PROCESSUS STOCHASTIQUE
ANALYSE BIVARIEE

PR. M. DJAMEL MOUSS

Laboratoire d'Automatique et Productique LAP

L'ANALYSE BI VARIEE

L'analyse bi variée étudie des populations suivant deux caractères (ou variables) statistiques X et Y .. Les deux variables observées peuvent être aussi bien quantitatives que qualitatives.

Les tableaux de données seront à deux dimensions : des jeux de données à deux colonnes ou des tableaux d'effectifs à deux entrées. Dans le cas d'une variable quantitative, on pourra faire des calculs d'indicateurs (moyenne, écart-type, etc.) en fonction des modalités de l'autre variable.

Tableau de Contingence

Lorsque les jeux de données comportent toutes les observations, ils se présentent sous la forme de tables dans lesquelles les observations sont représentées en lignes et les variables sont représentées en colonnes.

Désignons par X et Y les deux variables qui peuvent être qualitatives ou quantitatives et qui peuvent ne pas être de même nature.

La variable X possède p modalité et seront désignés par $x_1, x_2, \dots, x_i, \dots, x_p$; les q modalités de Y sont désignées par $y_1, y_2, \dots, y_j, \dots, y_q$.

Remarque: La i modalité d'une variable désigne le centre de la classe i dans le cas d'une variable quantitative continue

Y / X	y_1	y_2	…	y_j	…	y_q	Total
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2q}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iq}	$n_{i.}$
…	…	…	…	…	…	…	…
x_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}	$n_{p.}$
Total	$n_{.1}$	$n_{.1}$	…	$n_{.1}$	…	$n_{.1}$	$n_{..}$

Les sommes partielles qui figurent dans la dernière ligne et la dernière colonne les sommes partielles quand une variable reste fixe. Elles constituent une distribution uni-variée qu'on appelle distribution marginale du nombre des variables.

L'effectif n_{ij} désigne le nombre de fois où la modalité x_i de la variable X et la modalité y_j de la variable Y ont été observées simultanément.

On a donc:

$$n_{i\blacksquare} = \sum_{j=1}^{q} n_{ij} \qquad n_{\blacksquare j} = \sum_{i=1}^{p} n_{ij}$$

$$N = \sum_{j=1}^{q} \sum_{i=1}^{p} n_{ij}$$

$$\sum_{i=1}^{p} n_{i\blacksquare} = \sum_{j=1}^{q} n_{\blacksquare j} = N$$

La tableau de contingence peut être établit pour les fréquences

En posant :

$$f_{ij} = \frac{n_{ij}}{N}$$

Y / X	y_1	y_2	…	y_j	…	y_q	Total
x_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1q}	$f_{1.}$
x_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2q}	$f_{2.}$
…	…	…	…	…	…	…	…
x_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{iq}	$f_{i.}$
…	…	…	…	…	…	…	…
x_p	f_{p1}	f_{p2}	…	f_{pj}	…	f_{pq}	$f_{p.}$
Total	$f_{.1}$	$f_{.1}$	…	$f_{.1}$	…	$f_{.1}$	1

Avec pour les fréquences

$$f_{\blacksquare j} = \sum_{i=1}^{p} f_{ij} \qquad pour\ j \in [1; q]$$

$$f_{i\blacksquare} = \sum_{j=1}^{q} f_{ij} \qquad pour\ i \in [1; p]$$

$$\sum_{i=1}^{p} f_{i\blacksquare} = \sum_{j=1}^{q} f_{\blacksquare j} = 1$$

Moyenne Marginale

Les distributions marginales (qu'elles soient en ligne ou en colonne) sont des distributions uni-variées et donc on peut leur appliquer toutes les propriétés des distributions à une variable. En particulier, si les variables sont quantitatives, il n'y a aucune difficulté à calculer leur moyenne. On les appelle moyennes marginales.

Pour la variable *x*

$$\bar{x}_{Marg} = \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare} * x_i = \sum_{i=1}^{p} f_{i\blacksquare} * x_i$$

Pour la variable *y*

$$\bar{y}_{Marg} = \frac{1}{N}\sum_{j=1}^{q} n_{\blacksquare j} * y_j = \sum_{j=1}^{q} f_{\blacksquare j} * x_i$$

Moyenne Conditionnelle

Une autre approche intéressante consiste à regarder le tableau de contingence ligne par ligne (ou colonne par colonne).

