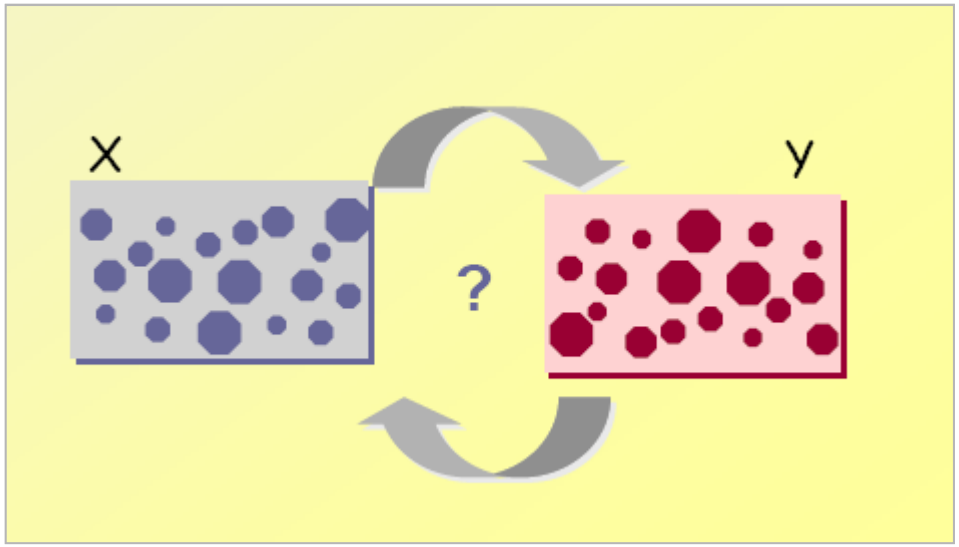
Lecture n° 5.

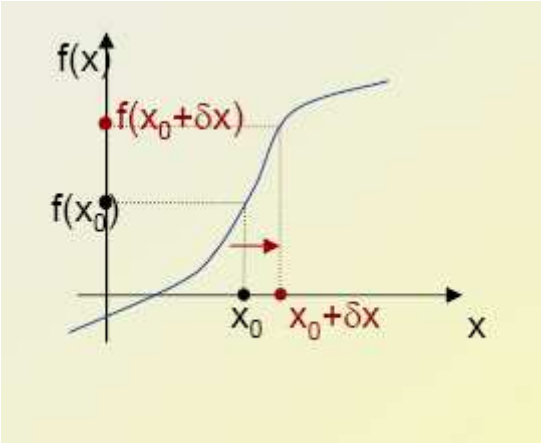
Correlation.



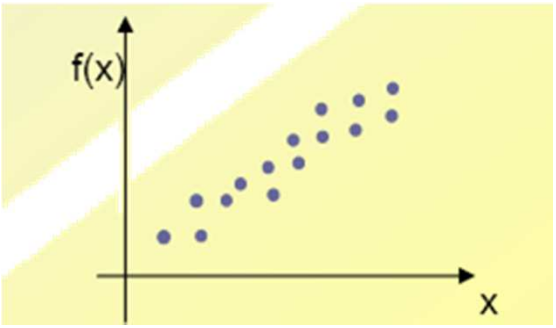
Liens entre les variables..

D'habitude, les individus constituant une population possèdent plusieurs caractères.
Ces caractères ne sont pas indépendant,
mais liés entre eux, souvent de manière complexe.

Lien fonctionnel:



Lien stochastique (probabiliste):



changement d'une variable indépendante provoque
un changement de distribution de probabilité
de la variable dépendante

lien statistique (corrélation):
changement de la variable **x** provoque
une variation bien déterminée de **y**

Lien stochastique ou lien statistique?

Théorème: Si entre les variables le lien stochastique n'existe pas,
alors il n'y a pas de corrélation entre ces variables non plus.

Le théorème inverse n'est pas vérifié:

EXEMPLE: pour un nombre donne des valeurs d'une variable
leur moyenne est strictement déterminable,
mais la même moyenne peut être obtenue
à partir d'une autre combinaison des données.

Exemple: Prenons la variable « poids ».
La moyenne de valeurs « 62 » et « 68 » est égale à 65.
mais on peut aussi obtenir la moyenne 65 à partir d'autres valeurs:
60 et 70
63 et 67
61 et 69
……

Régression ou corrélation?

