guide_pharmacie@yahoo.fr

Université de Batna

Faculté de Médecine

Département de Pharmacie

2005/2006

Cours de Statistique

1ère Année Pharmacie

Chapitre II : Les séries statistiques doubles

D'après le cahier de :

I. Hadef

Created by: T. Djebaili

Chapitre II : Les séries statistiques doubles :

Définition :

Une série statistique double ou série à 2 variables est un ensemble de résultats issus de l'observation de 2 caractères numériques sur une même population

exp :

1/ La taille et le poids d'un groupe d'enfants.
2/ le salaire est la qualification d'un ensemble de salaire.
3/ la température est la pression d'un milieu à différentes heure.

* La série statistique des couples (x_i, y_i) sera notée (x, y)

Notation et représentation des séries statistiques doubles

Une série statistique double peut être donnée comme l'énumération d'un certain nombre de résultats, on distingue deux cas :

1/ les effectifs des valeurs égaux à 1.
2/ Les effectifs des couples (x_i, y_j) égaux n_{ij}

1/ L'effectif des valeurs (x_i, y_j) égaux à 1 :

Dans ce cas une série statistique se présente comme suit :

X	x_1	x_2	. . . . .	x_i
Y	Y_1	x_2	. . . . .	Y_i

La moyenne, la variance et l'écart type

Les variables statistique X et Y ont chacune une moyenne, variance et l'écart type

* $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$, $V(X) = \sum \frac{x_i^2}{n} - \bar{X}^2$

$6 = \sqrt{V}$

* $\bar{Y} = \sum_{i=1}^{m} \frac{Y_i}{n}$, $V(Y) = \sum_{i=1}^{m} \frac{Y_i^2}{m} - \bar{Y}^2$

$6(Y) = \sqrt{V}$

2/ Les effectifs des valeurs (des couples) (x_i, y_j) egaux n_{ij} :

Soit X et Y deux séries statistiques definie sur la même population, * prenant les valeurs $x_1, x_2 \dots x$

*Y " " " $y_1, y_2, \dots y_n$:

Definition : On appelle effectifs partiels n_{ij} du couples (X_i, Y_j) le nombre d' individus de la population pour les quelles la serie X prend la valeur X_i et la serie Y prend la valeur Y_j.

Tableau d'effectif :

X \ Y	Y_1	Y_2	- - -	Y_j		Y_s	total $n_{i.}$
X_1	n_{11}	n_{12}	- . . .	n_{1j}	-	n_{1s}	$n_{1.}$
⋮							
X_i	n_{i1}	n_{i2}	- . -	n_{ij}		n_{is}	$n_{i.}$
⋮							
X_r	n_{r1}	n_{r2}		n_{rj}	.	n_{rs}	$n_{r.}$
totale $n_{.j}$	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	n .

Dans n_{ij} le 1er indice i correspond à la ligne
- Le 2ème indice j correspond à la colonne.

La définition : Si n_{ij} est l'effectif partiel d'un couple (x_i, y_j) et n l'effectif total. le rapport $f_{ij} = \frac{n_{ij}}{n}$ s'appelle la fréquence partielle du couple (X_i, Y_j).

Remarque : $\sum_{j=1}^{p} \left(\sum_{i=1}^{r} n_{ij} \right) = n$.

la somme des effectifs partiels correspondant à la valeur X_i est égale à l'effectif des individus pour lesquels X prend la valeur X_i et on note $n_{i\bullet}$

$$n_{i\bullet} = \sum_{j=1}^{j=s} n_{ij} = n_{i1} + n_{i2} + \dots n_{is}$$

De la même manière on définie $n_{\bullet j}$

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij} = n_{1j} + n_{2j} + n_{3j} + \dots n_{rj}$$

Caractéristiques d'une série à deux variables quantitatives :

Soit (X, Y) une série statistique double quantitatives définies sur une même population

