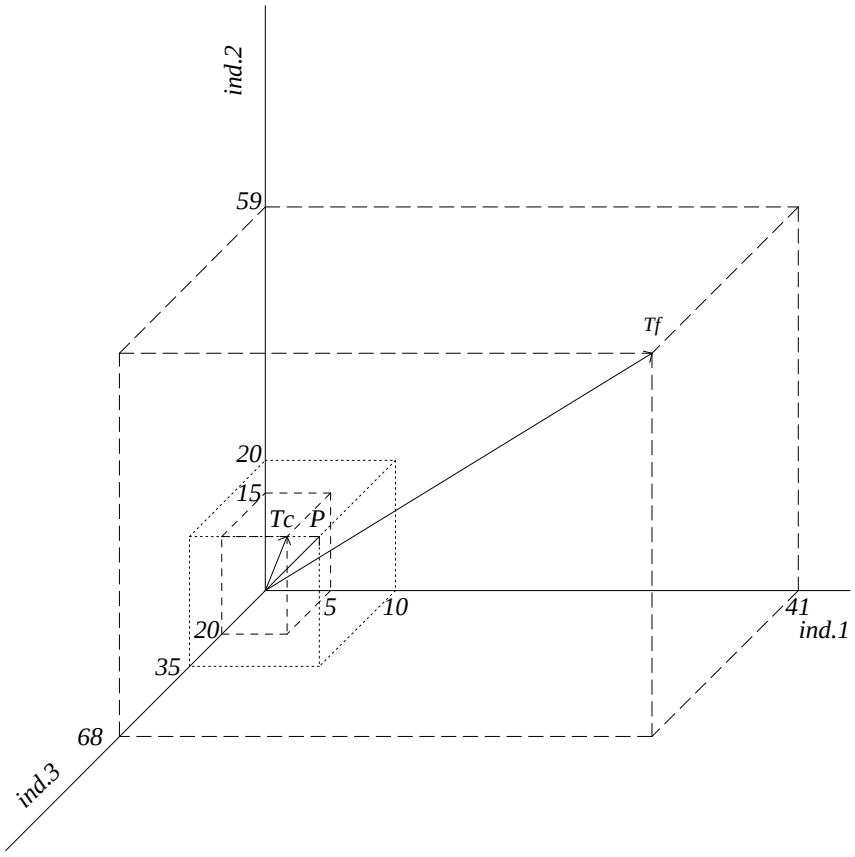
La liaison pression-température



La liaison pression-température

On représente le vecteur des pressions et le vecteur des températures exprimées en degrés centigrades.

La liaison pression-température



La liaison pression-température

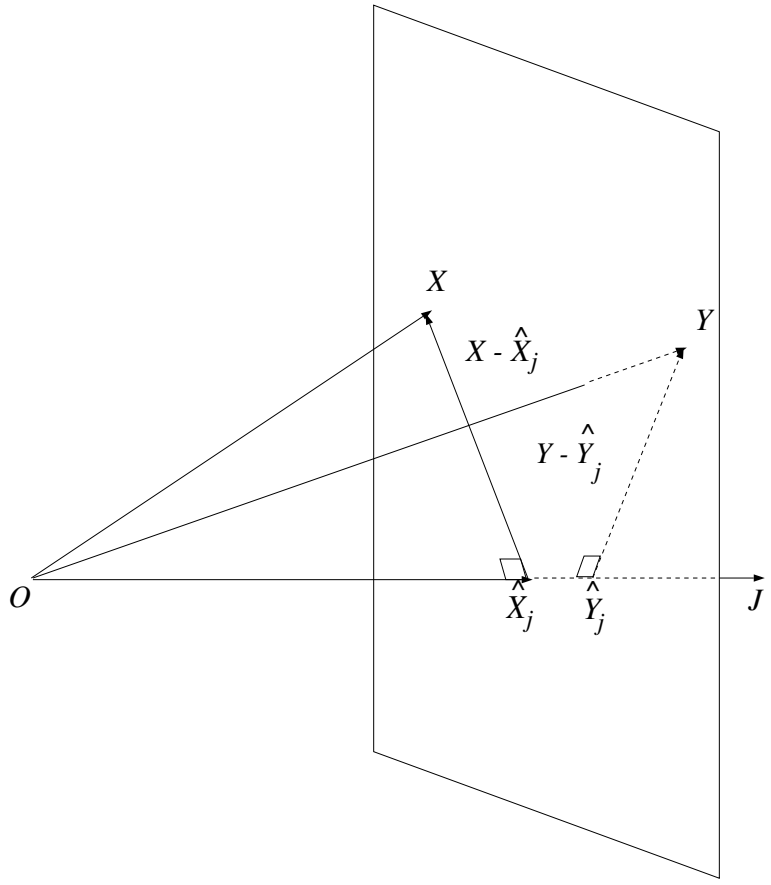
- On représente le vecteur des pressions et le vecteur des températures exprimées en degrés Fahrenheit.
- On fait remarquer que les deux vecteurs T_c et T_f ne sont pas colinéaires. Ils ne forment pas un même angle avec le vecteur P des pressions.