Corrélation: permet d'estimer le degré (la force) de relation (de lien) entre les variables. Elle répond à la question:

> à quel point une équation choisie (une loi analytique) décrit correctement une relation entre les variables (observables).

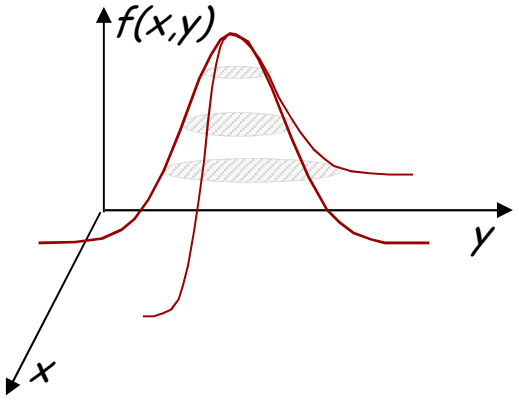
Régression (ou estimation): permet d'estimer une variable (dépendante) à partir d'une (ou plusieurs) variables indépendantes.

> Elle donne des outils statistiques pour, en outres, ajustement des courbes à des loi théoriques.

Distribution bidimensionnelle de probabilité.

- Quand le résultat d'une expérience aléatoire (mesure) est une paire de valeurs (x_i,y_i) alors les valeurs (x,y) constituent une variable aléatoire bidimensionnelle

$$P(a_1 < x < b_1, a_2 < y < b_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x,y)dxdy$$



- Distribution limite d'une variable si l'autre peut prendre une valeur quelconque

$$P(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$P(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

- Si les variables x et y sont indépendantes, alors

$$f(x,y) = f(x)\cdot f(y)$$

Moment d'une distribution.

- Moment de l'ordre **s** de la distribution unidimensionnelle de variable x

$$m_s = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^s$$

⟹ moment de l'ordre 2 = variance de x

- Moment central **r, s** de la distribution de la variable bidimensionnelle (x,y):

$$m_{r,s} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r (y_i - \bar{y})^s$$

  - moment de l'ordre 2,0 = variance de x

$$m_{2,0} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - moment de l'ordre 0,2 = variance de variable y

$$m_{0,2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  - moment de l'ordre 1,1 = **covariance** de variable (x,y)

$$m_{1,1} = \mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance.

- De la manière simplifiée, covariance peut être calculée de la formule:

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\overline{xy}-\bar{x}\cdot\bar{y}$$

$$\begin{aligned}\operatorname{cov}(x,y)=S_{xy}&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\\&=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\cdot\bar{y})=\\&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i-\bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i+\bar{x}\cdot\bar{y}=\\&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\cdot\bar{y}-\bar{x}\cdot\bar{y}+\bar{x}\cdot\bar{y}=\overline{xy}-\bar{x}\cdot\bar{y}\end{aligned}$$

- Covariance de variables indépendantes est nulle:

$$\operatorname{cov}(x,y)_{INDEP}=0$$

$$\begin{aligned}\operatorname{cov}(x,y)=S_{xy}&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\bar{x})(y_i-\bar{y})f(x,y)dxdy=\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\bar{x})(y_i-\bar{y})f(x)f(y)dxdy=\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\bar{x})(y_i-\bar{y})f(x)f(y)dxdy=\\&=\int_{-\infty}^{\infty}(x_i-\bar{x})f(x)dx\int_{-\infty}^{\infty}(y_i-\bar{y})f(y)dy=0\end{aligned}$$

Coefficient de corrélation.

- La valeur de la covariance peut indiquer la présence (ou absence) des corrélations entre les variables:
  - relation directe entre les variables $\Rightarrow \operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$
  - relation réciproque entre les variables $\Rightarrow \operatorname{cov}(x,y) < 0$
  - il n'y a pas de corrélation entre les variables $\Rightarrow \operatorname{cov}(x,y) = 0$

- La covariance n'est pas le meilleur indicateur de corrélation entre les variables:
  - elle varie (a priori) entre $^{-}\infty$ et $+\infty$;
  - elle est exprimée en unités à la fois de x et de y;

- Le coefficient de corrélation:

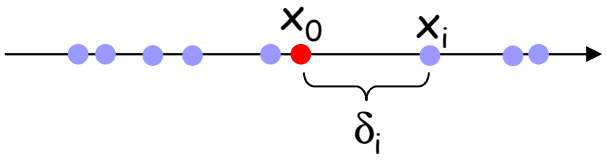
$$r = \frac{Cov(X,Y)}{S(X)\cdot S(Y)}$$

  - il varie entre -1 et 1;
  - il n'a pas de dimensions.