La moyenne de X : $\bar{X} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} X_i$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{r} n_{i\bullet} X_i$$

la moyenne de Y

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} y_j$$

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^{r} n_{\bullet j} y_j$$

La variance de X :

$$V(X) = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} (x_i - \overline{X})^2$$

$$= \frac{1}{n} \sum_{i=1}^{r} n_{i.} (x_i - \overline{X})^2$$

Après le développement de la forme :

$$V(X) = \frac{1}{n} \sum_{i=1}^{r} n_{i.} x_i^2 - \overline{X}^2$$

L'écart type de X : $\sigma_X = \sqrt{V(X)}$

La variance Y : $V(Y) = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} (y_j - \overline{Y})^2$

$$V(Y) = \frac{1}{n} \sum_{j=1}^{s} n_{.j} (y_j - \overline{Y})^2$$

$$V(Y) = \left(\frac{1}{n} \sum_{j=1}^{s} n_{.j} y_j^2\right) - (\overline{Y})^2$$

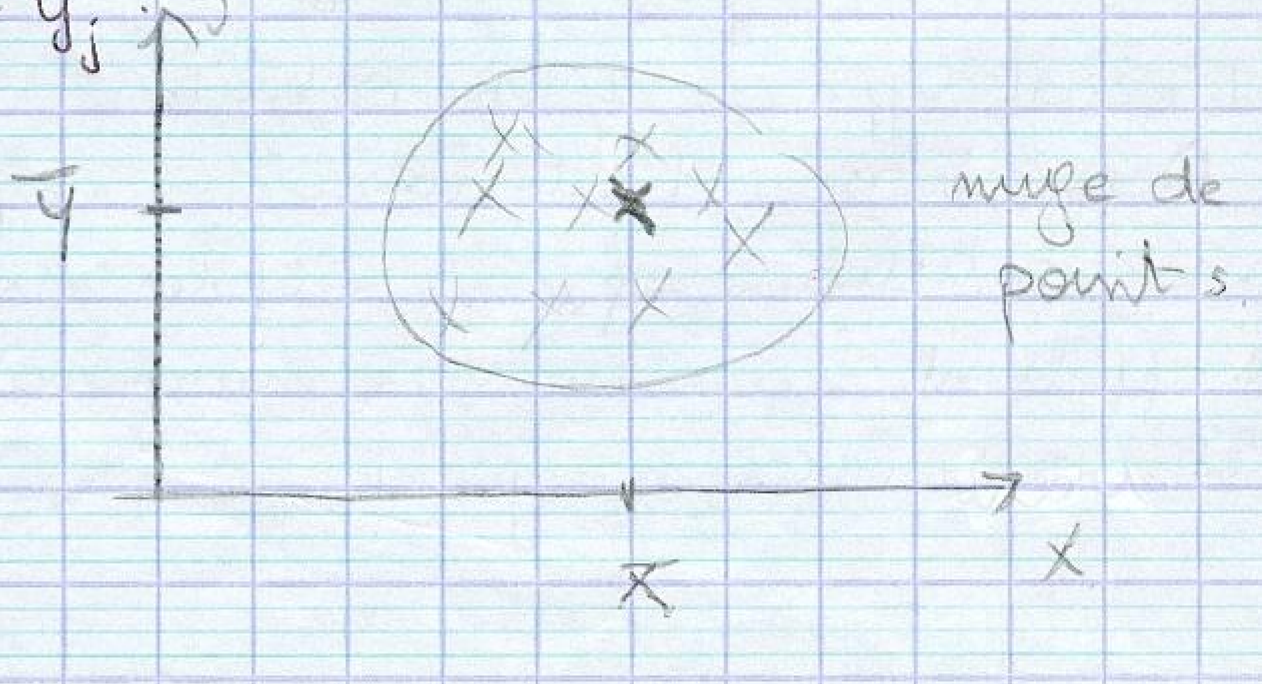
L'écart type de Y se calcule : $\sigma_Y = \sqrt{V(Y)}$.