Donc, le changement d'unité de mesure de la température (une homothétie plus une translation) modifie la relation pression-température.

Les angles que font les vecteurs pression et températures ne sont pas de bonnes mesures de la liaison entre ces variables.

Dans ces conditions, que peut-on faire ?

Notion de corrélation
L'espace des variables centrées



Notion de corrélation
L'espace des variables centrées

- Existe-t-il un sous espace vectoriel dans lequel il est possible de trouver une mesure angulaire insensible à des translations et à des homothéties sur les variables ?
- La réponse est oui, c'est le sous espace des **variables centrées**.
- On va le montrer sur l'exemple de la pression et des températures.

La liaison pression-température
L'espace des variables centrées

	P	T_c	T_f
$ind.1$	10	5	41
$ind.2$	20	15	59
$ind.3$	35	20	69

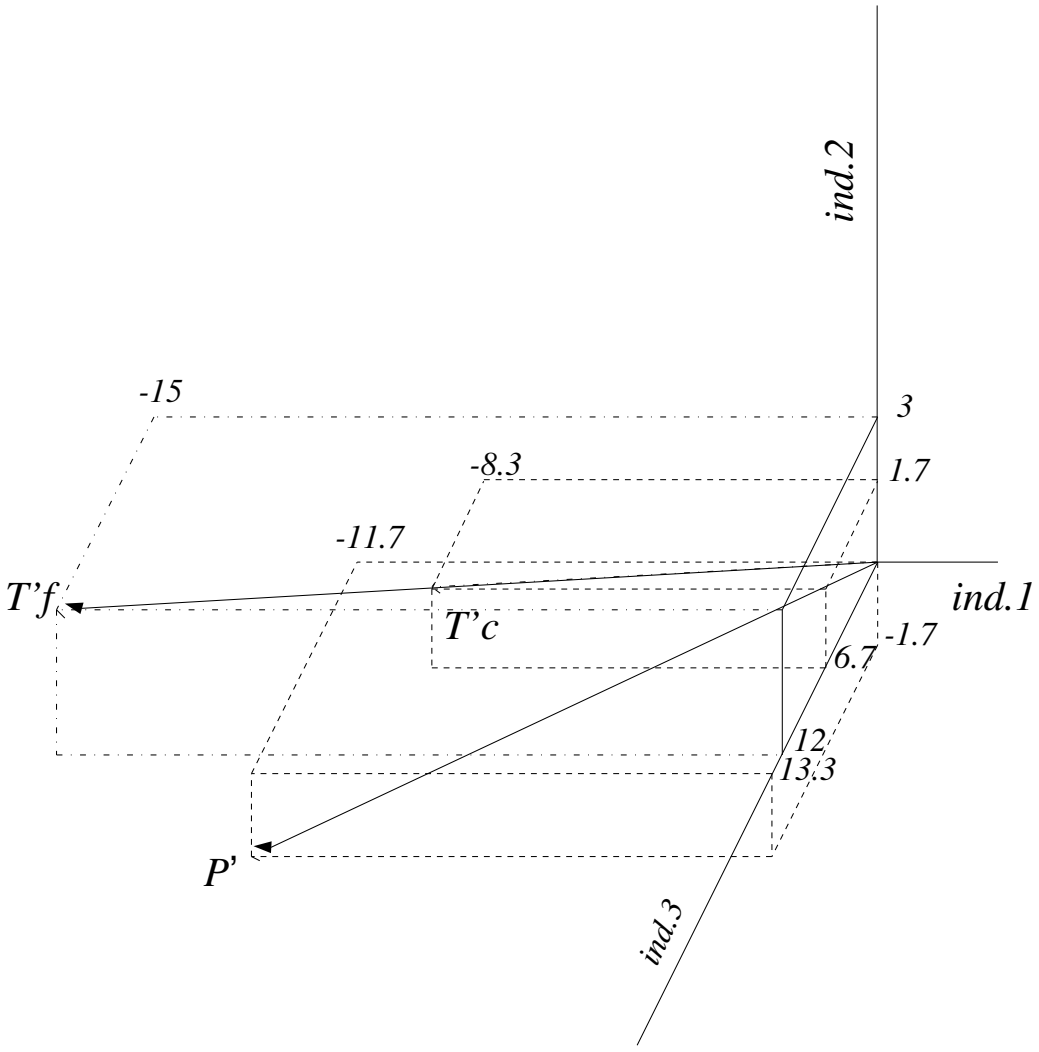
Centrons les variables

	P'	T'_c	T'_f
$ind.1$	-11.7	-8.4	-15
$ind.2$	-1.7	1.7	3
$ind.3$	13.4	6.7	12

La liaison pression-température
L'espace des variables centrées

- On présente le tableau des données initiales et celui des variables centrées.
- On va représenter sur un même graphique les variables centrées de la pression et des températures.

La liaison pression-température
L'espace des variables centrées

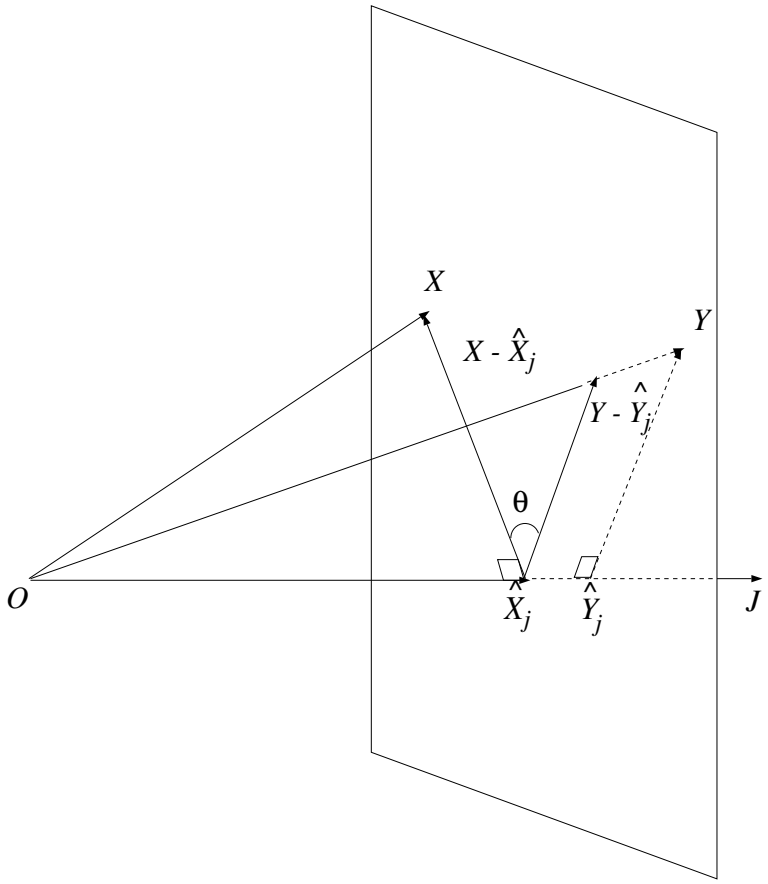


La liaison pression-température
L'espace des variables centrées

On observe que :

- Les variables centrées des températures exprimées dans deux unités différentes sont représentées par des vecteurs colinéaires.
- L'angle entre les variables centrées de la pression et des températures est insensible à une transformation sur les données initiales.