Méthode des moindres carrés.

- A cause d'une multitude de facteurs d'influence, la vraie valeur de la variable étudiée, x_0, est rarement observée durant mesurages.

- Il est alors raisonnable de demander
  quelle est la probabilité d'obtenir une valeur x_i
  qui diffère de δ_i de la valeur x_0:



$$P(\delta_i) = f(x_i - x_0)dx$$

probabilité d'obtenir une série des résultats $x_1, x_2, \ldots x_n$

$$P(x_1 \cap x_2 \cap .... \cap x_n) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} P(\delta_i) = \prod_{i=1}^{n} f(x_i - x_0)dx$$

- on souhaite avoir l'évaluation x_i la plus précise de x_0,
  alors on recherche x_0 tel, que

$$\prod_{i=1}^{n} P(x_i) = \text{maximum} \Rightarrow$$

si la variable x suit la loi normale, alors

$$\prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - x_0)^2}{\sigma}\right) = \text{maximum quand}$$

$$\sum_{i=1}^{n} (x_i - x_0)^2 = \text{minimum}$$

Méthode des moindres carrés.

$$\sum_{i=1}^{n}(x_i - x_0)^2 = \text{minimum}$$

cette condition sera vérifiée si

$$\frac{\partial}{\partial x}\sum_{i=1}^{n}(x_i - x_0)^2 = 0$$

$$2\sum_{i=1}^{n}(x_i - x_0) = 0$$

$$\sum_{i=1}^{n}(x_i - x_0) = 0$$

$$x_0 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

l'évaluation la plus précise de x_0 est la valeur moyenne des x_i

La méthode de moindres carrés a des multiples applications pratiques.
En particulier, elle permet d'analyser des corrélations entre les variables.

Méthode des moindres carrés et la corrélation.

- Nous allons considérer les variables comme corrélées, si changement d'origine et d'échelle transforme la distribution d'une variable dans l'autre:

$$(x_i - \bar{x})\lambda^* + \bar{y} = y_i$$

  - si cette relation est vérifiée pour toutes les paires (x_i,y_i), alors nous parlons d'une corrélation linéaire idéale.
  - si la valeur λ^* varie de la manière fonctionnelle $\lambda^* = f(x)$, alors nous parlons d'une corrélation non-linéaire.

- Les corrélations idéales ne sont jamais observées, à cause des incertitudes de mesure:

$$S_x^2 = \frac{1}{n-1}(x_i - \bar{x})^2 \qquad S_y^2 = \frac{1}{n-1}(y_i - \bar{y})^2$$

changeons les variables x et y en variables centrées réduites:

$$\xi_i = \frac{x_i - \bar{x}}{S_x} = \frac{\delta x_i}{S_x}$$

$$\eta_i = \frac{y_i - \bar{y}}{S_y} = \frac{\delta y_i}{S_y}$$

$$\xi_i S_x \lambda^* = \eta_i S_y$$

Méthode des moindres carrés et la corrélation.

$$\xi_i S_x \lambda^* = \eta_i S_y$$

$$\xi_i S_x \lambda^* - \eta_i S_y = 0 \Longleftarrow \text{si corrélation est idéale}$$

$$\xi_i S_x \lambda^* - \eta_i S_y = \varepsilon \Longleftarrow \text{si corrélation n'est pas parfaite}$$

- D'habitude on considère que une seule variable dans la paire (p.ex. y ou la variable centrée réduite η) est entachée de l'incertitude.

  Alors, on cherche à minimiser cette incertitude (minimiser la variance de la variable):

$$S_\eta^2 = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((\eta_i - \lambda^* \frac{S_x}{S_y}\xi_i)\right)^2$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{n}(\eta_i)^2 - 2\lambda^*\sum_{i=1}^{n}\eta_i\xi_i + (\lambda^*)^2\sum_{i=1}^{n}(\xi_i)^2\right\}$$

$$= 1 + (\lambda^*)^2 - \frac{1}{n}2\lambda^*\sum_{i=1}^{n}\eta_i\xi_i$$

Méthode des moindres carrés et la corrélation.

$$= 1 + (\lambda^*)^2 - \frac{1}{n} 2\lambda^* \sum_{i=1}^{n} \eta_i \xi_i$$

posons $\rho = \frac{1}{n} \sum_{i=1}^{n} \eta_i \xi_i$

$$= 1 + (\lambda^*)^2 - 2\lambda^* \rho = 1 - \rho^2 + (\rho - \lambda)^2$$

Cette expression est minimale si $\rho = \lambda$.

