

Sur une population  $\Omega$  de  $n$  individus, on dispose de recensement de deux caractères qualitatifs  $X, Y$ :

par exemple le parti pour lequel ils ont voté pour une élection donnée et leur catégorie socio-professionnelle (C.S.P.).  
Le recensement  $\rightarrow$  tableaux de contingence  
i comme vote j - comme CSP

on pose  $F_{ij} = \frac{n_{ij}}{n}$  ( $n_{ij}$  - nombre d'individus ayant voté  $i$  et  $j$  comme CSP)

La probabilité qu'un individu  $w$  de  $\Omega$  répond  $(X=i, Y=j)$  est  $F_{ij} = P(X=i, Y=j)$   
(L'importance de chaque individu est la même  $\rightarrow$  la probabilité de chaque ind.  $w$  est  $P(w) = \frac{1}{n}$ )

$X \backslash Y$	1	j	l	
1	$F_{11}$	$F_{1j}$	$F_{1l}$	$F_{1.}$
i	$F_{i1}$	$F_{ij}$	$F_{il}$	$F_{i.}$
k	$F_{k1}$	$F_{kj}$	$F_{kl}$	$F_{k.}$
.	$F_{.1}$	$F_{.j}$	$F_{.l}$	1

La fréquence de la réponse ( $X=i$ ) ou sa probabilité, est noté  $F_i$  ou  $P(X=i)$

$$P(X=i) = F_i = \frac{n_{i1} + \dots + n_{il}}{n} = \frac{n_i}{n}$$

$$= F_{i1} + \dots + F_{il} = P(X=i, Y=1) + \dots + P(X=i, Y=l)$$

Probabilité conditionnelle à ( $X=i$ )

$$P(\omega/X=i) = \frac{1}{n_i} \quad \text{pour } X(\omega)=i$$

$$= 0, \quad \text{pour } X(\omega) \neq i$$

La fréquence relative de la CSP  $j$  parmi les électeurs du parti  $i$ , est

$$\frac{F_{ij}}{F_i} = \frac{n_{ij}}{n_i}$$

On dit que la probabilité ( $Y_j$ ) conditionnelle à ( $X=i$ ) est:

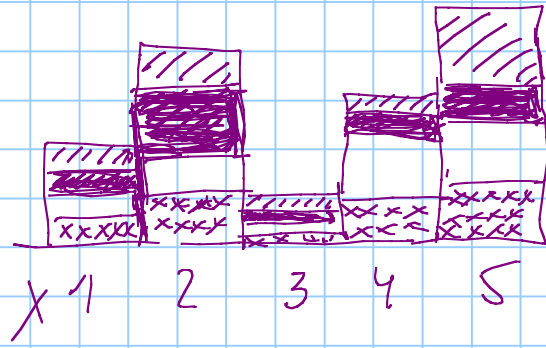
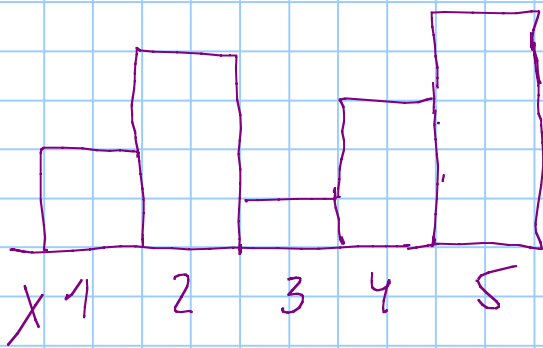
$$P(Y=j | X=i) = \frac{P(X=i, Y=j)}{P(X=i)} = \frac{n_{ij}}{n_i}$$
$$= \sum_{(\omega, Y(\omega)=j)} P(\omega/X=i)$$

La loi de  $Y$  conditionnelle à ( $X=i$ ) est l'ensemble des fréquences relatives



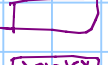

$$\left\{ \frac{F_{ij}}{F_{.i}} = P(Y=j | X=i); j=1, \dots, l \right\}$$

C'est un profil du caractère Y dans le parti i.

