

## LES FICHES CUISINE DE TONTON LULU

QU'EST CE QU'UN PETIT GROS ?

ou

Les statistiques ça peut aussi faire réfléchir !

par Jacques LUBCZANSKI

Nous publions dans ce numéro la première de deux "fiches cuisine" de J. Lubczanski qui anime à L'ENSET le séminaire "Enseignement et mathématiques". L'ensemble de ces fiches sera publié chez Cédic en 1986. L'Ouvert remercie l'auteur d'avoir bien voulu l'autoriser à publier celles relatives aux statistiques.

### L'OUVERT

Non, les statistiques ne sont pas forcément d'ennuyeuses colonnes de chiffres ou des graphiques hachurés pleins de pourcentages.

Et pas non plus un aléa de la théorie des probabilités, plus ou moins compréhensible. Mais c'est aussi l'occasion de répondre à des questions, à l'aide d'outils mathématiques simples et variés : raisonnement autant que calcul ; géométrie autant qu'algèbre.

La question du jour : "qu'est-ce qu'un petit gros ?" est celle-ci : une population étant donnée par les poids et tailles des individus qui la composent, peut-on déterminer un sous-ensemble des "petits gros" ?

Avec quel critère mathématique ?

Avec quelle validité ?

Le problème évoqué ici est celui de la classification d'individus à deux variables, d'être à deux dimensions ; ce problème se pose non seulement aux statisticiens mais aussi à tous les scientifiques (biologistes, médecins, sociologues...), et souvent avec plus de deux variables : les méthodes sont alors les mêmes que celles que je vais exposer, mais on "voit" moins bien ce qu'on fait.

AU DEBUT ETAIT LE NUAGE... (représentation des données) 

Pour se fixer les idées, et pour que vous puissiez vérifier les calculs, donnons-nous une fois pour toutes une population qu'on va étudier dans tout ce qui suit. Prenons-la de faible effectif : 10 individus répondant aux doux noms de A, B, C, D, E, F, G, H, I et J dont voici les poids et tailles :

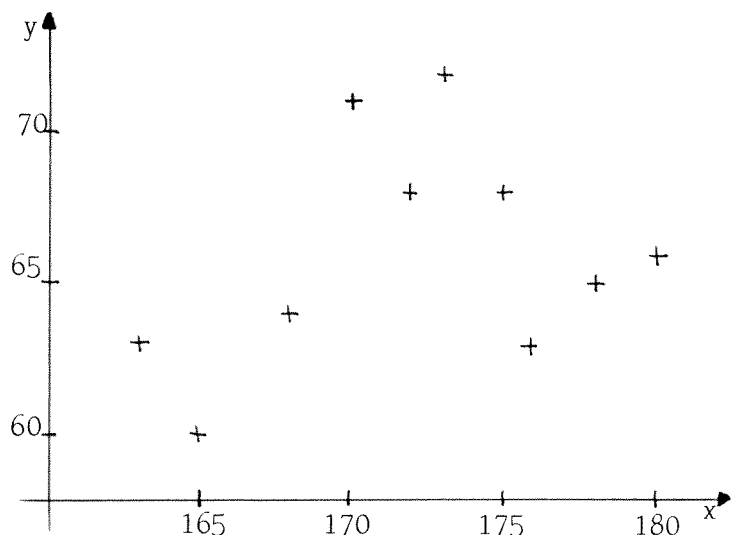
| nom         | A   | B   | C   | D   | E   | F   | G   | H   | I   | J   |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| taille (cm) | 163 | 168 | 170 | 165 | 173 | 172 | 176 | 178 | 180 | 175 |
| poids (kg)  | 63  | 64  | 71  | 60  | 72  | 68  | 63  | 65  | 66  | 68  |

On commence par faire un dessin, ou plutôt une carte de la population : chaque individu est représenté par un point de coordonnées  $(\begin{smallmatrix} x \\ y \end{smallmatrix})$ , où  $x$  est la taille et  $y$  le poids de l'individu.

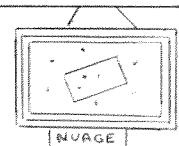
(On a donc fait le choix d'une échelle).

L'ensemble de ces points s'appelle le nuage des points représentant notre population.

La première préoccupation du statisticien est de faire une description de ce nuage. Et décrire, en l'occurrence, signifie résumer : on va résumer le nuage par une figure géométrique plus simple...



LE NUAGE VU PAR L'ART ABSTRAIT ! (Réduction des données)



### 1. Centre du nuage

Le centre du nuage en sera le point moyen : celui dont l'abscisse est la moyenne des abscisses et dont l'ordonnée est la moyenne des ordonnées.

Dans notre exemple si on note  $\bar{x}$  et  $\bar{y}$  ces valeurs, on obtient  $\bar{x} = 172$  et  $\bar{y} = 66$ .

Soit  $\Omega$  le centre du nuage.

Désormais, on va se placer dans un repère d'origine  $\Omega$ , d'axes parallèles aux axes initiaux.