Dans le système des variables centrées réduites, l'équation

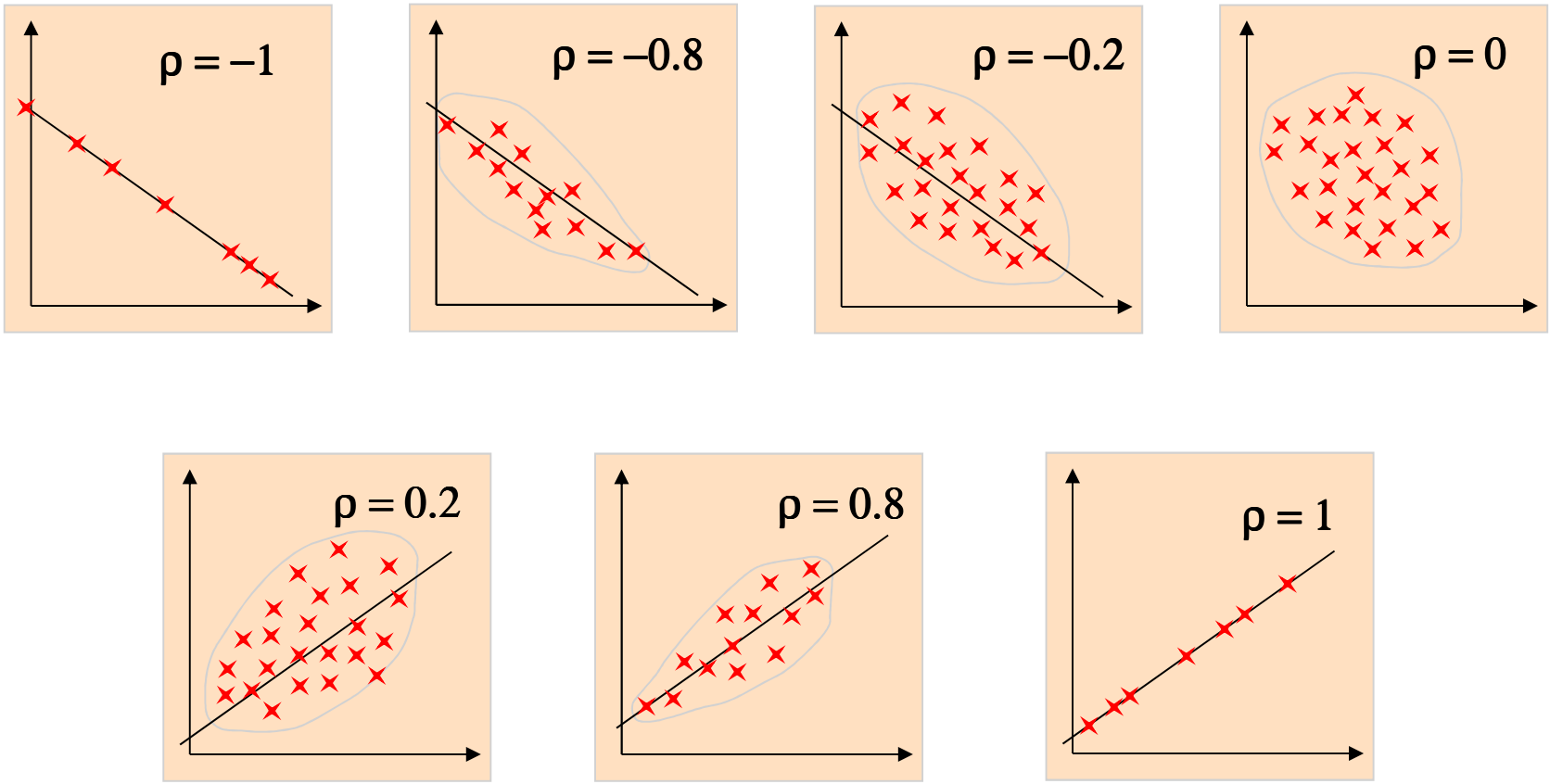
$$(x_i - \bar{x})\lambda^* + \bar{y} = y_i$$

devient $\eta_i - \rho \xi_i = 0$ ⟸ droite de régression

coefficient de corrélation

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \eta_i \xi_i = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \frac{\mathrm{cov}(x, y)}{S_x S_y}$$

Représentation graphique de corrélation.