Des représentations telles que celle appelée histogramme



Legende:

	Y=1
	Y=2
	Y=3
	Y=4

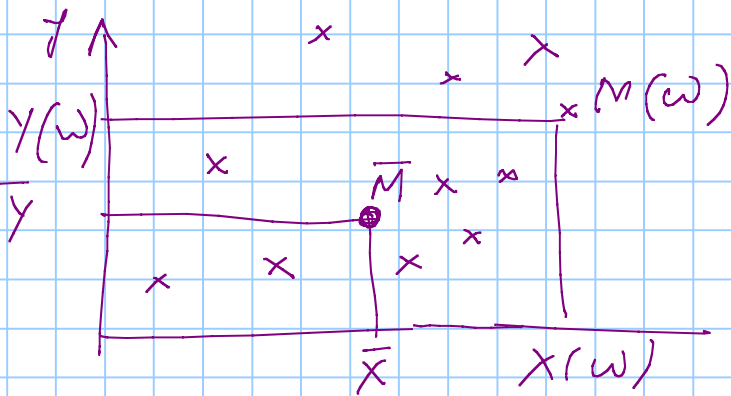
Echelle: en cm<sup>2</sup>  
représente une  
fréquence de 1/4

## 2. Recensement de caractère quantitatifs

La question du type quel est votre âge? ou quel est votre revenu annuel, ont un # fini des réponses, mais ici la valeur numérique des réponses a son importance.

L'histogramme de X est souvent remplacé par un histogramme en

Bâtons : On porte en abscisse des diverses valeurs de  $X$ , et on trace au dessus de  $x$  un bâton dont la longueur est proportionnelle à  $P(X=x)$ . On représente aussi un nuage de  $n$  points (dont certains peuvent être confondus) dans un repère cartésien du plan  $(Ox, Oy)$ . A chaque individu  $\omega$  corresp. un pt.  $M(\omega)$  d'abscisse  $X(\omega)$  et d'ordonnée  $Y(\omega)$ .



Des questions: quel est l'age moyenne de population?  $\bar{X} = \frac{1}{n} \sum_{\omega \in \Omega} X(\omega)$

le revenu moyenne  $\bar{Y} = \frac{1}{n} \sum_{\omega \in \Omega} Y(\omega)$

La pyramide des âges est-elle très large? Les inég. économiques sont-elles grandes?

→ Etude de dispersion des valeurs de  $X$  et de  $Y$  autour de moyennes  $\bar{X}$  et  $\bar{Y}$ .

On mesure la dispersion par

## les variances

$$\sigma^2(X) = \frac{1}{n} \sum_{\omega \in \Omega} (X(\omega) - \bar{X})^2, \quad \sigma^2(Y) = \frac{1}{n} \sum_{\omega \in \Omega} (Y(\omega) - \bar{Y})^2$$

$\sigma(X) = \sqrt{\sigma^2(X)}$  est l'écart-type de  $X$

L'influence mutuelle?

$\bar{M} = (\bar{X}, \bar{Y})$  - centre de masse  
- barycentre des pts

du nuages affectés de poids égaux.

Si les caractères  $X$  et  $Y$  ont tendance à évoluer dans le même sens,

$(X(\omega) - \bar{X})$  et  $(Y(\omega) - \bar{Y})$  ont tendance à avoir le même signe. Pour analyser:

La covariance  $\Gamma(X, Y) = \frac{1}{n} \sum (X(\omega) - \bar{X})(Y(\omega) - \bar{Y})$

positive  $\rightarrow$  évolution dans <sup>le</sup> même sens

négative  $\rightarrow$  év. dans un sens contraire

le changement d'unité de mesure pour l'une ou l'autre des variables ne devrait pas changer la manière dont on mesure leur lien: la corrélation de  $X$  et  $Y$ :

$$\rho(X, Y) = \frac{\Gamma(X, Y)}{\sigma(X) \sigma(Y)}$$

Droite de régression de  $Y$  par rapport à  $X$

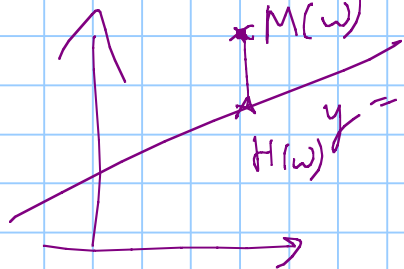
Exemple:  $Y$  - revenu annuel  
 $X$  - revenu annuel qu'avaient

les parents au même âge.