Si  $X$  et  $Y$  sont les coordonnées d'un point dans ce nouveau repère, on a donc :

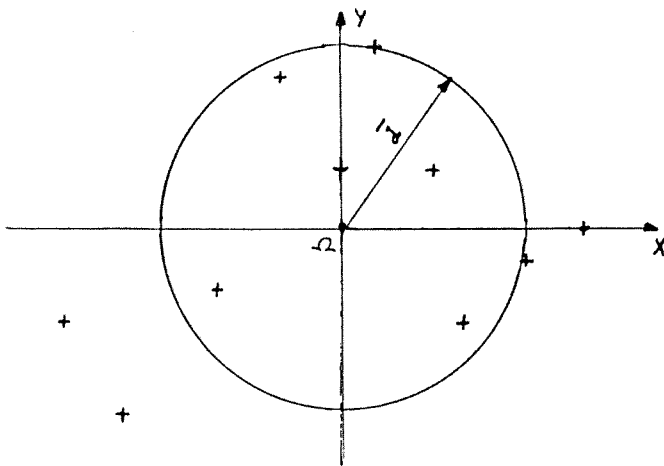
$$X = x - \bar{x} \text{ et } Y = y - \bar{y}.$$

On dit aussi qu'on a centré les variables  $x$  et  $y$  :  $X$  et  $Y$  sont des variables centrées ; leur moyenne est nulle.

Tout se passe comme si on repartait des données en  $X$  et  $Y$  :

| nom | A  | B  | C  | D  | E  | F  | G  | H  | I | J |
|-----|----|----|----|----|----|----|----|----|---|---|
| X   | -9 | -4 | -2 | -7 | +1 | 0  | 4  | 6  | 8 | 3 |
| Y   | -3 | -2 | +5 | -6 | +6 | +2 | -3 | -1 | 0 | 2 |

## 2. Résumé du nuage par un cercle



L'idée est simple : on calcule les distances  $\Omega A, \Omega B, \Omega C \dots \Omega J$ , puis on calcule leur moyenne  $\bar{d}$  :

Le cercle de centre  $\Omega$  et de rayon  $\bar{d}$  est un résumé du nuage.

On trouve ici  $\bar{d} \approx 5,9$ .

Si pour calculer la moyenne des distances, on n'utilise plus la moyenne arithmétique mais la moyenne quadratique (c'est-à-dire moyenne des carrés puis racine de cette moyenne), on trouve alors  $\bar{d}' \approx 6,4$ .

Et on peut prouver que  $\Omega$  est le point du plan qui minimise  $\bar{d}'$ . La valeur minimale de  $\bar{d}'$  s'appelle la variance totale du nuage. On y reviendra plus loin.

## 3. Résumé par un rectangle parallèle aux axes

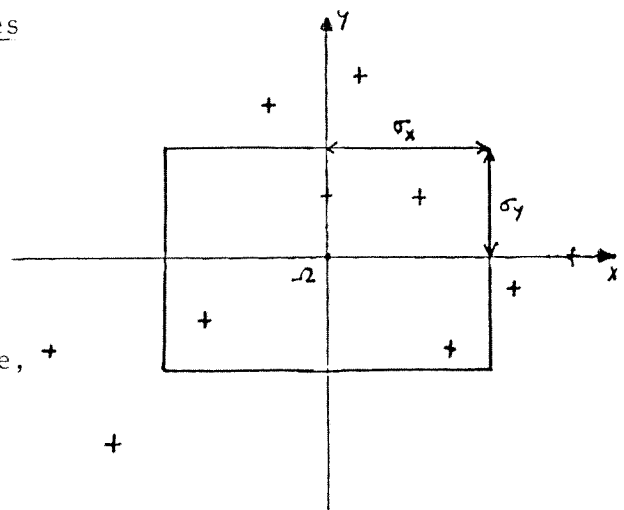
Au lieu de chercher la distance moyenne à l'origine, on cherche les distances moyennes à chacun des axes.

Pour ne pas à avoir à faire de calculs sur les valeurs absolues  $|X|$  et  $|Y|$ , on utilise la moyenne quadratique :

on calcule la moyenne des  $X^2$ , puis sa racine, notée  $\sigma_x$ .

Et de même pour les  $Y^2$ , d'où  $\sigma_y$ .

Ici on trouve :  $\sigma_x \approx 5,25$  et  $\sigma_y \approx 3,6$ .



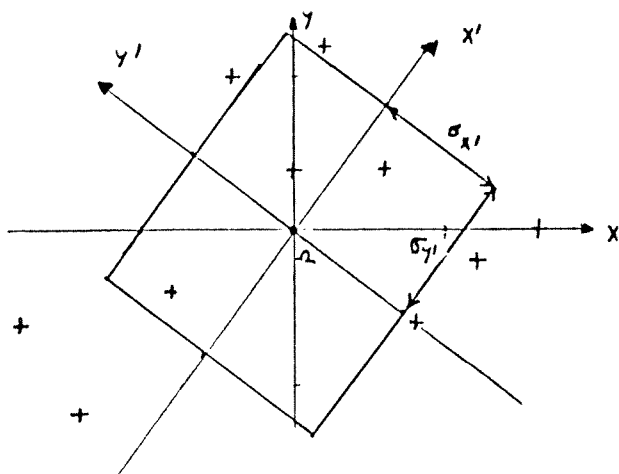
Le rectangle centré en  $\Omega$ , de côtés  $2\sigma_x$  et  $2\sigma_y$  résume le nuage.

$\sigma_x$  s'appelle l'écart type de X (et aussi de x);  $\sigma_y$  l'écart type de Y (et de y).

