UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

MODULE METHODOLOGIQUE M2

Statistique Descriptive : Cas bidimensionnel

UVED
Université Virtuelle Environnement & Développement Durable

30 AVRIL 2014
VERSION 300414

STATISTIQUE DESCRIPTIVE : CAS BIDIMENSIONNEL

Cécile Mallet

SOMMAIRE

2. Cas bidimensionnel

On considère ici un ensemble d'observations, constitué des mesures X et Y de deux grandeurs réelles. Soient deux variables X et Y étudiées sur le même échantillon. On étudie les liaisons entre les variables. Existe-t–il une relation fonctionnelle entre X et Y ? Existe-t-il une relation affine ? C'est l'étude des corrélations. Il faut noter que la corrélation n'implique pas la causalité. Une cause commune peut en particulier faire varier simultanément deux variables. Les méthodes varient selon la nature des variables étudiées, on ne considère ici que les variables numériques. Supposons donc n individus et deux variables X et Y. On a donc n couples (xi ;yi)

2.1 Phase exploratoire préliminaire

- ✓ L'***étude statistique simple*** de chaque variable a pour but de détecter les erreurs de saisie ou de codage des données, les valeurs aberrantes qui peuvent avoir une très grande influence sur la liaison entre variables
- ✓ Le ***diagramme de dispersion*** (nuage de points , scatter plot) est une représentation cartésienne du tableau de donnée

X	Y
28	70
23	68
52	90
42	75
27	68
29	80
43	78
34	70
40	80
28	72

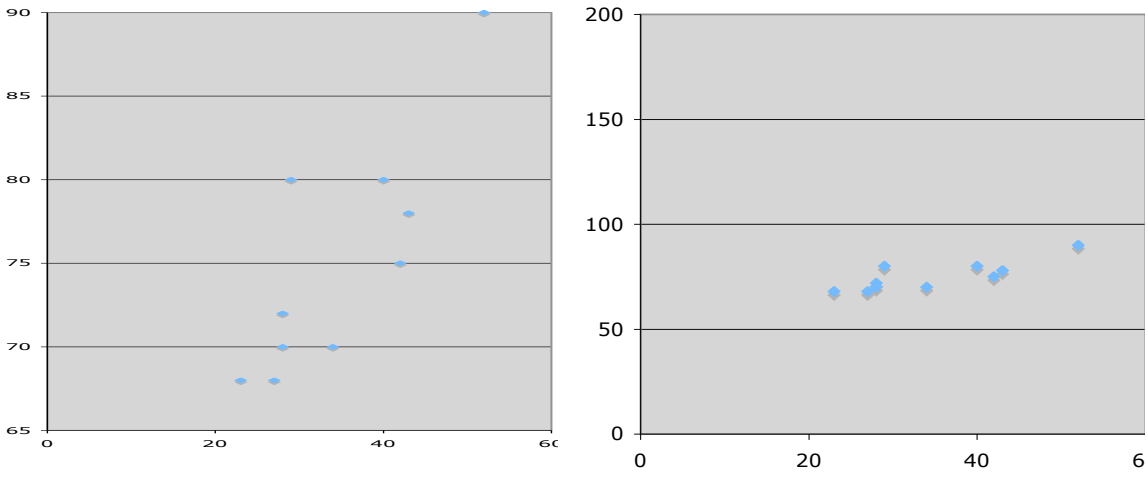

Figure 1 : tableau de données et diagrammes de dispersion réalisés avec deux échelles différentes.

Il apparaît dans l'exemple ci-dessus qu'il existe une relation positive entre les variables. Si x augmente, y augmente. Le choix d'échelles adaptées est important pour mettre en évidence une éventuelle relation entre X et Y. Deux cas sont possibles. Si les variables sont homogènes (même unité) on choisira la même échelle sur les deux axes. Si les variables sont hétérogènes, il est recommandé de représenter les variables centrées et réduites.

Le diagramme de dispersion est une étape indispensable pour décider s'il existe une relation entre les variables et pour déterminer le type d'équation approprié pour décrire la relation.

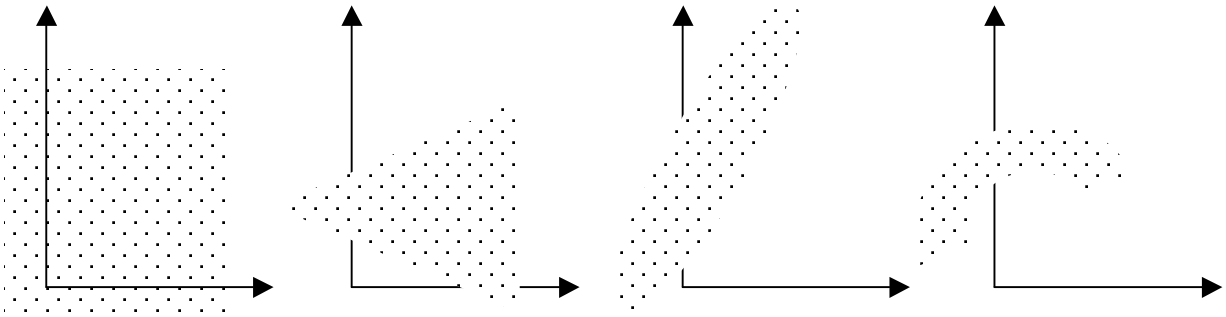

Figure 2 : diagramme de dispersion : étude graphique de la corrélation.