On pense a priori que la covariance est positive. Dans quel sens on se trompe, en supposant qu'il y a une dépendance linéaire  $Y = aX + b$  ?

L'erreur :

$$\delta(a, b) = \frac{1}{n} \sum_{\omega \in \Omega} (Y(\omega) - aX(\omega) - b)^2$$



"distance de  $Y$  au caractère  $aX + b$ ."

On cherche à déterminer la droite qui minimise  $\delta(a, b)$  : droite de régression. Le gradient  $\delta(a, b)$  s'annule :

$$\frac{\partial}{\partial a} \delta(a, b) = -\frac{2}{n} \sum_{\omega \in \Omega} X(\omega)(Y(\omega) - aX(\omega) - b) = 0$$

$$\frac{\partial}{\partial b} \delta(a, b) = -\frac{2}{n} \sum_{\omega \in \Omega} (Y(\omega) - aX(\omega) - b) = 0$$

$$\Leftrightarrow l = \bar{Y} - a\bar{X}$$

$$l \sum_{\omega \in \Omega} X(\omega)(Y(\omega) - \bar{Y}) = a \sum_{\omega \in \Omega} X(\omega)(X(\omega) - \bar{X})$$

$$\text{or } \bar{X} \sum_{\omega \in \Omega} (Y(\omega) - \bar{Y}) = \bar{X} \sum_{\omega \in \Omega} (X(\omega) - \bar{X}) = 0$$

$$\text{si } X \neq \text{const}, \quad a = \frac{\Gamma(X, Y)}{\sigma^2(X)}, \quad l = \bar{Y} - a\bar{X}$$

$$y - \bar{Y} = \frac{\Gamma(X, Y)}{\sigma^2(X)} (x - \bar{X}) \text{ est la droite}$$

de regression de  $Y$  p.r. à  $X$  (si  $X \neq \text{const}$ )

La distance de  $Y$  au  $\bar{Y} + \frac{\Gamma(X, Y)}{\sigma^2(X)} (X - \bar{X})$   
est:

$$\frac{1}{n} \sum_{\omega \in \Omega} \left[ Y(\omega) - \bar{Y} - \frac{\Gamma(X, Y)}{\sigma^2(X)} (X(\omega) - \bar{X}) \right]^2$$

$$= \frac{1}{n} \sum_{\omega \in \Omega} (Y(\omega) - \bar{Y})^2 -$$

$$- 2 \frac{\Gamma(X, Y)}{\sigma^2(X)} \frac{1}{n} \sum_{\omega \in \Omega} (Y(\omega) - \bar{Y}) (X(\omega) - \bar{X})$$

$$+ \frac{\Gamma^2(X, Y)}{\sigma^4(X)} \frac{1}{n} \sum_{\omega \in \Omega} (X(\omega) - \bar{X})^2$$

$$= \sigma^2(Y) - \frac{\Gamma^2(X, Y)}{\sigma^2(X)} = \sigma^2(X) [1 - \rho^2(X, Y)]$$

Cette distance est d'autant plus grande que la variance de  $Y$  est grande et que la corrélation de  $X, Y$  est petite.