Les carrés des écarts types  $\sigma_x^2$  et  $\sigma_y^2$  sont les variances de X et Y (et aussi de x et y).

La somme  $\sigma_x^2 + \sigma_y^2$  est la variance totale du nuage.

#### 4. Résumé par un rectangle quelconque



On continue à chercher les distances moyennes à chacun des axes, mais pour des axes quelconques passant par  $\Omega$  : en d'autres termes, on reprend le calcul du paragraphe précédent, après avoir fait tourner les axes.

Faisons-le sur un exemple : prenons comme nouvel axe  $\Omega X'$  la droite  $Y = 4/3X$  et pour nouvel axe  $\Omega Y'$  la droite  $Y = -3/4X$ .

$$\text{On a : } X' = 3/5X + 4/5Y$$

$$Y' = -4/5X + 3/5Y$$

ce qui permet de calculer les coordonnées des points du nuage dans le repère ( $\Omega$ ,  $\Omega X'$ ,  $\Omega Y'$ ) :

| nom  | A    | B  | C   | D  | E   | F   | G  | H    | I    | J    |
|------|------|----|-----|----|-----|-----|----|------|------|------|
| $X'$ | -7,8 | -4 | 2,8 | -9 | 5,4 | 1,6 | 0  | 2,8  | 4,8  | 3,4  |
| $Y'$ | 5,4  | 2  | 4,6 | 2  | 2,8 | 1,2 | -5 | -5,4 | -6,4 | -1,2 |

Alors le calcul donne  $\sigma_{X'} \approx 4,9$  et  $\sigma_{Y'} \approx 4,25$  : d'où un nouveau rectangle.

. Remarquons d'abord qu'on aurait pu obtenir ces dernières valeurs beaucoup plus vite :

$$\sigma_{X'}^2 = \overline{X'^2} = \overline{(3/5 X + 4/5 Y)^2} = \frac{9}{25} \overline{X^2} + \frac{24}{25} \overline{XY} + \frac{16}{25} \overline{Y^2} = \frac{9}{25} \sigma_X^2 + \frac{24}{25} \overline{XY} + \frac{16}{25} \sigma_Y^2$$

Il aurait donc suffi de calculer la moyenne des XY : ce nombre s'appelle la covariance de X et Y et se note  $\sigma_{XY}$ .

Ici  $\sigma_{XY} = 6,1$  d'où  $\sigma_{X'}^2 \approx \frac{9}{25} \times 27,6 + \frac{24}{25} \times 6,1 + \frac{16}{25} \times 12,8 \approx 24,1$  on retrouve

bien  $\sigma_{X'} \approx 4,9$ .

. Comparons le rectangle obtenu à celui du paragraphe précédent : il est plus "carré", moins aplati.

Lequel, de ces deux rectangles, donne une meilleure description du nuage ?

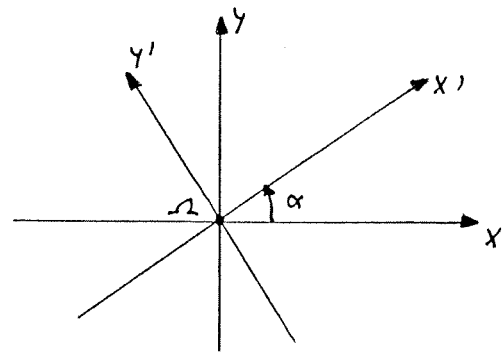
Pour le statisticien, la meilleure description est celle qui est la plus interprétable ; si, par exemple, le rectangle était complètement aplati, réduit à un segment de droite, l'interprétation serait aisée : il existerait alors une relation du premier degré entre X et Y (l'équation de la droite supportant le rectangle) : la connaissance de X nous donnerait Y et vice versa. Dans ce cas (idéal) on dirait que X et Y sont parfaitement corrélés (et alors le nombre  $\frac{\sigma_{XY}}{\sigma_X \sigma_Y}$  qui s'appelle le coefficient de corrélation de X Y vaudrait 1 ou -1).

C'est donc dans cette direction qu'on va chercher : on va faire tourner les axes de façon à obtenir le rectangle le plus aplati possible.

Ici notre choix des axes  $\Omega X'$  et  $\Omega Y'$  n'est pas terrible, puisque le rectangle obtenu est moins aplati que celui qu'on avait.

##### 5. Recherche du rectangle le plus aplati

Soit  $\alpha$  une mesure de l'angle ( $\Omega X, \Omega X'$ ) :  
 exprimons, en fonction des données et de  $\alpha$   
 les côtés du rectangle qui résume le nuage,  
 dans le repère ( $\Omega X', \Omega Y'$ ) :



$$\text{On a : } X' = \cos \alpha \, X + \sin \alpha \, Y$$

$$Y' = -\sin \alpha \, X + \cos \alpha \, Y$$

$$\sigma_{X'}^2 = \overline{X'^2} = \overline{(X \cos \alpha + Y \sin \alpha)^2} = \overline{X^2 \cos^2 \alpha + 2XY \sin \alpha \cos \alpha + Y^2 \sin^2 \alpha}$$

$$\text{Or } \cos^2 \alpha = \frac{1}{2} + \frac{1}{2} \cos 2\alpha ; 2 \sin \alpha \cos \alpha = \sin 2\alpha \text{ et } \sin^2 \alpha = \frac{1}{2} - \frac{1}{2} \cos 2\alpha$$

$$\text{D'où } \sigma_{X'}^2 = \frac{1}{2} (\sigma_X^2 + \sigma_Y^2) + \sigma_{XY} \sin 2\alpha + \frac{1}{2} (\sigma_X^2 - \sigma_Y^2) \cos 2\alpha$$

$\sigma_{X'}$  prendra une valeur extrême en même temps que  $\sigma_{XY} \sin 2\alpha + \frac{1}{2} (\sigma_X^2 - \sigma_Y^2) \cos 2\alpha$ .  
 Cette expression est du type :  $a \sin 2\alpha + b \cos 2\alpha$ .

