



الكلية متعددة التخصصات الناحور

ⵜⴰⵎⴻⵔⴰⵏⵜ ⵜⴰⵎⴻⵔⴰⵏⵜ ⵜⴰⵏⴰⵔⵉⵜ | 1978
Faculté Pluridisciplinaire de Nador

STATISTIQUE

Dépendance linéaire entre deux caractères

Filière: SMC S4

Professeur: Toufik Chaayra

Statistique descriptive bivariée

- Série statistique bivariée
- On s'intéresse à deux variables x et y . *Ces deux variables sont mesurées sur les n unités d'observation.*
- Pour chaque unité, on obtient donc deux mesures. La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu :
 - $(x_1, y_1), \dots, (x_n, y_n)$

Exemple

On mesure le poids Y et la taille X de 20 individus.

y_i	x_i	x_i*y_i
60	155	9300
61	162	9882
64	157	10048
67	170	11390
68	164	11152
69	162	11178
70	169	11830
70	170	11900
72	178	12816
73	173	12629
75	180	13500
76	175	13300
78	173	13494
80	175	14000
85	179	15215
90	175	15750
96	180	17280
96	185	17760
98	189	18522
101	187	18887
Total	77,45	172,9
		13391,105

Analyse des variables

- Les variables x et y peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

- La moyenne \bar{x}, \bar{y}
- La variance s_x^2, s_y^2

Covariance

- La *covariance* est définie:

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

- Remarque:
 - La covariance peut prendre des valeurs positives, négatives ou nulles.
 - Quand $x_i = y_i$; pour tout $i = 1, \dots, n$; la covariance est égale à la variance.

Théorème

- La covariance peut s'écrire aussi sous la forme suivante:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \overline{xy}$$

- Démonstration: à faire

Corrélation

- Le *coefficient de corrélation* est la *covariance* divisée par les deux *écart-types marginaux*:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Corrélation

- Le coefficient de corrélation mesure la dépendance linéaire entre deux variables X et Y :
- Si le coefficient de corrélation est égal à 1 ou proche de un, il y a une forte dépendance linéaire entre X et Y .
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut
- cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.

Exemple

On mesure le poids Y et la taille X de 20 individus.

y_i	x_i	$x_i \cdot y_i$
60	155	9300
61	162	9882
64	157	10048
67	170	11390
68	164	11152
69	162	11178
70	169	11830
70	170	11900
72	178	12816
73	173	12629
75	180	13500
76	175	13300
78	173	13494
80	175	14000
85	179	15215
90	175	15750
96	180	17280
96	185	17760
98	189	18522
101	187	18887
Total	77,45	172,9
		13391,105

$$\bar{Y} = \frac{1}{20} \sum_{i=1}^{20} y_i = 77,45$$

$$\bar{X} = \frac{1}{20} \sum_{i=1}^{20} x_i = 172,9$$

$$\frac{1}{20} \sum_{i=1}^{20} x_i y_i = 13491,65$$

$$s_{xy} = Cov(X, Y) = \frac{1}{20} \sum_{i=1}^{20} x_i y_i - \bar{X}\bar{Y} = 100,545$$

$$s_y = \sqrt{Var(Y)} = \sqrt{\frac{1}{20} \sum_{i=1}^{20} y_i^2 - \bar{Y}^2} = 12,74$$

$$s_x = \sqrt{Var(X)} = \sqrt{\frac{1}{20} \sum_{i=1}^{20} x_i^2 - \bar{X}^2} = 9,469619$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0,83 = 83\%$$