On définit alors une moyenne (Fréquence) conditionnelles. On obtient cette caractéristique en fixant une variable sur une ligne(une colonne)

Pour les fréquences, nous introduisons une nouvelle notation (proche des probabilité conditionnelle. Cette notation $f_{j/i}$ se lit « *f indice j sachant i)*

$$f_{j/i} = \frac{n_{ij}}{n_{i\blacksquare}} \qquad f_{i/j} = \frac{n_{ij}}{n_{\blacksquare j}}$$

Chaque distribution conditionnelle (en ligne ou en colonne) peut être considérée isolément comme une distribution uni-variée. On peut donc lui associer n'importe lequel des indicateurs usuels des distributions à une variable. On obtient ainsi en particulier p moyennes pour les distributions conditionnelles de Y sachant X et q moyennes pour les distributions conditionnelles de X sachant Y .

Pour la variable x

$$\bar{x}_j = \frac{1}{n_{\blacksquare j}} \sum_{i=1}^{p} n_{ij} * x_i = \sum_{i=1}^{p} f_{i/j} * x_i$$

Pour la variable y

$$\bar{y}_i = \frac{1}{n_{i\blacksquare}} \sum_{j=1}^{q} n_{ij} * y_j = \sum_{j=1}^{q} f_{j/i} * y_j$$

Variance Marginale

Les variances marginales sont définit comme suit

Pour la variable x

$$V(x) = \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare}\left(x_i - \bar{x}_{Marg}\right)^2$$

$$V(x) = \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare}\, x_i^2 - \left(\bar{x}_{Marg}\right)^2 = \sum_{i=1}^{p} f_{i\blacksquare}\, x_i^2 - \left(\bar{x}_{Marg}\right)^2$$

Pour la variable y

$$V(y) = \frac{1}{N}\sum_{j=1}^{q} n_{\blacksquare j}\left(y_j - \bar{y}_{Marg}\right)^2$$

$$V(y) = \frac{1}{N}\sum_{i=1}^{p} n_{\blacksquare j}\, y_j^2 - \left(\bar{y}_{Marg}\right)^2 = \sum_{i=1}^{p} f_{\blacksquare j}\, y_j^2 - \left(\bar{x}_{Marg}\right)^2$$

Variance Conditionnelle

Pour la variable x

$$V_j(x) = \frac{1}{n_{\blacksquare j}} \sum_{i=1}^{p} n_{ij} \left(x_i - \overline{x}_j\right)^2 = \sum_{i=1}^{p} f_{i/j} * \left(x_i - \overline{x}_j\right)^2$$

$$V_j(x) = \frac{1}{n_{\blacksquare j}} \sum_{i=1}^{p} n_{ij}\, x_i^2 - \left(\overline{x}_j\right)^2 = \sum_{i=1}^{p} f_{i/j}\, x_i^2 - \left(\overline{x}_j\right)^2$$

Variance Conditionnelle

Pour la variable y

$$V_i(y) = \frac{1}{n_{i\blacksquare}} \sum_{i=1}^{q} n_{ij} (y_j - \bar{y}_i)^2 = \sum_{i=1}^{q} f_{j/i} * (y_j - \bar{y}_i)^2$$

$$V_i(y) = \frac{1}{n_{i\blacksquare}} \sum_{i=1}^{q} n_{ij}\, y_j^2 - (\bar{y}_i)^2 = \sum_{i=1}^{q} f_{j/i}\, y_j^2 - (\bar{y}_i)^2$$

Relation entre les moyennes

La relation entre Moyenne Marginale et Moyenne Conditionnelle est:

Pour la variable x

$$\overline{x}_{Marg} = \frac{1}{N}\sum_{j=1}^{q} n_{\blacksquare j}\,\overline{x}_j$$

Pour la variable y

$$\overline{y}_{Marg} = \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare}\,\overline{y}_i$$