On peut donc la transformer en  $A \cos (2\alpha - \varphi)$  où  $A = \sqrt{a^2 + b^2}$  ;  $\cos \varphi = \frac{b}{A}$  ;  $\sin \varphi = \frac{a}{A}$ .

Alors  $\sigma_{X'}$  sera minimum lorsque  $\cos (2\alpha - \varphi) = -1$  et maximum lorsque  $\cos (2\alpha - \varphi) = +1$ .

Pour connaître la valeur de  $\sigma_{Y'}$ , remarquons qu'il suffit de remplacer, dans l'expression de  $\sigma_{X'}$ ,  $\alpha$  par  $\alpha + \frac{\pi}{2}$  : alors  $2\alpha - \varphi$  est remplacé par  $2\alpha - \varphi + \pi$  : le cosinus qui valait -1 vaut 1 et celui qui valait 1 vaut -1.

Il s'ensuit que lorsque  $\sigma_{X'}^2$  est minimum,  $\sigma_{Y'}^2$  est maximum, et vice versa.

On a bien affaire au rectangle le plus aplati possible. Calculons-le pour notre population :

$$\begin{aligned}\sigma_{X'}^2 &= 20,2 + 6,1 \sin 2\alpha + 7,4 \cos 2\alpha \\ &= 20,2 + A \cos(2\alpha - \varphi)\end{aligned}$$

où  $A \approx 9,59$  et  $\varphi \approx 40^\circ$

$\sigma_{X'}^2$  est maximum quand  $2\alpha - \varphi = 0$

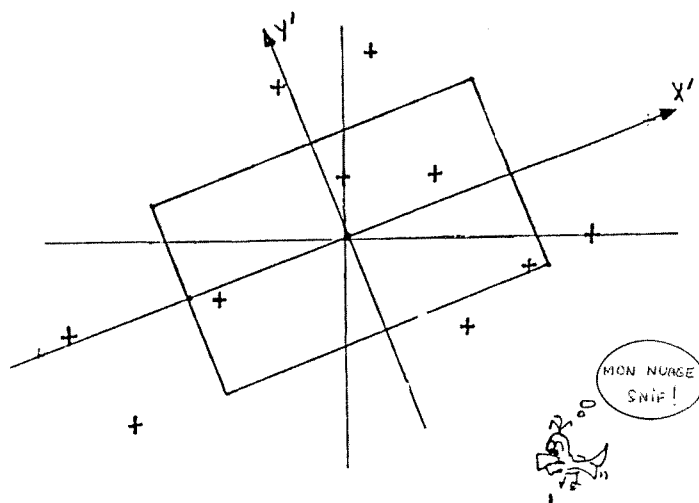
soit  $\alpha = \frac{\varphi}{2} \approx 20^\circ$

Alors  $\sigma_{X'}^2 \approx 20,2 + 9,59 \approx 29,79$

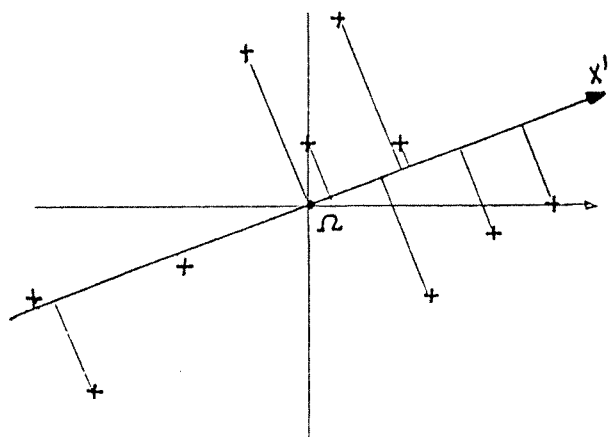
d'où  $\sigma_{X'} \approx 5,5$ .

Et  $\sigma_{Y'}^2 \approx 20,2 - 9,59 \approx 10,61$

d'où  $\sigma_{Y'} \approx 3,3$ .



## 6. Appellations contrôlées !



Avec la méthode exposée ci-dessus, nous entrons dans un domaine d'appellations contrôlées !

La droite  $\Omega$   $X'$  est la droite la plus proche des points du nuage : la distance moyenne (quadratique) des points du nuage à cette droite est la plus petite possible ( $\sigma_{Y'} \approx 3,3$ ).

Elle s'appelle droite des moindres carrés du nuage ou aussi droite de régression orthogonale. Nous verrons

au paragraphe suivant ce que peut être une régression pas orthogonale.