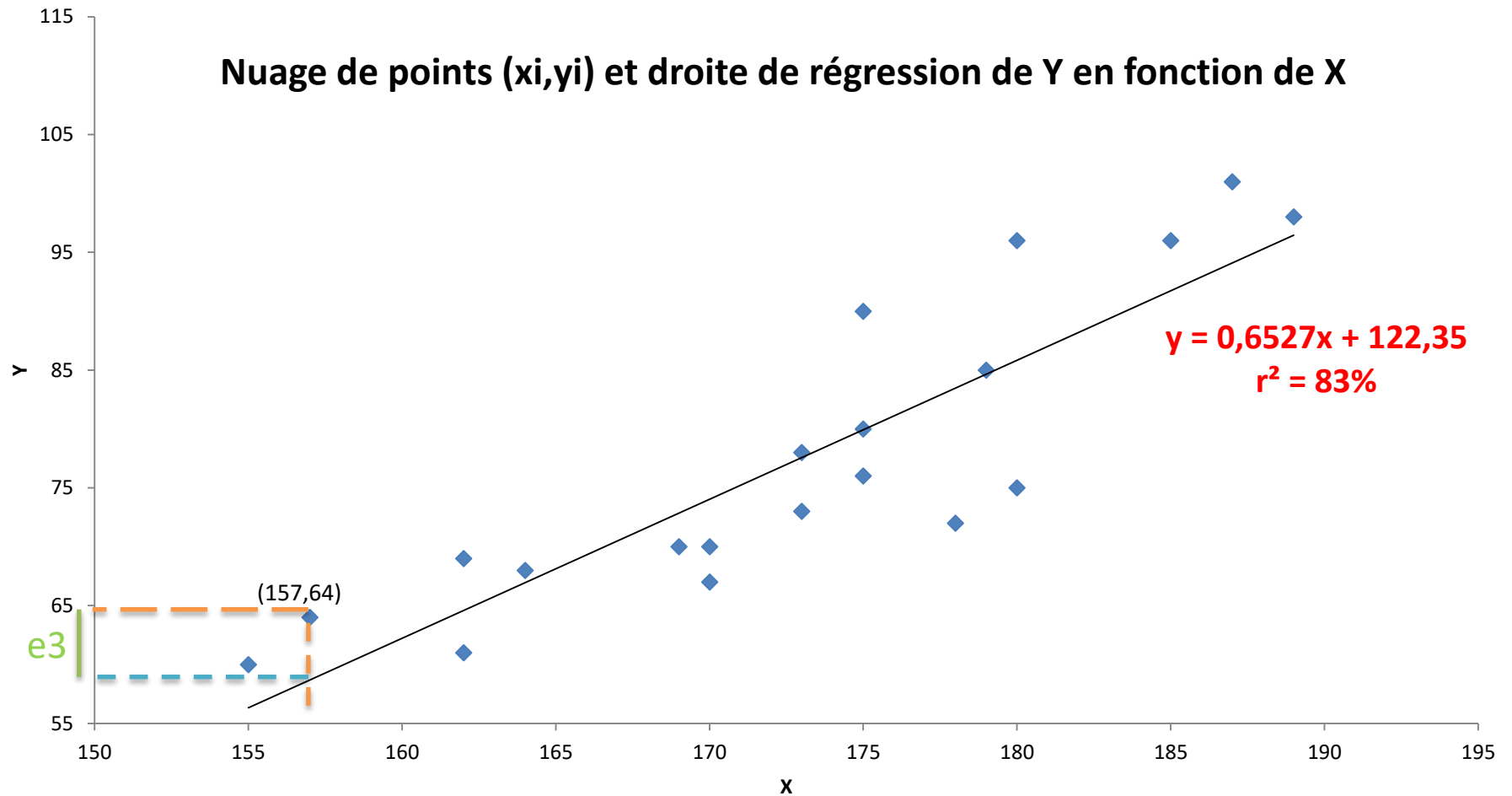
Conclusion: puisque le coefficient de corrélation est proche de 1 alors il y a une forte corrélation directe entre la variable 'poids Y ' et la variable 'taille X '.

Droite de régression

- *La droite de régression est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés.*
- *On considère que la variable X est explicative et que la variable Y est dépendante. L'équation d'une droite est*

$$y = bx + a$$

Représentation de nuages de points (x_i, y_i) et droite de régression de Y en fonction de X



Droite de régression

- Le problème consiste à identifier une droite qui ajuste bien le nuage de points. Si les coefficients *a* et *b* étaient connus, on pourrait calculer les résidus de la régression définie par:

$$e_i = y_i - bx_i - a$$

Droite de régression

- Pour déterminer la valeur des coefficients a et b on utilise le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des résidus :

$$M(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

Théorème

- Les coefficients a et b sont donnés par:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad a = \bar{Y} - b\bar{X}$$

Démonstration

Puisque les estimateurs a et b minimisent $M(a, b)$, alors

$$\frac{dM(a, b)}{da} = 0$$

$$\frac{dM(a, b)}{db} = 0$$

On a

$$\frac{dM(a,b)}{da} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{dM(a,b)}{db} = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0$$

Alors

$$\frac{dM(a,b)}{da} = \sum_{i=1}^n (y_i - a - bx_i) = \sum_{i=1}^n y_i - na - b \sum_{i=1}^n x_i = 0$$

$$\frac{dM(a,b)}{db} = \sum_{i=1}^n x_i (y_i - a - bx_i) = \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0$$

$$\frac{dM(a,b)}{da} = \sum_{i=1}^n (y_i - a - bx_i) = n\bar{y} - na - bn\bar{x} = 0$$

$$\frac{dM(a,b)}{db} = \sum_{i=1}^n x_i (y_i - a - bx_i) = \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0$$