Definition : étant donné un couple (X, Y) de série statistique quantitatif les moyennes $\overline{X}$, $\overline{Y}$ s'appellent les moyennes marginales, V(X) et V(Y) les variances marginales.

$x \backslash y$		$n_{i.}$	$x_i n_{i.}$
y_j $n_{.j}$		n	

Nuage de point.

Dans un repère cartésien on represente la série statistique à 2 variables par les points M_{ij} de coordonnées x_i et y_j. L'ensemble de ces point s'appelle le nuage de points de la serie double (x, y). C'est une representation importants des liens qu'il y a entre les x_i et y_j



Le points moyen : le point moyen de la série (X, Y) est le point $\overline{M}(\overline{X}, \overline{Y})$ dont les coordonnées sont les moyennes $\overline{X}$ et $\overline{Y}$ des x_i et y_j.

Definition de la covariance :

Si (X, Y) désigne un couple des séries statistiques quantitatif prenant respectivement.

$x_1, x_2, \ldots\ldots x_r$

$y_1, y_2, \ldots\ldots, y_s$.

n_{ij} l'effectif partiel du (x_i, y_j)

On appelle covariance du couple (X, Y) et on note

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}(x_i - \overline{X})(y_j - \overline{Y}).$$

$$= \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}\, x_i y_j \,) - (\overline{X} \cdot \overline{Y})$$

$$= \sigma_{XY}$$

Propriétés :

1/ $\operatorname{cov}(X, X) = V(X)$, $\operatorname{cov}(Y, Y) = V(Y)$

2/ $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.

proposition : si X et Y sont indépendantes $\Rightarrow$ $\operatorname{cov}(X, Y) = 0$.

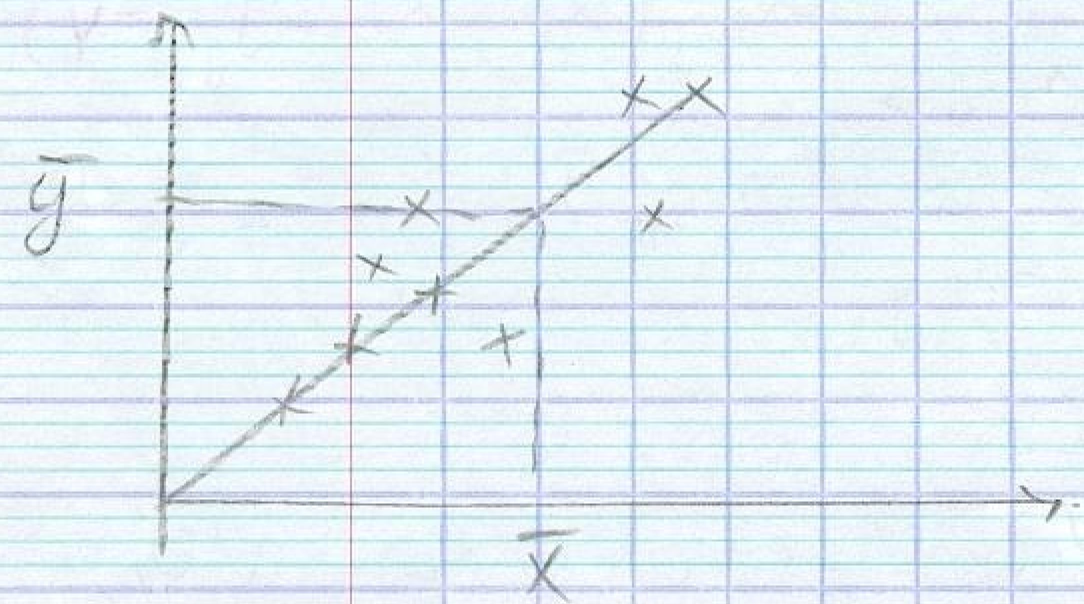
L'ajustement : (x, y) une série double.

$M_{ij} = (x_i, y_j)$ L'ensemble des points de la série

Définition : Ajuster un ensemble des points consiste à déterminer une courbe simple (aussi proche que possible de l'ensemble M_{ij}.