Si on prenait la distance moyenne (arithmétique), on trouverait une autre droite la plus proche des points du nuage. Mais le calcul fait intervenir le minimum d'une somme de valeurs absolues, qui est techniquement difficile à traiter.

Enfin, on peut remarquer que la somme  $\sigma_{X'}^2 + \sigma_{Y'}^2$  ne dépend pas de  $\alpha$  : elle vaut  $\sigma_X^2 + \sigma_Y^2$ . En d'autres termes, cette quantité, la variance totale du nuage, est invariante par rotation des axes, ce qui explique que quand  $\sigma_{X'}^2$  est maximum, alors  $\sigma_{Y'}^2$  est minimum. Cela signifie aussi que la diagonale de tous les rectangles a la même longueur  $2 \sqrt{\sigma_X^2 + \sigma_Y^2}$ .

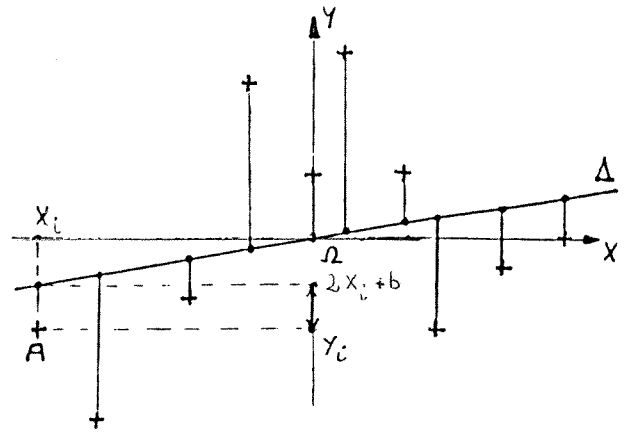
Pour les curieux, je signale que le nuage possède un autre invariant par rotation, qui est  $\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2$ . (Pour les savants : derrière ceci se cache une matrice et ses valeurs propres...).

### 7. Résumé par un parallélogramme

Comme dans le paragraphe précédent, on va chercher à minimiser la distance moyenne des points du nuage à une droite  $\Delta$ .

La différence tient à ce qu'on va mesurer les distances, non plus orthogonalement à  $\Delta$ , mais parallèlement à l'axe  $\Omega Y$ .

Ici encore, la moyenne quadratique va être utilisée, de préférence à la moyenne arithmétique.



Supposons que l'équation de  $\Delta$  soit du type  $Y = aX + b$ .

Si  $(X_i, Y_i)$  sont les coordonnées d'un point du nuage, la distance "verticale" de ce point à la droite  $\Delta$  est  $|Y_i - (aX_i + b)|$  : voir le dessin ci-dessus, pour le point A.

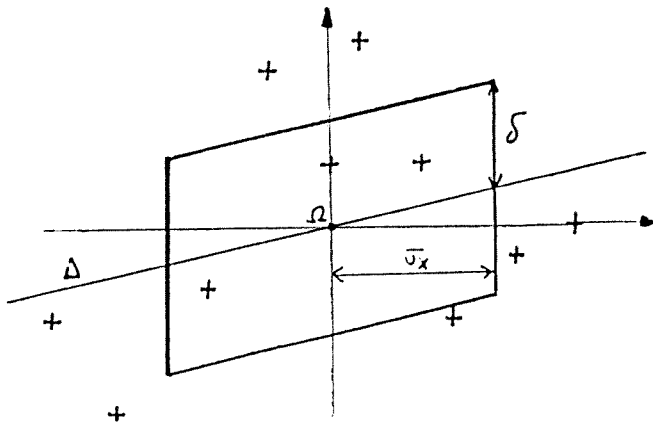
La moyenne des carrés de ces distances est :  $\overline{[Y - (aX + b)]^2} = \overline{Y^2 - 2Y(aX + b) + (aX + b)^2}$

Soit :  $Y^2 - 2aXY - 2bY + a^2 X^2 + 2abX + b^2$

Or  $\bar{Y} = \bar{X} = 0$ , on trouve donc :  $\sigma_Y^2 - 2a \sigma_{XY} + a^2 \sigma_X^2 + b^2$ .

Remarquons d'abord que, pour que cette expression soit minimale, il faut que  $b = 0$ , c'est-à-dire que  $\Delta$  passe par le centre du nuage  $\Omega$ , ce qui n'était pas a priori évident.

Ensuite, il nous reste un trinôme du second degré en  $a$ , qui admet un minimum pour  $a = \frac{\sigma_{XY}}{\sigma_X^2}$  : la droite  $\Delta$  cherchée a donc pour équation  $Y = \frac{\sigma_{XY}}{\sigma_X^2} \cdot X$ .



Le minimum vaut alors  $\sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$

et la distance moyenne est

$$\sigma = \sqrt{\sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}}$$

Dans notre exemple, on trouve  $Y \approx 0,22 X$  et la distance  $\sigma$  vaut à peu près 3,4.