Notion de corrélation



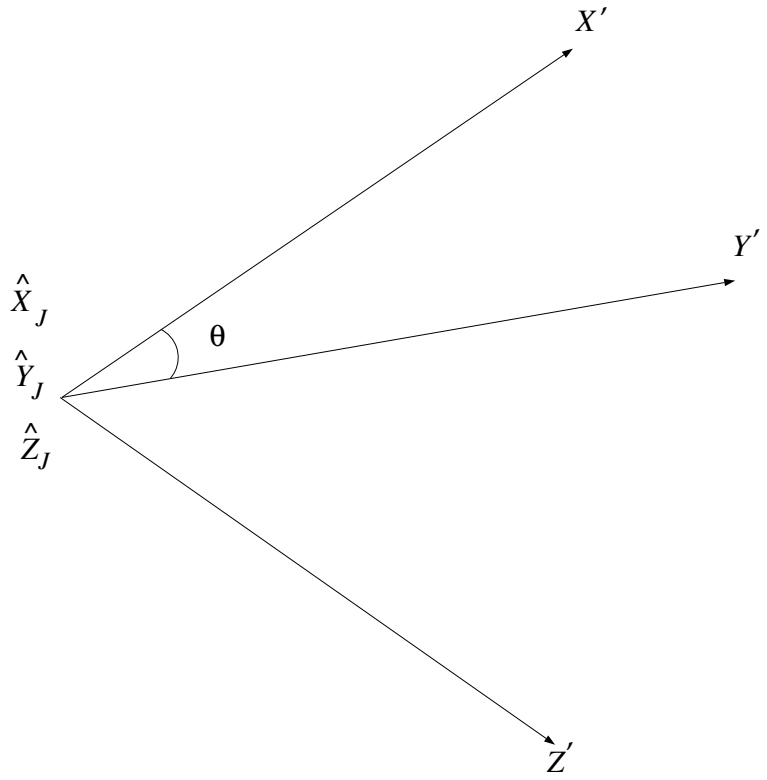
Notion de corrélation

- On se place dans le sous espace orthogonal à J et on utilise comme mesure de l'intensité de la relation entre X et Y le cosinus de l'angle θ entre $X - \hat{X}_J$ et $Y - \hat{Y}_J$.
- C'est cet angle θ que l'on a visualisé dans l'exemple pression-température, il est insensible à tout changement d'échelle et translation sur les X et les Y.
- Le sous espace des variables centrées est le sous espace complémentaire orthogonal à J.
- On a :

$$cos(\theta) = \frac{< X - \hat{X}_J, Y - \hat{Y}_J >}{\sqrt{< X - \hat{X}_J, X - \hat{X}_J >< Y - \hat{Y}_J, Y - \hat{Y}_J >}}$$

$$cos(\theta) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \rho(X,Y)$$

Notion de corrélation

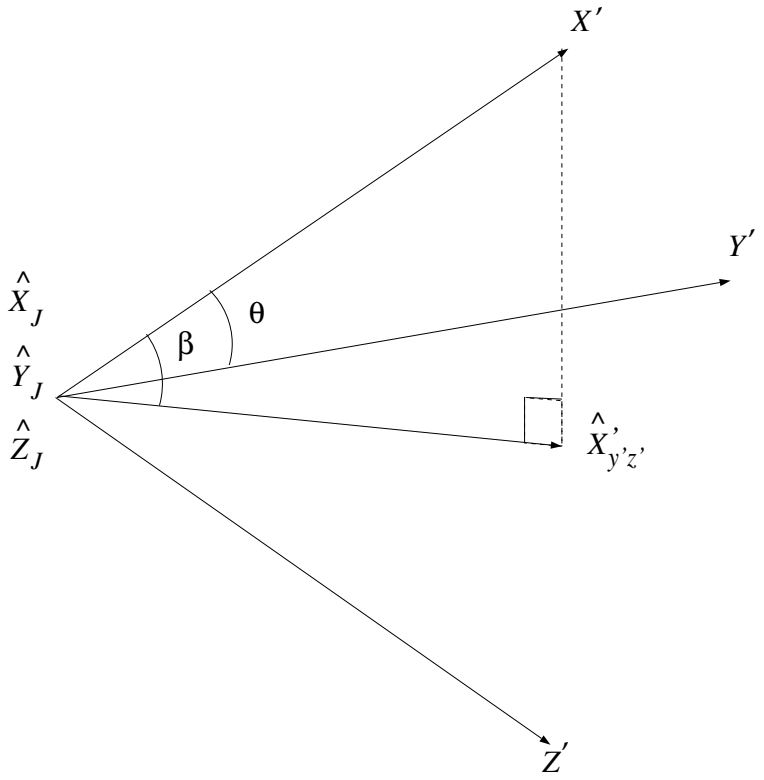


Notion de corrélation (Corrélation simple)