Estimation ponctuelle du coeff. de corrélation.

- Si dans une population P on étudie simultanément deux grandeurs physiques X,Y telles, que:
  - ou bien X est une variable contrôlée et Y est une variable dépendante,
  - ou bien le couple {X,Y} suit une distribution normale bidimensionnelle,

  alors la meilleurs estimation ponctuelle du coefficient de corrélation dans la population XY est sa valeur calculée à l'issue de l'expérience:

$$\rho = r = \frac{\mathrm{cov}(X,Y)}{S_X S_Y}$$

S_X – écart type de l'échantillon $\{x_1, x_2, \ldots, x_n\}$
S_Y – écart type de l'échantillon $\{y_1, y_2, \ldots, y_n\}$

- L'exactitude de l'estimation dépend, en réalité, du nombre d'informations disponibles, et alors, d'une manière plus précise,

$$\rho = r\left(1 + \frac{1-r}{2(n-3)}\right)$$

Estimation du coefficient de corrélation par intervalle de confiance.

- De la population P on retire n fois des échantillons $\{x_i, y_i\}$ de taille n ($n \geq 20$).
- Pour chaque échantillon, on détermine la valeur de r et z.

$$z = \frac{1}{2}\ln\left(1+\frac{1+r}{1-r}\right)$$

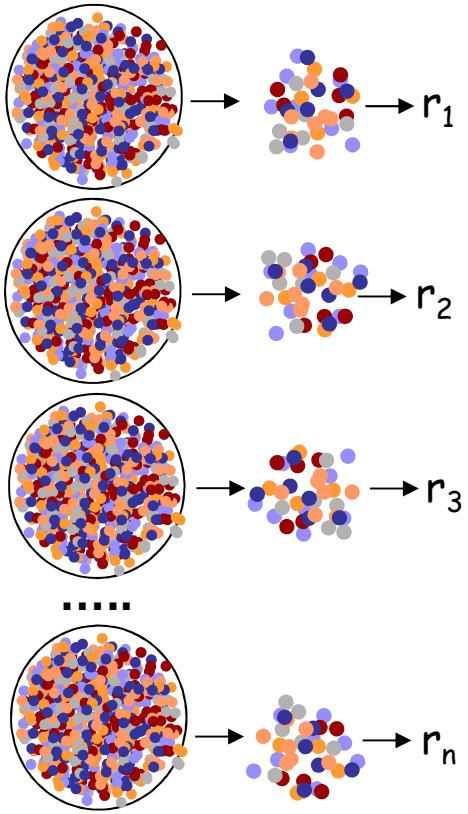
- On démontre que, si le nombre de mesures n >20, la variable aléatoire z suit la loi normale N(μ, σ):

$$\mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \qquad \sigma = \frac{1}{\sqrt{n-3}}$$

Au niveau de confiance α, l'intervalle de confiance de ρ est donné par la formule:

$$\rho \in \left(z - \frac{z_\alpha}{\sqrt{n-3}}, z + \frac{z_\alpha}{\sqrt{n-3}}\right)$$

Comparaison des coefficients de corrélation.



- Dans k populations P_i, $P_2 \ldots P_n$ les variables X,Y sont liées avec des coefficients de corrélation ρ_1, $\rho_2 \ldots \rho_n$.
- De ces populations on extrait k échantillons E_i de taille n_i et on détermine des coefficients de corrélations entre les grandeurs x et y: r_i.

PROBLEME: y a-t-il une corrélation entre les X et Y dans les échantillons ? Est-elle comparable?

Hypothèse H_0: Dans toutes les populations les variables X et Y sont liées avec le même coefficient de corrélation: $r_1 = r_2 = \ldots = r_n = \rho$

Comparaison des coefficients de corrélation.

EXECUTION du TEST: pour tous les échantillons, on calcule z_i et la moyenne pondérée de z_i:

$$z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right) \Rightarrow \bar{z} = \frac{\sum_i (n_i - 3) z_i}{\sum_i (n_i - 3)}$$

Sous l'hypothèse H_0, si les échantillons E_i sont grands ($n_i > 30$), la variable Y^2

$$Y^2 = \sum_i (n_i - 3)(z_i - \bar{z})^2 =$$

$$= \sum_i (n_i - 3) z_i{}^2 - \bar{z}^2 \sum_i (n_i - 3)$$

suit la loi de χ^2 à ν=(k-1) degrés de liberté.