L'ajustement linéaire : lorsque le nuage de points d'une série double semble avoir une forme rectiligne on cherche à approcher par une fonction affine. (ajustement linéaire)

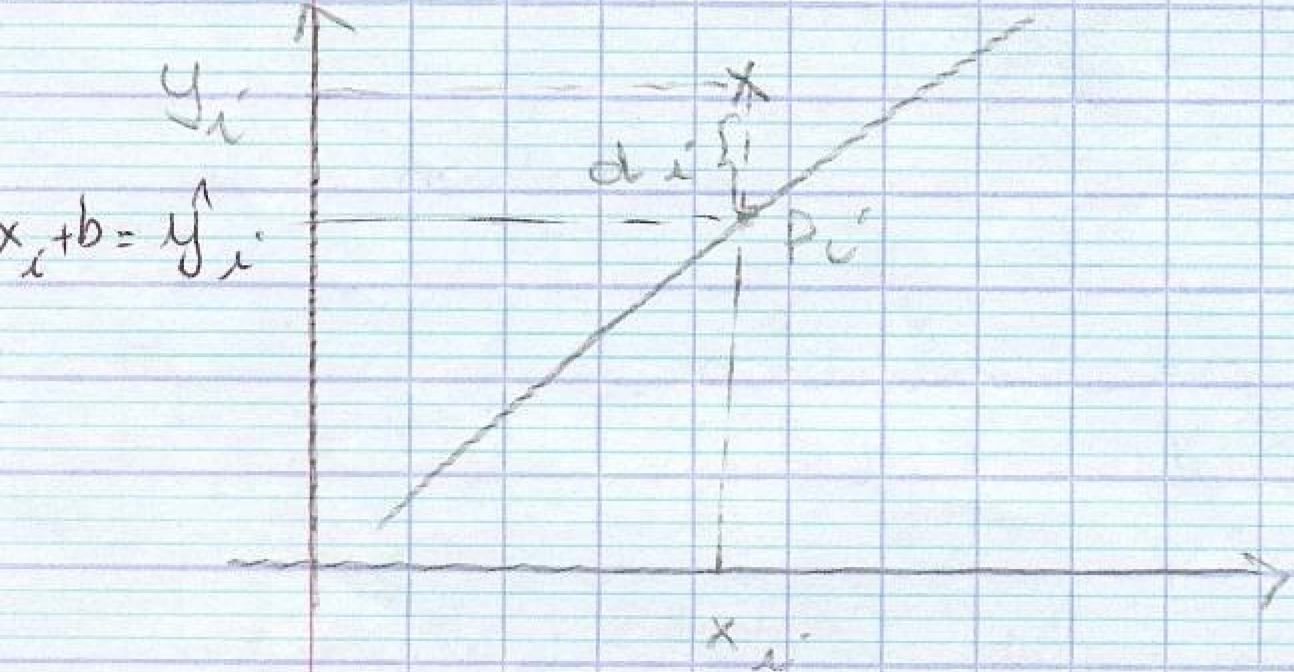
On dit que l'on effectue une ajustement linéaire de ce nuage de points et on note : $y = ax + b$ l'équation de la droite recherché.



Méthode de moindres carrés :

On cherche une équation $y = ax + b$ de la droite (Δ) telle que la somme de ses distances au differentes points representant soit minimal

- La distance choisie est le carré de la difference des ordonnées entre chaque points et le points de la droite ayant le même abscisse.



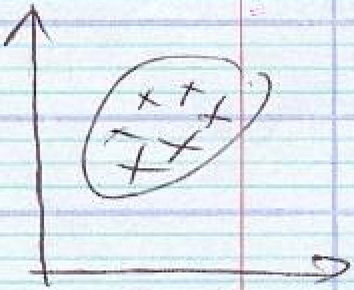
$M_i(x_i, y_i)$, $P_i(x_i, \hat{y}_i)$.