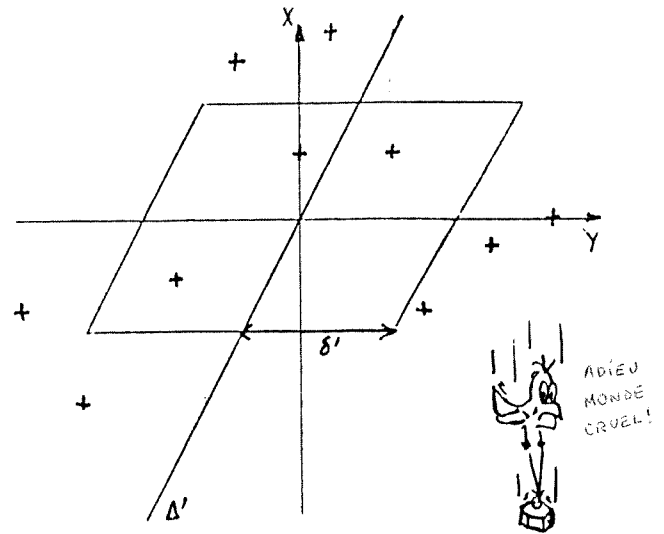
Alors c'est un parallélogramme centré en  $\Omega$  de côté vertical  $2\sigma$ , de dimension horizontale  $\sigma_X$ , qui résume le nuage.

L'appellation contrôlée de  $\Delta$  est "droite de régression de Y en X", et la méthode s'appelle méthode de régression linéaire.

Bien entendu, on peut échanger les rôles de X et Y et calculer la droite de régression de X en Y, dont l'équation est  $X = \frac{\sigma_{XY}}{\sigma_Y^2} Y$  et la distance moyenne est alors  $\sqrt{\sigma_X^2 - \frac{\sigma_{XY}^2}{\sigma_Y^2}}$ .

On obtient une autre droite et un autre parallélogramme pour résumer le nuage.  
 Dans notre exemple, on trouve  $X \approx 0,484$   
 et  $\sigma' \approx 5$ .

Voilà : on a fait le tour des principales  
 façons de décrire, de résumer un nuage.  
 Laquelle choisir ?  
 Et après cette phase de réduction des  
 données, quelle analyse peut-on faire ?



### CRITIQUE DE L'ART ABSTRAIT ! (Analyse des données)



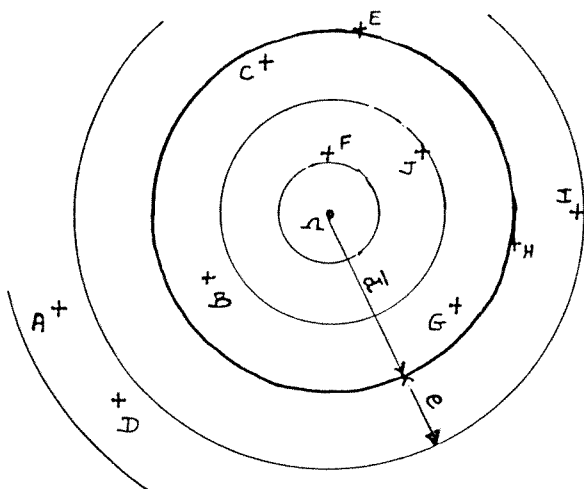
L'analyse des données, l'interprétation de la description qu'on a faite du nuage, consiste à répartir les individus de notre population en catégories.

On va chercher à réaliser une partition de la population en "classes".

Il existe de nombreuses méthodes ; limitons-nous aux méthodes issues des descriptions que nous avons faites ; les autres méthodes feront l'objet d'une prochaine "fiche cuisine".

### LE NUAGE COUPÉ EN TRANCHES (Analyse par lignes de niveau)

#### 1. Description par un cercle



On trace des cercles concentriques, de centre  $\Omega$ , et dont les rayons sont en progression arithmétique.

Ici, on a tracé le cercle "résumé" de rayon  $\bar{d}$  et les cercles de rayon  $\bar{d} - 2e$ ,  $\bar{d} - e$ ,  $\bar{d} + e$ ,  $\bar{d} + 2e$ .

On obtient la partition suivante :

$\{F, J\}$  ;  $\{B, C, G\}$  ;  $\{E, I, H\}$  ;  $\{A, D\}$  .

On peut classer les individus selon leur éloignement de  $\Omega$  :

$F ; J ; B ; C ; G ; H ; E ; I ; D ; A$ .

Au lieu de tracer les cercles, on peut se contenter de calculer les distances à  $\Omega$  : pour un point M de coordonnées  $\begin{pmatrix} X \\ Y \end{pmatrix}$ , on a  $\bar{d} = \sqrt{X^2 + Y^2}$ .

Cela nous permet de substituer les lignes de niveau à un indice  $\bar{d}$ .

Pour un individu de taille  $T$  (cm) et de poids  $P$  (kg), on calculera donc l'indice

$$\bar{d} = \sqrt{(T - 172)^2 + (P - 66)^2}$$

soit encore  $\bar{d} = \sqrt{T^2 - 344T + P^2 - 132P + 33940}$  : de quoi impressionner les foules !

Derrière cette formule savante se cache alors le postulat qu'il y a un poids idéal et une taille idéale dont il "faut" se rapprocher : plus votre indice est élevé, plus vous êtes anormal !

Cette affirmation d'un idéal unique, semble passée de mode, pour ce qui est du poids et de la taille. Elle ne permet pas de séparer les "petits gros" des autres.

## 2. Description par un rectangle :

Qu'a-t-on fait ?

On a observé l'ombre du nuage sur un axe qui tourne.