Si le module de la corrélation est 1,  $Y$  est une fn. affine de  $X$ ,

Dans tous les cas:  $|\rho(X, Y)| \leq 1$

$$\left( \text{car } |\Gamma(X, Y)| \leq \sigma(X) \sigma(Y) \right)$$

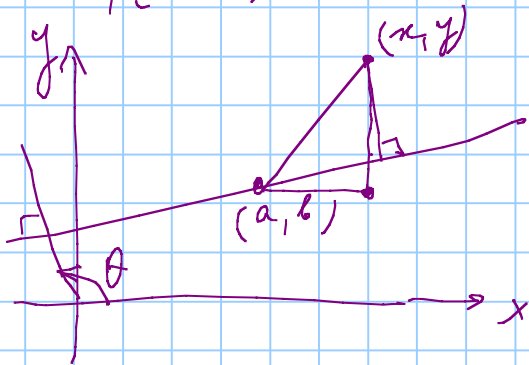
## Axes principaux d'un nuage

On mesure la dispersion du nuage autour d'une droite  $\Delta$  par la somme des carrés des distances des pts du nuage à la droite.

Soit  $\Delta_{a,b,\theta}$  une droite passant par  $(a,b)$  orthogonal au vecteur  $(\cos\theta, \sin\theta)$

La distance de  $(x,y)$  à  $\Delta_{a,b,\theta}$  est

$$|(x-a)\cos\theta + (y-b)\sin\theta|$$



L'angle  $\theta$  étant fixé

$\varphi$  est min pour

$$a = \bar{X}, \quad b = \bar{Y}$$

$$\varphi(a,b,\theta) = \frac{1}{n} \sum_{w \in \Omega} \left[ (x(w) - a)\cos\theta + (y(w) - b)\sin\theta \right]^2$$

$$\varphi'(\theta) = \frac{2}{n} \sum_{w \in \Omega} \left[ (x(w) - \bar{X})\cos\theta + (y(w) - \bar{Y})\sin\theta \right]$$

$$\cdot \left[ -(x(w) - \bar{X})\sin\theta + (y(w) - \bar{Y})\cos\theta \right]$$

$$= 2 \cos 2\theta \Gamma(x,y) + \sin 2\theta [\sigma^2(y) - \sigma^2(x)]$$

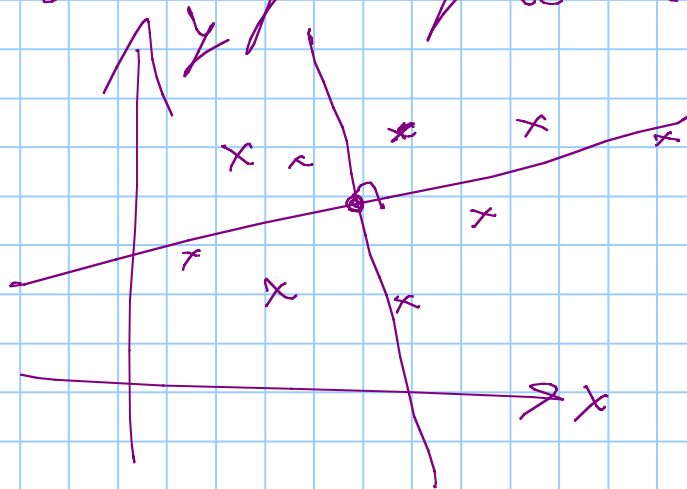
$\varphi$  est const si  $X$  et  $Y$  ne sont pas corrélées et ont même variance

le nuage a un aspect circulaire

Si non  $\varphi$  s'annule pour 2 droites

orthogonales: max de  $\varphi$  et min de  $\varphi$ .

La dist. min = la droite de regression orthogonale du couple  $(x, y) \rightsquigarrow$  le nuage s'allonge le plus l'axe principale de nuage.



1.2.4.  $p$  caractères quantitatifs:

Analyse en composantes principales (ACP)

Les images de proj sur les axes de regression sont "le plus represent"

Grouper les caractères

$$v(x) = \sum_{i=1}^p v_i x_i$$

$$\bar{v}(x) = \frac{1}{n} \sum_{h \in \Omega} v(x(h)) = \sum_{i=1}^p v_i \bar{x}_i = v(\bar{X})$$