Relation entre les variances

La relation entre Variance Marginale et Variance Conditionnelle est:

La variance marginale est égale à la somme de :

- la moyenne pondérée des variances conditionnelles
- la variance pondérée des moyennes conditionnelles

Pour la variable x

$$V(x) = \frac{1}{N}\sum_{j=1}^{q} n_{\blacksquare j}\, V_j(x) + \frac{1}{N}\sum_{j=1}^{q} n_{\blacksquare j}\left(\bar{x}_j - \bar{x}_{Marg}\right)^2$$

$$V(x) = \overline{V_j(x)} + V\left(\bar{x}_j\right)$$

Pour la variable y

$$V(y) = \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare}\, V_i(y) + \frac{1}{N}\sum_{i=1}^{p} n_{i\blacksquare}\left(\bar{y}_i - \bar{y}_{Marg}\right)^2$$

$$V(y) = \overline{V_i(y)} + V\left(\bar{y}_i\right)$$

Exemple

Considérons pour cet exemple une population composé de couples de différents âges. On s'intéresse à leur âge lors de leur mariage. Posons X l'âge du mari et Y l'âge de la femme. L'étude statistique donne le tableau suivant:

	[15,20[	[20,25[	[25,30[	[30,35[	Total
[20,25[	11	11	2	0	24
[25,30[	9	13	5	1	28
[30,35[	2	11	6	2	21
[35,40[	0	4	16	7	27

		[15,20[	[20,25[	[25,30[	[30,35[	$n_{i\blacksquare}$	$n_{i\blacksquare} * x_i$	x_i^2	$n_{i\blacksquare} * x_i^2$		Moyenne y_j
		17.5	22.5	27.5	32.5						
[20,25[	22.5	11	11	2	0	24	540	506.25	10450	495	20.63
[25,30[	27.5	9	13	5	1	28	770	756.25	14175	620	22.14
[30,35[	32.5	2	11	6	2	21	682.5	1056.3	12831.25	512.5	24.4
[35,40[	37.5	0	4	16	7	27	1012.5	1406.3	21518.75	757.5	28.06
$n_{\blacksquare j}$		22	39	29	10	100	3005		58975		
$n_{\blacksquare j} * y_j$		385	877.5	797.5	325	2385					
y_j^2		306.25	506.25	756.25	1056.3	2625					
$n_{\blacksquare j} * y_j^2$		14487.5	32643.75	33631.25	12712.5	93475					
		560	1112.5	977.5	355	3005					
Moyenne x_i		25.46	28.53	33.71	35.5						

$$\bar{x}_{Marg} = \frac{3005}{100} = 30.05 \qquad \bar{y}_{Marg} = \frac{2385}{100} = 23.85$$

$\bar{x}_1 = 25.45$ $\bar{x}_2 = 28.53$ $\bar{y}_1 = 20.63$ $\bar{y}_2 = 22.14$

$\bar{x}_3 = 33.71$ $\bar{x}_4 = 35.50$ $\bar{y}_3 = 24.40$ $\bar{y}_4 = 28.06$

$$V(x) = \frac{93475}{100} - (30.05)^2 = 31.75 \qquad V(y) = \frac{58975}{100} - (23.85)^2 = 20.93$$

$$V_1(x) = \frac{14487.5}{22} - (25.46)^2 = 10.31 \qquad V_2(x) = \frac{32643.75}{39} - (28.53)^2 = 23.06$$

$$V_3(x) = \frac{33631.25}{29} - (33.71)^2 = 23.33 \qquad V_4(x) = \frac{12712.5}{10} - (35.5)^2 = 11.00$$

$$V_1(y) = \frac{10450}{24} - (20.63)^2 = 9.82 \qquad V_2(y) = \frac{14175}{28} - (22.14)^2 = 16.07$$

$$V_2(y) = \frac{14175}{28} - (22.14)^2 = 16.07 \qquad V_4(y) = \frac{21518.75}{27} - (28.06)^2 = 9.63$$