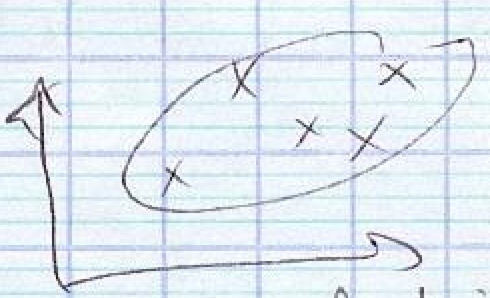
$\sum_i d_i^2 = \sum (y_i - \hat{y}_i)^2$ soit minimum.

Notion de correlation linéaire :

la méthode de Moindres carrés peut être utilisé pour n'importe quelle série double. il existe une droite d'estimation par la méthode moindre carrés pour s'assurer d'une façon objective et non visuelle que l'ajustement est valide ou non. On calcule le coefficient de corrélation linéaire.

Correlation. Soit (x, y) une série double. On peut voir sur le nuage de points si les 2 variables sont correlées ou non.





le Coefficient de correlation :

$$r = \frac{cov(x, y)}{\sigma x \cdot \sigma y} \quad -1 \leq r \leq 1$$

s'il est voisin en valeur de 1 l'ajustement est valide. $(0{,}70 \leq |r| \leq 1)$

Application de la methode de moindres-carrés

Regression y en x = $(\Delta) : \hat{y} = ax + b$.

le problème est donc de calculer a et b.

pour que la somme des termes $(y_i - \hat{y}_i)^2$ est minimum

Cela revient à chercher le minimum de la fonction Φ de deux variables a, b

$$\Phi(a, b) = \Sigma(y_i - ax_i - b)^2$$
$$= \Sigma(y_i - \hat{y}_i)^2$$

Φ admet un minimum si $\frac{\partial \Phi}{\partial a} = 0$ et $\frac{\partial \Phi}{\partial b} = 0$

$$\begin{cases} \dfrac{\partial \Phi}{\partial a} = 2 \displaystyle\sum_{i=1}^{n} -x_i (y_i - a x_i - b) = 0 \\ \dfrac{\partial \Phi}{\partial b} = 2 \displaystyle\sum_{i=1}^{n} -(y_i - a x_i - b) = 0 \end{cases}$$

$$\begin{cases} \displaystyle\sum_i -x_i y_i + a \sum_i x_i^2 + b \sum_i x_i = 0 \quad \ldots (1) \\ \displaystyle -\sum_i y_i + a \sum_{x=1}^{n} x_i + \sum_i b = 0 \quad \ldots (2) \end{cases}$$

$$(2) \Leftrightarrow -n\bar{y} + a n \bar{x} + n b = 0$$

$$\Leftrightarrow b = \bar{y} - a\bar{x} \Leftrightarrow \bar{y} = b + a\bar{x}$$

$$(1) \Leftrightarrow -\sum_i x_i y_i + a \sum_i x_i^2 + (\bar{y} - a\bar{x}) \sum_i x_i = 0$$

$$\Rightarrow -\sum_i x_i y_i + a \sum x_i^2 + (\bar{y} - a\bar{x}) n\bar{x} = 0$$

$$a = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$a = \dfrac{cov(x,y)}{V(x)}$. (Δ) la droite de regression y en x

$(\Delta) : y = ax + b$ / $a = \dfrac{cov(x,y)}{V(x)}$, $b = \bar{y} - a\bar{x}$.