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

Par la suite

$$\sum_{i=1}^n x_i y_i - (\bar{y} - b\bar{x}) \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i - b \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) - \bar{y} \sum_{i=1}^n x_i = 0$$

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

$$\sum_{i=1}^n x_i y_i - bns_x^2 - n\bar{y}\bar{x} = 0$$

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

$$ns_{xy} - bns_x^2 = 0$$

ET alors

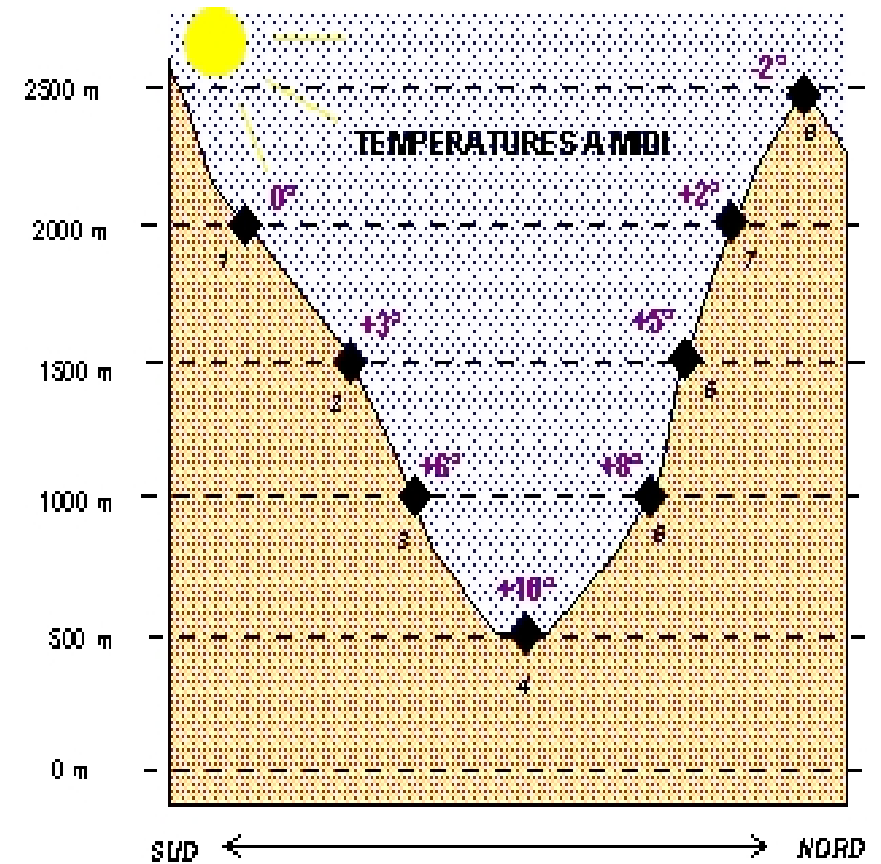
$$a = \bar{y} - b\bar{x}$$

$$b = \frac{s_{xy}}{s_x^2}$$

Exemple

Tableau : Paramètres caractéristiques de la température (Y) et de l'altitude (X) de 8 stations météorologiques d'une vallée alpine (données imaginaires)

i	(Xi)	(Yi)	(Xi-mX)	(Yi-mY)	(Xi-mX)(Yi-mY)
1	2000	0	500	-4	-2000
2	1500	3	0	-1	0
3	1000	6	-500	2	-1000
4	500	10	-1000	6	-6000
5	1000	8	-500	4	-2000
6	1500	5	0	1	0
7	2000	2	500	-2	-1000
8	2500	-2	1000	-6	-6000
moyenne	mX = 1500	mY = 4	0	0	-2250
écart-type	612	3.8	-	-	-