COORELATION & REGRESSION

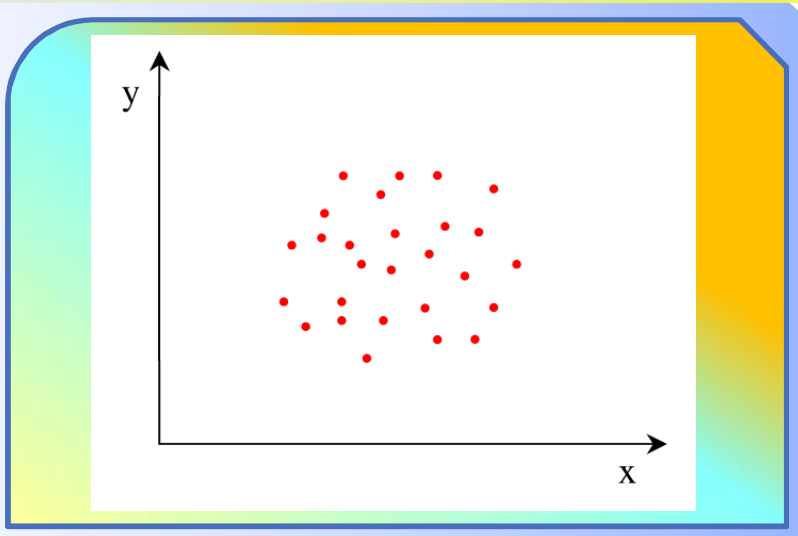
la **corrélation** entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance .
Cette corrélation est très souvent réduite à la corrélation ***linéaire*** entre variables quantitative, c'est-à-dire l'ajustement d'une variable par rapport à l'autre par une relation affine obtenue par régression linéaire.

Pour cela, on calcule un ***coefficient de corrélation linéaire***, quotient de leur covariance par le produit de leurs écart-types .
Son signe indique si des valeurs plus hautes de l'une correspondent « en moyenne » à des valeurs plus hautes ou plus basses pour l'autre. La valeur absolue du coefficient, toujours comprise entre 0 et 1, ne mesure pas l'intensité de la liaison mais la prépondérance de la relation affine sur les variations internes des variables.

Un coefficient nul n'implique pas indépendance, car d'autres types de corrélation sont possibles.

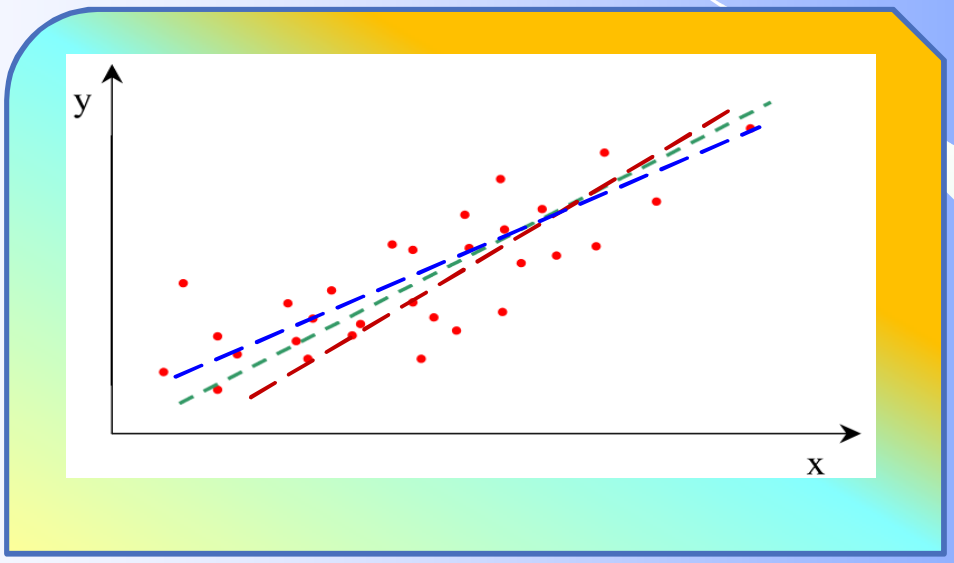
Le fait que deux variables soient « fortement corrélées » ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre.