En permutant les rôles de x et y on obtient :

la droite (Δ') de regression x en y : $x = a'y + b'$

$$\begin{cases} b' = \bar{x} - a'\bar{y} \\ a' = \dfrac{cov(x,y)}{V(y)}. \end{cases}$$

$a \cdot a' = r^2$

$\Rightarrow |r| = \sqrt{aa'}$

Retour sur le coefficient de corrélation linéaire :

Les 2 droites de regression trouvées sont :

$y = ax + b \qquad a = \dfrac{cov(x, y)}{V(x)}$

$x = a'y + b' \qquad a' = \dfrac{cov(x, y)}{\sigma(y)}$

$$a a' = \frac{cov^2(x, y)}{V(x) \cdot V(y)} = \frac{cov^2(x, y)}{\sigma^2(x) \cdot \sigma^2(y)} = r^2.$$

lorsque les droites sont identiques les coefficient a' et $\frac{1}{a}$ sont egaux : $\frac{1}{a} = a' \Rightarrow a a' = 1 \Rightarrow r^2 = 1 \Rightarrow |r| = 1$

si les droites sont proches $|r|$ est voisin de 1, ce qui correspond à un ajustement valide. si (r) n'est pas très different de zero, c'est que a et a' sont loin d'être inverse l'une de l'autre, et par consequent l'adjustement est non valide

L'adjustement non lineaire :

Il peut arriver que le nuage de points representant une série double ne soient pas alignés, mais soient voisin d'une courbe connu

1/ L'adjustement à l'aide d'une fonction exponentielle : $y = B \cdot A^x$. (x_i [illegible], y [illegible])

si à des valeurs x_i en progression arithmetique correspondant à des valeurs y_i sensiblement en progression géométrique.

On peut ajuster la série par une fonction exponentielle

$Y = B \cdot A^x \iff \ln Y = \ln(B \cdot A^x)$

$\ln Y = \ln B + x \ln A$ de la forme : $y = b + a x$

avec : $\begin{cases} y = \ln Y \\ a = \ln A \end{cases}$

$\begin{cases} a = \dfrac{\operatorname{cov}(x, y)}{V(x)} \\ b = \bar{y} - a\bar{x} \end{cases}$ $\begin{cases} a = \ln A \\ b = \ln B \end{cases} \iff \begin{cases} A = e^a \\ B = e^b \end{cases}$

$$Y = B \cdot A^x = e^b (e^a)^x$$

2/ L'ajustement à l'aide d'une fonction puissance

$Y = B X^a$.

Si une série statistique varie de tel sorte que les valeurs x_i et y_i sont sensiblement en progression géométrique. On peut ajuster par une fonction puissance d'équation : $Y = B X^a$.

$\ln Y = \ln B + a \ln X$ de la forme : $y = b + a x$

avec $\begin{cases} y = \ln Y \\ x = \ln X \\ b = \ln B \end{cases}$

d'où : $a = \frac{cov(x,y)}{V(x)} = \frac{cov(\ln X, \ln Y)}{V(\ln X)}$,

$b = \bar{y} - a\bar{x} = \overline{\ln Y} - a\overline{\ln X}$

$\ln B = b \Leftrightarrow B = e^{b}$

$y = Bx^{a} = e^{b}x^{a}$.

3/ L'ajustemtent à l'aide d'une fonction parabole

si la presentation graphique de la serie semble permettre l'ajustement par une parabole d'équation

$y = a + bX + cX^2$.

le problème consisterait à recherche a, b, c de telle sorte que $\sum (y_i - a - bx_i - cx_i^2)^2$ soit minimum /.

on définit Φ une fonction de 3 variables.

$$\Phi(a,b,c) = \sum (y_i - a - bx_i - cx_i^2)^2$$

Φ est minimum lorsque :

$\frac{\partial \Phi}{\partial a} = 0$, $\frac{\partial \Phi}{\partial b} = 0$, $\frac{\partial \Phi}{\partial c} = 0$ alors.

$$\begin{cases} 2\sum -(y_i - a - bx_i - cx_i^2) = 0 \\ 2\sum -x_i(y_i - a - bx_i - cx_i^2) = 0 \\ 2\sum -x_i^2(y_i - a - bx_i - cx_i^2) = 0 \end{cases}$$

$$\rightarrow \begin{cases} na + b\sum x_i + c\sum x_i^2 = \sum y_i \\ a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = \sum x_i y_i \\ a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 = \sum x_i^2 y_i \end{cases}$$

c'est un système de 3 équations à 3 variables