Ou, ce qui revient au même, on a tourné autour du nuage, et on a observé sa silhouette.

Celle-ci se déforme, dépend de notre position, de notre point de vue.

En particulier il y a deux positions extrêmes :

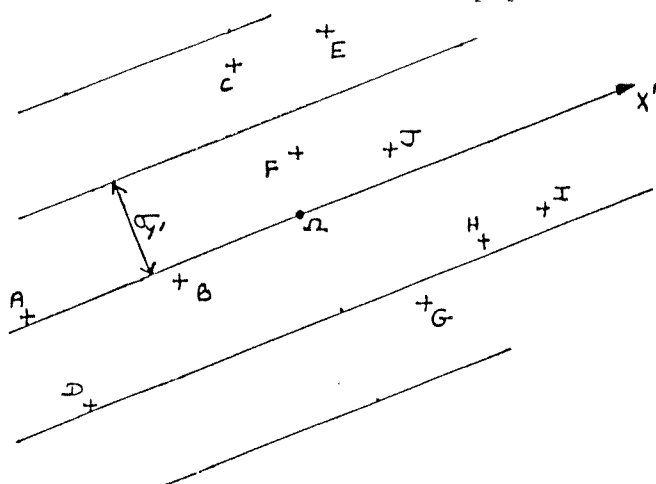
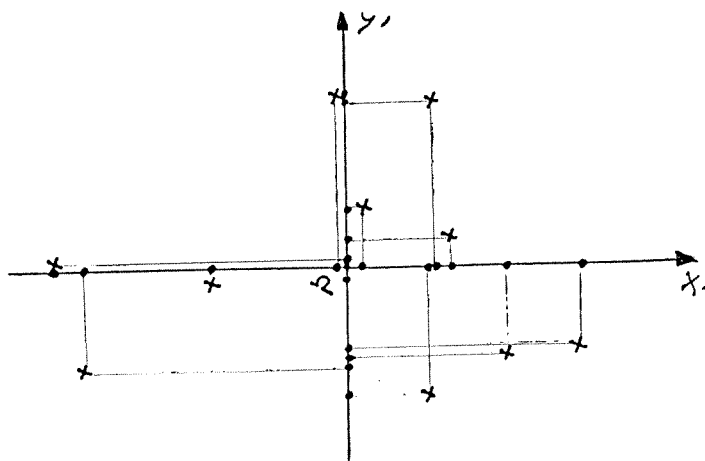
l'une où la silhouette est la plus ramassée, et l'autre où elle est, au contraire, la plus dispersée.

Par contre, on a vu que la variance totale ne dépendait pas du "point de vue" adopté.

La valeur maximale de la variance de l'ombre du nuage,  $\sigma_{X'}$ , valait 29,79, soit 74 % de la variance totale : on dit que l'axe  $\Omega X'$  explique 74 % de la variance du nuage.

Et on ne peut pas faire mieux.

Quant aux 26 % de variance qui restent, ce sont ceux qui vont nous permettre de classer les individus de notre population !



Les lignes de niveau sont des droites parallèles à  $\Omega X'$ , distantes entre elles de  $\sigma_{Y'}$ .

On obtient la partition suivante :

$\{C, E\}$  ;  $\{A, F, J\}$  ;  $\{D, B, H, I\}$  ;  $\{G\}$  .

Et la classification (selon les valeurs de  $|Y'|$ ) :

|       |      |      |      |      |      |      |      |      |      |      |
|-------|------|------|------|------|------|------|------|------|------|------|
| nom : | A    | B    | J    | F    | I    | H    | D    | G    | E    | C    |
| Y' :  | 0,26 | 0,51 | 0,85 | 1,88 | 2,74 | 2,99 | 3,24 | 4,19 | 5,30 | 5,38 |

Le principe est simple : une fois qu'on a "expliqué" la ressemblance entre les individus par la direction  $\Omega X'$ , on les classe par leur distance à cet axe.

En d'autres termes, l'équation de  $\Omega X'$  représente alors la relation "idéale" entre X et Y : à savoir dans notre cas  $Y' = 0 \Leftrightarrow -\sin \alpha X + \cos \alpha Y = 0 \Leftrightarrow 0,940Y - 0,342X = 0$ .

On peut ici encore remplacer le tracé des lignes de niveau par le calcul d'un indice  $|Y'|$  :

$$|Y'| = |0,940 Y - 0,342X| = |0,940(P - 66) - 0,342 (T - 172)| \quad \text{si P est le poids} \\ = |0,940 P - 0,342 T - 3,192| \quad \text{et T la taille.}$$

Derrière cette formule (savante ?) se cache donc le postulat qu'il existe une relation linéaire entre la taille et le poids, dont il faut se rapprocher.

Là encore, plus votre indice est élevé, plus vous êtes anormal !

Ce genre de formule est très répandu (en général sans les valeurs absolues), sous la forme  $P = aT + b$  représentant le poids "idéal" en fonction de la taille. Les nombres a et b dépendent évidemment de l'échantillon qui a servi à le calculer : est-il représentatif ? et de quoi ?

Peut-on définir un ensemble des petits gros ?

Oui, et c'est facile :

Il suffit de trouver un indice qui croît avec le poids et décroît avec la taille.