Soit deux V.A quantitatives, représentons dans un plan les valeur de la V.A X sur l'axe des « x » et les valeurs de la V.A Y sur l'axe des « y »



Cherchons dans un premier temps a établir la relation (si elle existe) entre ces deux variable sous la forme d'une droite $y = a.x + b$ (On l'appelle la droite de régression des y en x Noté $D_y(x)$)

Calculer le coefficient de corrélation entre deux variables numériques revient à chercher à résumer la liaison qui existe entre les variables à l'aide d'une droite. On parle alors d'un ajustement linéaire.

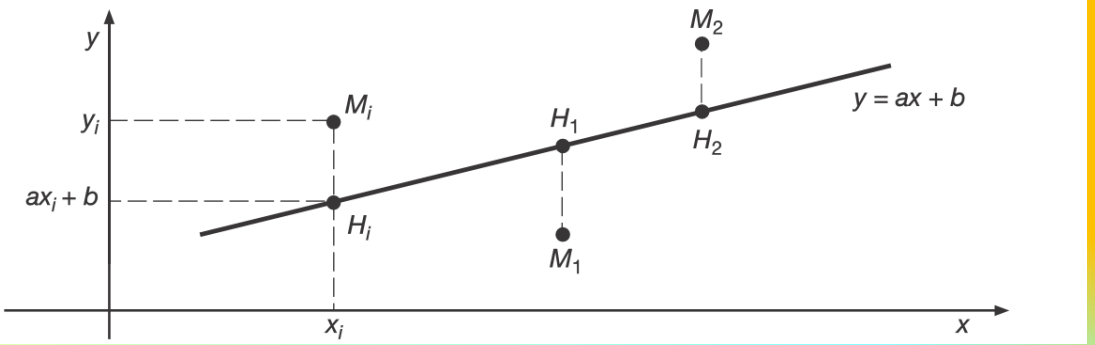


Un ajustement à main levé conduira à une multitude de proposition. La solution est de rechercher analytiquement la meilleurs droite.

Comment calculer les caractéristiques de cette droite ? En faisant en sorte que l'erreur que l'on commet en représentant la liaison entre nos variables par une droite soit la plus petite possible. Le critère formel le plus souvent utilisé, mais pas le seul possible, est de minimiser la somme de toutes les erreurs effectivement commises au carré. On parle alors d'ajustement selon la méthode des moindres carrés.

Méthode des moindres carrés

Position du problème : Nous avons que le montre la figure une suite d'observation $M_i(x_i,y_i)$. La droite que nous cherchons ne passe pas par tous les points (désignons les par H_i (x_i, ax_i+b)



Pour une même abscisse i la distance H_iM_i peut être soit positive (exemple M_2) soit négative (exemple M_1).La somme de ces différences peut être nulle sans pour autant donner la meilleur droite (à des termes négatifs et positif).

La solution est de prendre la somme des carrés de ces distance et chercher à la minimiser (d'où le nom des moindres carrés)

$$\sum_{i=1}^{n} |H_i M_i|^2 \ soit\ Minimum$$

Les distances sont comptées parallèlement à l'un des axes des coordonnées ; nous avons choisi ici l'axe des ordonnées. Il s'agit de déterminer la droite d'équation $y = ax_i + b$ telle que:

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2 \ soit\ Minimum$$

$$\sum_{i=1}^{n} ((y_i - ax_i) - b)^2 = \sum_{i=1}^{n} ((y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2)$$

$$= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b \sum_{i=1}^{n} (y_i - ax_i) + nb^2$$

Les facteurs a et b sont alors données par les expression

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\bar{y} = a\,\bar{x} + b$$

Coefficient de Corrélation

Le coefficient de corrélation ne mesure pas l'intensité de la liaison mais la prépondérance de la relation affine sur les variations internes des variables.

un **coefficient de corrélation linéaire** est le quotient de leur covariance par le produit de leurs écart-types. Son signe indique si des valeurs plus hautes de l'une correspondent « en moyenne » à des valeurs plus hautes ou plus basses pour l'autre. La valeur absolue du coefficient, toujours comprise entre 0 et 1,

Il est donné par l'expression:

$$r_{xy} = \frac{COV(X,Y)}{\sqrt{V(X) * V(Y)}} \qquad r_{xy} = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} * \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