$\bar{X} = (\bar{x}_1, \dots, \bar{x}_p)$  - le barycentre de nuage

Soit  $v$  et  $w \in \mathbb{R}^p$  la covariance  $v(x)$  et  $w(x)$  est

$$T(v(x), w(x)) = \frac{1}{n} \sum (v(x(h)) - v(\bar{X})) \cdot$$

$$\begin{aligned}
 (w(x(\omega)) - w(\bar{x})) &= \\
 &= \sum_{i=1}^p \sum_{j=1}^p v_i w_j \left( \frac{1}{n} \sum_{\omega \in \Omega} (x_i(\omega) - \bar{x}) (y_j(\omega) - \bar{y}) \right) \\
 &= [v_1, \dots, v_p] \Gamma \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}
 \end{aligned}$$

$\Gamma$  la matrice de covariance,

$\Gamma = \Gamma(x_i, x_j)_{1 \leq i, j \leq p}$   
 c'est une forme bilinéaire  
 symétrique

forme quadratique associée:

$$v \mapsto \Gamma(v(x), v(x)) = v^2(v(x))$$

variance de  $v(x)$  - positive, nulle si  
 $v(x)$  est const.

$v(x)$  explique d'autant mieux  
 la dispersion des individus que  
 sa variance est grande.

Mais  $v$  intervient comme  
 pondération de  $\lambda$  pour utiliser  
 un critère invariant pas multipl.  
 de  $v$  par une const on cherche  
 à maximiser

$$\frac{v^2(v(x))}{\sum_{i=1}^p v_i^2} = \frac{[v_1, \dots, v_p] \Gamma \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix}}{\sum_{i=1}^p v_i^2}$$

L'espace  $\mathbb{R}^p$  étant muni du produit

scalaire  $\langle v, w \rangle = \sum_{i=1}^p v_i w_i$

il existe une base orthonormée de vect. propres de  $\Gamma$  associées aux valeurs propres positives ordonnées en ordre décroissant  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

On obtient:

Prop. 1.2.1.: Soit  $\lambda_1 \geq \dots \geq \lambda_p$  les valeurs propres de  $\Gamma$  et  $u_1, \dots, u_p$  une base orthonormée de vect. propres de  $\Gamma$  assoc.

Alors: a)  $\sigma^2(u_i(x)) = \lambda_i$   
 b)  $\Gamma(v(x), w(x)) = \sum_{i=1}^p \lambda_i \langle v, u_i \rangle \langle w, u_i \rangle$

donc les caractères  $u_1(x), \dots, u_p(x)$  sont deux à deux non corrélés;

c)  $\sigma^2(u_i(x)) = \lambda_i = \sup_{v \neq 0} \frac{\sigma^2(v(x))}{\sum_{i=1}^p v_i^2}$

$\sigma^2(u_i(x)) = \lambda_i = \sup \left\{ \frac{\sigma^2(v(x))}{\sum_{i=1}^p v_i^2}; v \neq 0 \right\}$

non corrélés avec  $u_1(x), \dots, u_{i-1}(x)$

On dit que  $u_i(x)$  est un  $i^{\text{ème}}$  facteur principal (qui est déterminé, au signe près, de manière unique si  $\lambda_i$

est une valeur propre simple.

Si  $v$  un vect de  $(\mathbb{R}^p)$ . Si on translate le nuage en lui ajoutant  $v$  (par exemple si on le centre en ajoutant  $-\bar{X}$ )

On ne change pas les covariances  
 $\Gamma(v(x+v), w(x+v)) = \Gamma(v(x), w(x))$   
Les facteurs principaux sont les mêmes.

Dispersion du nuage des  $n$  jets  
 $(X(\omega))_{\omega \in \Omega}$  autour de son baryc.  $\bar{X}$ ,  
= inertie du nuage

$$J = \frac{1}{n} \sum_{\omega \in \Omega} \|X(\omega) - \bar{X}\|^2$$

$(u_1, \dots, u_p)$  étant une base orthonormée,

$$J = \frac{1}{n} \sum_{\omega \in \Omega} \sum_{i=1}^p \langle X(\omega) - \bar{X}, u_i \rangle^2$$

$$= \sum_{i=1}^p \sigma^2(u_i(X)) = \sum_{i=1}^p \lambda_i$$

c'est la trace de  $\Gamma$ ,