Or la coordonnée Y' vérifie ces propriétés :

Plus Y' est élevé, plus on est petitgros ; pour  $Y' = 0$  on est normal.

Et pour Y' négatif, on entre dans la catégorie des grands maigres.

Il reste à fixer le seuil à partir duquel on est un petit gros. Pour notre population, le seuil  $\sigma_{Y'}$  est discriminatoire.

Alors on a la partition :  $\{C, E\}$  : petits gros

$\{A, B, D, F, I, J, H\}$  : normaux

$\{G\}$  : grands maigres.

Pour une analyse plus fine, on peut tracer un réseau plus serré de lignes de niveau : on obtiendra alors des classes plus nombreuses.

### 3. Description par un parallélogramme

Le principe est rigoureusement le même que précédemment : la direction des lignes de niveau est alors celle de la droite de régression linéaire utilisée.

Le tracé des lignes de niveau peut être remplacé par le calcul de l'indice  $i = |Y - aX|$ , distance du point de coordonnées  $\begin{pmatrix} X \\ Y \end{pmatrix}$  à la droite de régression de Y en X, si c'est cette dernière qu'on utilise : dans notre exemple on aura  $i = |Y - 0,22 X|$  d'où la classification différente de la précédente :

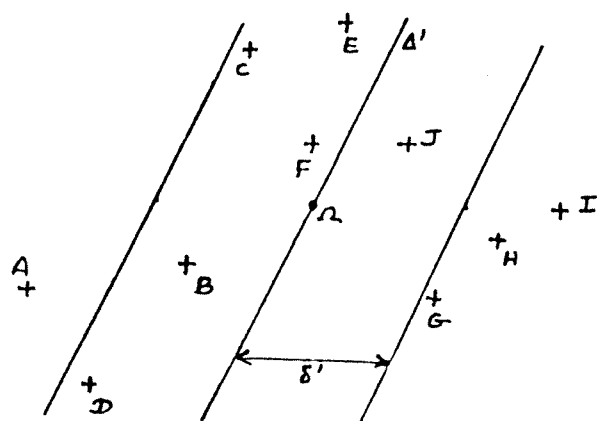
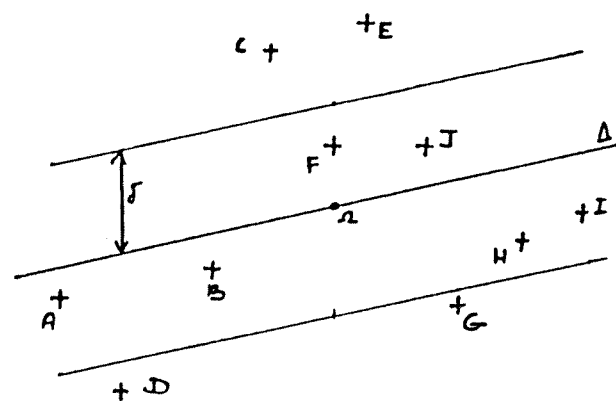
|       |      |      |      |      |   |      |      |      |      |      |
|-------|------|------|------|------|---|------|------|------|------|------|
| nom : | A    | B    | J    | I    | F | H    | G    | D    | C    | E    |
| i :   | 1,01 | 1,11 | 1,34 | 1,77 | 2 | 2,32 | 3,88 | 4,45 | 5,44 | 5,78 |

La partition de la population a changée :

$\{C, E\}$  ;  $\{F, J\}$  ;  $\{A, B, H, I\}$  ;  $\{D, G\}$

Et si les petits gros, au seuil d'un écart type  $\sigma$ , sont toujours  $\{C, E\}$ , D rejoint la catégorie des grands maigres.

A titre de curiosité, utilisons à présent l'autre droite de régression linéaire :



Nous voici face à une nouvelle partition :

$\{A\}$  ;  $\{B, C, D, E, F\}$  ;  $\{J\}$  ;  $\{G, H, I\}$

et à une nouvelle classification par les valeurs de  $i' = |X - a'Y|$  avec

$$a' = \frac{\sigma_{XY}}{\sigma_Y^2} \approx 0,48$$

|       |      |      |      |      |      |      |      |      |      |   |
|-------|------|------|------|------|------|------|------|------|------|---|
| nom : | F    | E    | J    | B    | D    | C    | G    | H    | A    | I |
| i' :  | 0,95 | 1,86 | 2,09 | 3,05 | 4,14 | 4,38 | 5,43 | 6,48 | 7,57 | 8 |

Quant aux petits gros, si on reste au seuil d'un écart type  $\sigma'$ , seul A en fait partie. Par contre, cette fois, le groupe des grands maigres est  $\{G, H, I\}$ .

Quelle description choisir alors ? Et quelle interprétation ?

La réponse du statisticien : toutes à la fois, mais en ne conservant que les conclusions communes.

Si on exclut la description par un cercle, qui ne semble pas convenir à notre problème, quelles sont les conclusions communes aux trois régressions ?

G est un "grand maigre" ; B, F et I sont "normaux".

Et c'est tout ! Avec un seuil un peu plus bas pour la dernière régression, C serait encore un petit gros : disons que C est presque un petit gros !

Non ! Les aventures du petit gros ne sont pas terminées !

Vous le retrouverez dans une prochaine fiche cuisine de tonton Lulu :

"Lutte des classes dans un nuage..."

A bientôt !