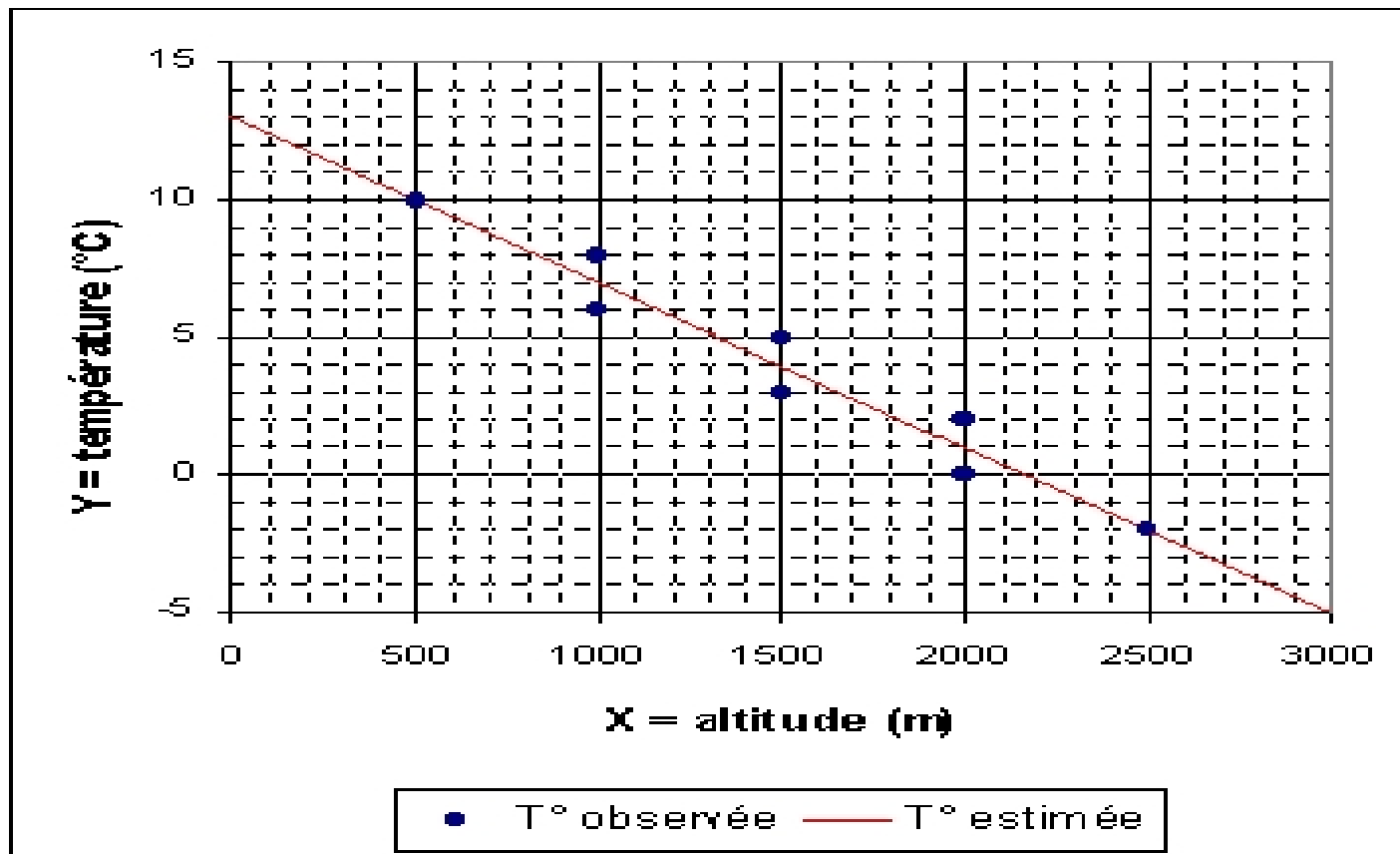
Exemple

On déduit de la valeur de la covariance (-2250) et de celle des deux écarts-type (654 pour X et 4.04 pour Y) l'existence d'une très forte corrélation linéaire négative entre les deux variables :

$$r(X,Y) = \text{Cov}(X,Y) / [\text{ect}(X) * \text{ect}(Y)] = -0.85.$$

Même si l'on tient compte du nombre réduits d'observation (8 stations météorologiques soit 7 degrés de liberté) cette corrélation apparaît hautement significative : il y a moins d'une chance sur 1000 que le hasard ait pu engendrer une corrélation aussi forte entre les deux variables X et Y. La forme du nuage de point croisant les valeurs de X et de Y est par ailleurs parfaitement linéaire

Figure : Droite de régression exprimant la température en fonction de l'altitude pour 8 stations météorologiques d'une vallée alpine (données imaginaires)



$$y = -0,005x + 11,87$$

Exemple

Ce qui justifie la recherche d'un ajustement à l'aide d'une droite.

Il reste à déterminer le **sens de la relation**, c'est-à-dire l'hypothèse faite sur la variable explicative (indépendante) et la variable à expliquer (dépendante). Dans l'exemple choisi, il paraît assez naturel de supposer que la température (Y) dépend de l'altitude (X) et non pas l'inverse, de sorte que l'on va chercher à la température Y en fonction de l'altitude X. Mais la détermination de la relation inverse ne serait pas totalement absurde et l'on pourrait imaginer ... qu'un alpiniste se serve d'un thermomètre pour déterminer l'altitude à laquelle il se trouve (en supposant que les conditions climatiques soient "normales" et qu'il n'y ait pas ce jour là de phénomène d'inversion thermique).