On veut visualiser les différents coefficients de corrélation (**simple**, **multiple** et **partielle**). Pour cela :

- On choisit 3 variables X, Y et Z et on se place dans l'espace des variables centrées. Les vecteurs représentés sont X', Y' et Z'.
- L'origine des vecteurs X', Y' et Z' est le point $(\hat{X}_J, \hat{Y}_J, \hat{Z}_J)$.
- Les angles que forment ces vecteurs entre eux ou avec des sous espaces vectoriels engendrés par ces vecteurs sont insensibles à des changements d'échelle et à des translations.
- Ici, on visualise θ l'angle dont le cosinus représente le coefficient de corrélation simple.

Notion de corrélation

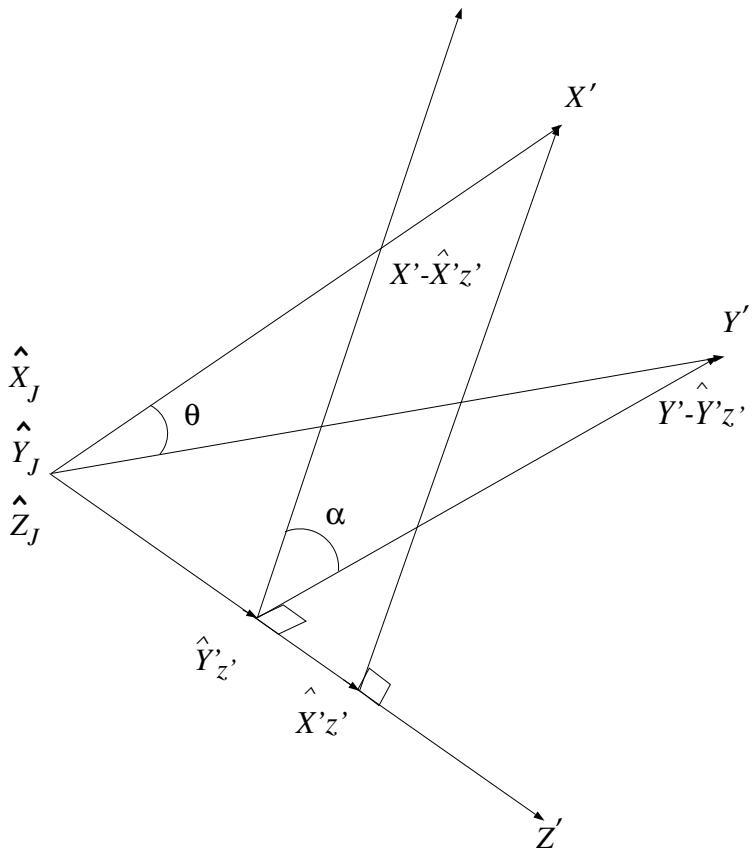


Notion de corrélation (Corrélation multiple)

Ici, on visualise le coefficient de corrélation **simple** et le coefficient de corrélation **multiple**.

- On note $\hat{X'}_{Y'Z'}$ la projection orthogonale du vecteur X' sur le sous espace (un plan) engendré par les vecteurs Y' et Z'.
- Le cosinus de l'angle β entre X' et $\hat{X'}_{Y'Z'}$ représente le coefficient de corrélation multiple entre X et le couple (Y, Z).

Notion de corrélation



Notion de corrélation (corrélation partielle)

Ici, on visualise le coefficient de corrélation **simple** et le coefficient de corrélation **partielle**.

- On note $\hat{X'}_{Z'}$ et $\hat{Y'}_{Z'}$ les projections orthogonales de X' et Y' sur Z'.
- Par définition, le coefficient de corrélation partielle entre X et Y est le coefficient de corrélation entre X et Y lorsque l'effet de Z est éliminé.
- C'est donc le cosinus de l'angle α entre $X' - \hat{X'}_{Z'}$ et $Y' - \hat{Y'}_{Z'}$ qui représente le coefficient de corrélation partielle car ces vecteurs sont orthogonaux à Z' (leur corrélation avec Z' est nulle).