Le coefficient de corrélation est compris entre -1 et +1. Plus il s'éloigne de zéro, meilleure est la corrélation

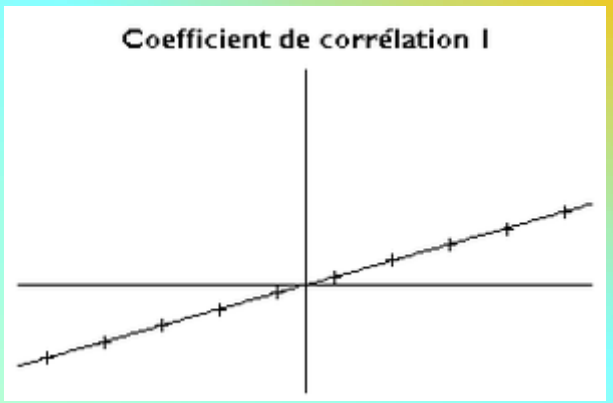
r = +1 corrélation positive parfaite
r =-1 corrélation négative parfaite
r = 0 absence totale de corrélation

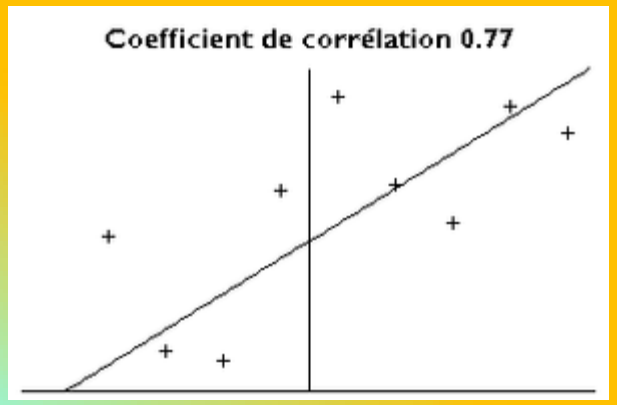
Remarque

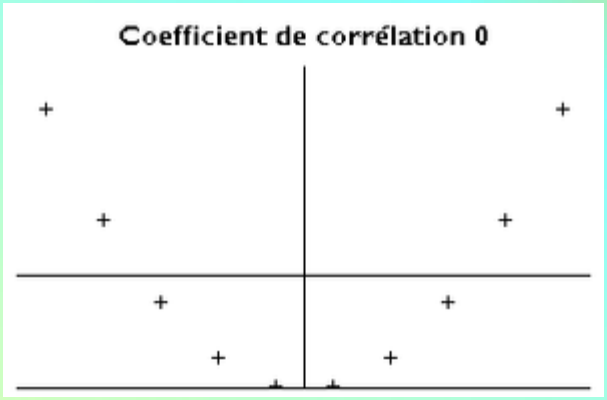
Si $D_y(x)$ s'exprime par $y = ax + b$ et $D_x(y)$ par $x = a'y + b'$ alors :

$$r_{xy}^2 = a * a'$$

Quelques exemples







Cas d'une variation exponentielle

Si le nuage de point définit par les couples (xi, yi) a la forme d'une courbe exponentielle

$$y = Be^{Ax}$$

En passant par le logarithme et en faisant un changement de variable

$$\ln y = \ln Be^{Ax} = \ln B + Ax \ln e$$

En posant $\ln y = Y$ $A = a$ $\ln B = b$

$$Y = ax + b$$

Nous obtenons alors une relation linéaire entre:

- ***la variable x***
- ***la variable Y = ln y***

Cas d'une variation puissance

Si le nuage de point définit par les couples (xi, yi) a la forme d'une courbe exponentielle

$$y = Bx^A$$

En passant par le logarithme et en faisant un changement de variable

$$\ln y = \ln Bx^A = \ln B + A \ln x$$

En posant $\ln x = X$ $\ln y = Y$ $A = a$ $\ln B = b$

$$Y = aX + b$$

Nous obtenons alors une relation linéaire entre :

- ***la variable $X = \ln x$***
- ***la variable $Y = \ln y$***