3 LES TESTS STATISTIQUES
Comparaison de deux populations

- P_1 et P_2 deux populations normales $N(\mu_1, \sigma^2)$ et $N(\mu_2, \sigma^2)$
- Y le vecteur des observations (n_1 pour P_1 et n_2 pour P_2)
- J_1 et J_2 deux vecteurs orthogonaux tels que :

$$< J_1, J_2 >= 0 \text{ et } J = J_1 + J_2$$

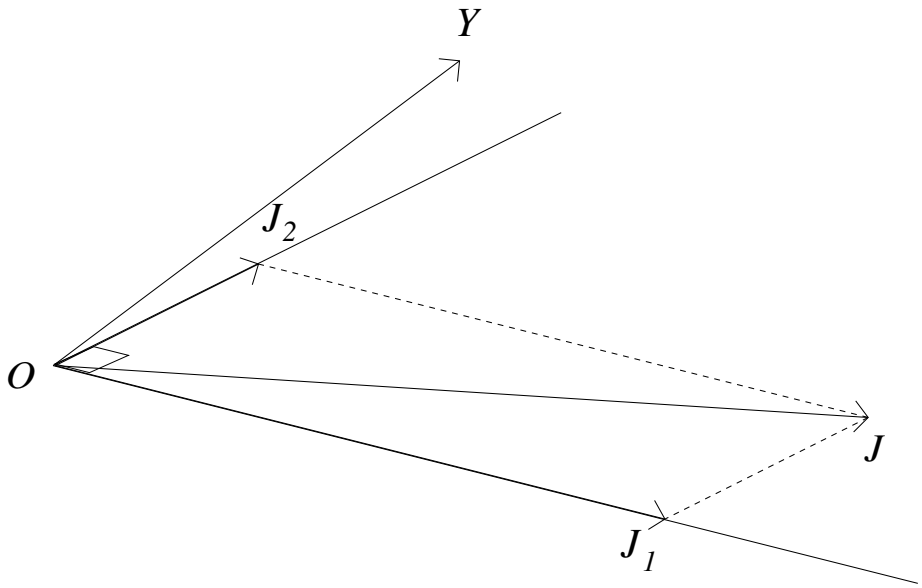
$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} \quad J_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad J_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad J = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

3 LES TESTS STATISTIQUES
Comparaison de deux populations
(*cf.* paragraphe 2.1.4)

Soient P_1 et P_2 deux populations normales de même variance σ^2 et de moyennes μ_1 et μ_2.

- On veut comparer ces deux populations c'est-à-dire répondre à la question : ces deux populations ont-elles des moyennes égales ($\mu_1 = \mu_2 = \mu$) ou ont-elles des moyennes différentes ($\mu_1 \neq \mu_2$) ?
- Pour répondre à cette question, on se place dans l'espace des variables de dimension $n = n_1 + n_2$.

Comparaison de deux populations

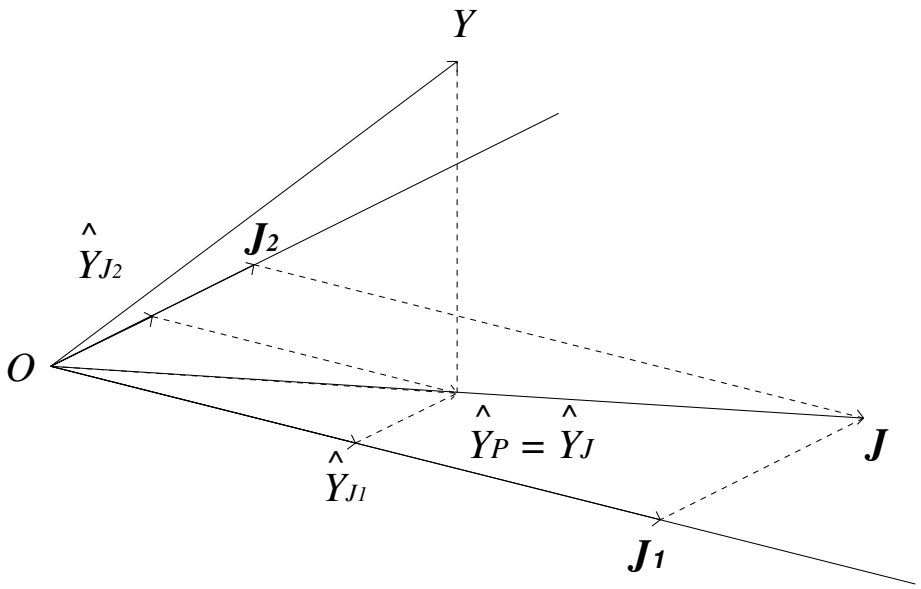


Comparaison de deux populations (Représentation des espaces)

- Pour chaque population (P_1 et P_2), les meilleurs résumés des observations sont les moyennes.
- Les moyennes sont les projections orthogonales de Y sur J_1 et J_2. Les sous espaces J_1 et J_2 sont orthogonaux ($< J_1, J_2 >= 0$).
- Le meilleur résumé de l'ensemble des données est la moyenne générale qui est la projection orthogonale de Y sur J.
- On se pose la question :

Les moyennes des deux populations sont-elles égales ?

Comparaison de deux populations (Les deux populations sont identiques)

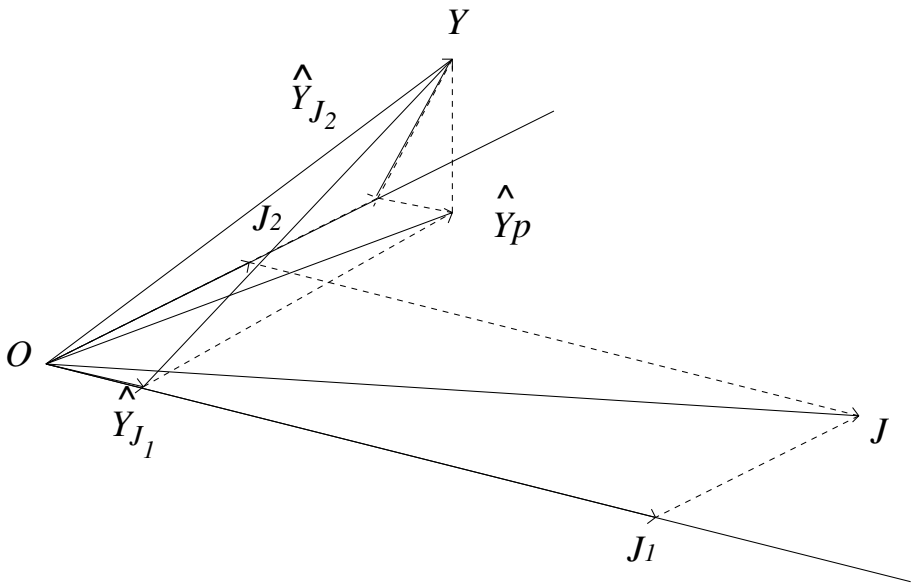


Comparaison de deux populations (Les deux populations sont identiques)

Si les deux populations P_1 et P_2 sont identiques :

- La moyenne des observations de P_1 est égale à la moyenne des observations de P_2 et ces deux moyennes sont égales à la moyenne calculée sur l'ensemble des observations.
- Sous cette condition, la projection orthogonale du vecteur Y des observations est $\hat{Y}_P = \hat{Y}_J$. Autrement dit, Y se projette sur J contenu dans P.

Comparaison de deux populations (Les deux populations sont différentes)



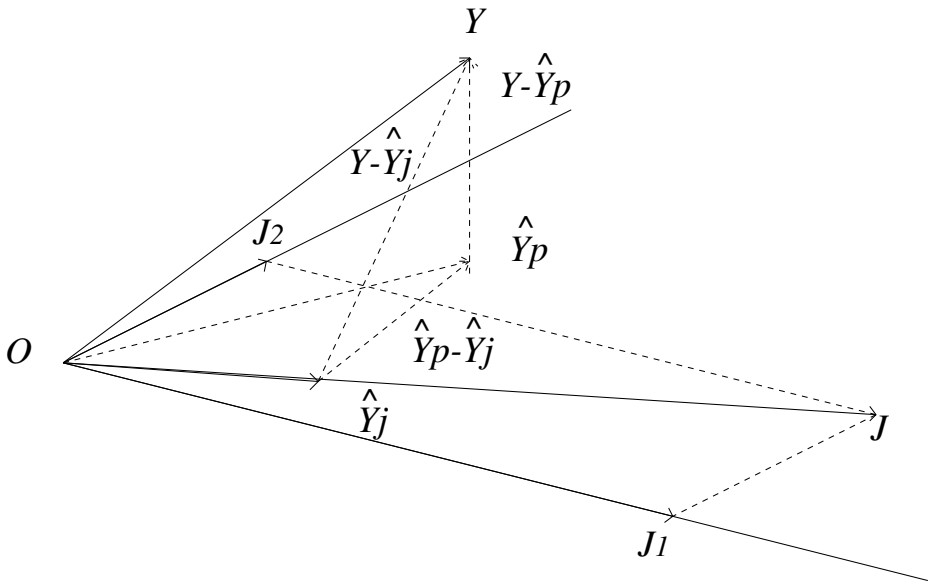
Comparaison de deux populations (Les deux populations sont différentes)

Si les deux populations P_1 et P_2 sont différentes :

- La moyenne des observations de P_1 est différente de la moyenne des observations de P_2 et ces deux moyennes sont différentes de la moyenne calculée sur l'ensemble des observations.
- Sous cette condition, la projection orthogonale du vecteur Y des observations est $\hat{Y}_P$. Ici, Y ne se projette pas sur la bissectrice J de l'angle (J_1, J_2) contenu dans P.
- La distance de $\hat{Y}_P$ à J sera d'autant plus grande que les deux moyennes seront différentes.
- On se pose la question :

Comment mesurer l'écart entre les moyennes des populations P_1 et P_2 ?

Comparaison de deux populations (Notion de test)



Comparaison de deux populations (Notion de test)

La différence entre les deux moyennes sera d'autant plus grande que $\hat{Y}_P$ sera plus éloigné de J.

- L'éloignement est mesuré par la distance $|\hat{Y}_p - \hat{Y}_J|$. Le vecteur $\hat{Y}_p - \hat{Y}_J$ est dans un sous espace affine du sous espace engendré par J_1 et J_2. Il est de dimension 1 et est orthogonal à J.
- La distance $|\hat{Y}_p - \hat{Y}_J|$ doit être appréciée par rapport à l'erreur $|Y - \hat{Y}_P|$ que l'on commet quand on projette Y sur le sous espace engendré par J_1 et J_2. Le vecteur $Y - \hat{Y}_P$ est dans l'orthogonal au sous espace engendré par J_1 et J_2, il a pour dimension $n_1 + n_2 - 2$.
- Pour apprécier la longueur du vecteur $\hat{Y}_P - \hat{Y}_J$ par rapport à celle de $Y - \hat{Y}_P$ il faut tenir compte des dimensions des espaces qui les contiennent (les degrés de liberté).

$$\text{On forme le rapport } : F = \frac{\frac{|\hat{Y}_P - \hat{Y}_J|^2}{1}}{\frac{|Y - \hat{Y}_J|^2}{n_1 + n_2 - 2}}$$

4 Résumé

	Représentation géométrique	**Expression statistique**
Vecteur X	segment orienté de O vers X	$x_1, \cdots, x_n$
Vecteur J	segment orienté de O vers J	$1, \cdots, 1$
Vecteur $\hat{X}_J = \bar{x}J$	Projeté orthogonal de X sur J	moyenne
$\lvert X - \hat{X}_J \rvert^2$	Carré distance entre X et $\hat{X}_J$	$(n-1)\times$ variance

4 Résumé

Représentation géométrique	Expression statistique
$\frac{< X - \hat{X}_J, Y - \hat{Y}_J >}{\lvert X - \hat{X}_J \rvert \lvert Y - \hat{Y}_J \rvert}$	
Cosinus angle $X - \hat{X}_J$ et $Y - \hat{Y}_J$	Corrélation simple entre X et Y
$\frac{\frac{\lvert \hat{Y}_P - \hat{Y}_J \rvert^2}{1}}{\frac{\lvert Y - \hat{Y}_P \rvert^2}{n_1 + n_2 - 2}}$	
Rapport carrés distances pondérées par d.d.l.	Fisher à 1 et $n_1 + n_2 - 2$ d.d.l.