Dans les exemples ci-dessus la forme du nuage indique pour les différents cas (de gauche à droite)

pas de liaison ; pas de liaison en moyenne ; corrélation linéaire positive ; corrélation non linéaire

La corrélation indique la dépendance en moyenne : la moyenne de y est fonction de x. si cette fonction est linéaire on parle de corrélation linéaire.

Histogramme 2D : L'inconvénient du diagramme de dispersion est qu'il ne tient pas compte de la densité des points et dans les cas ou l'on traite un nombre très important données il est préférable de représenter l'histogramme 2D, c'est à dire qu'on défini des classes de X et de Y et qu'on attribue une couleur correspondant à la fréquence de la classe.

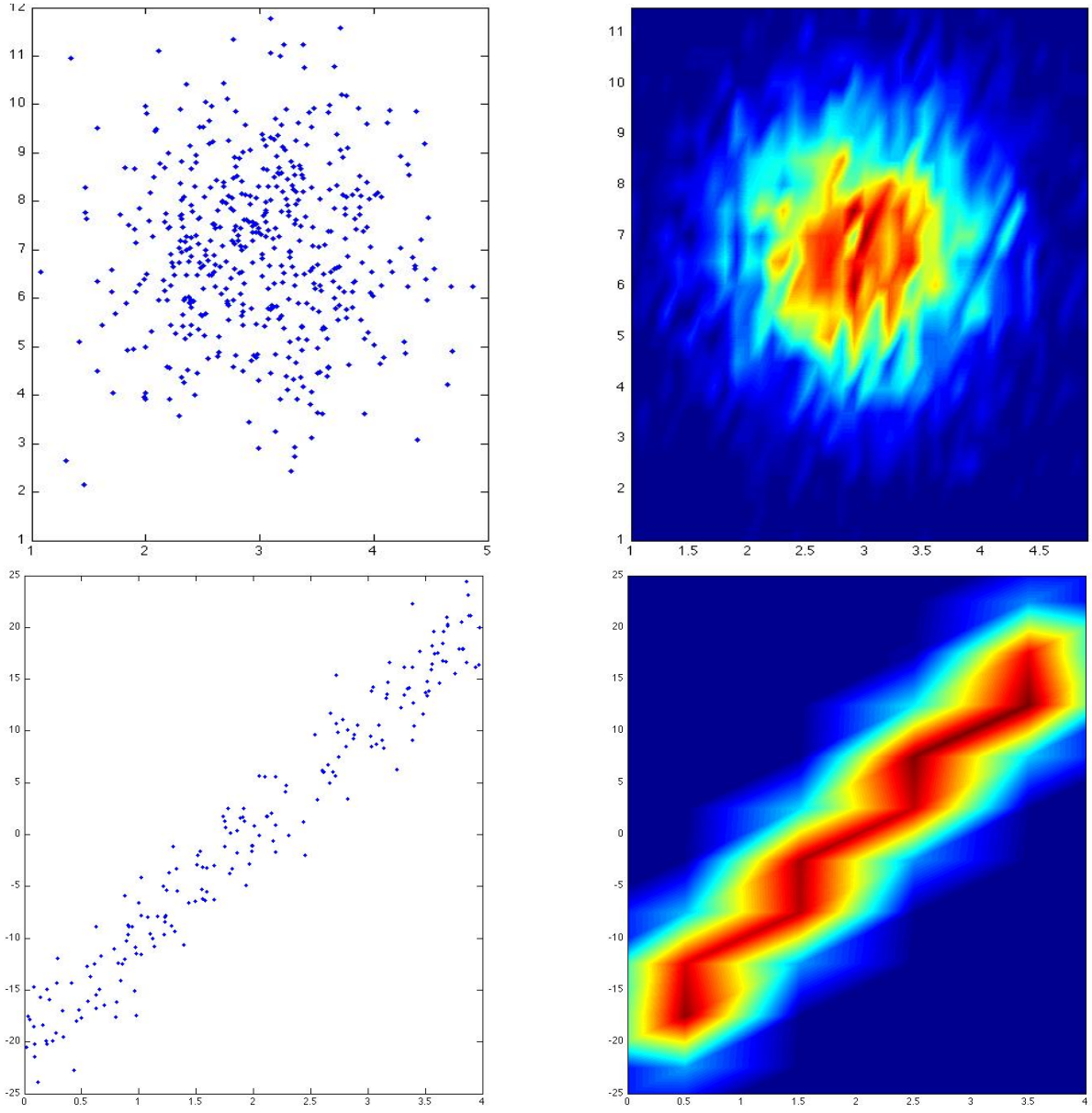

Figure 3 : diagramme de dispersion à gauche et histogramme 2D à droite.

- ✓ ***Centre de gravité*** : C'est le point de coordonnées G $(\overline{x},\overline{y})$.

2.2 Covariance et coefficient de corrélation linéaire

Il s'agit de définir un indice rendant compte numériquement de la manière dont les deux variables varient simultanément.

- On définit d'abord la ***covariance observée*** qui généralise la variance

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

La covariance est un indice symétrique $c_{XY}=c_{YX}$, elle peut prendre toutes les valeurs réelles. Elle dépend des unités de mesure dans lesquelles sont exprimées les variables considérées. Des nuages de points de forme et d'orientation identiques mais de dimension différentes donnent des covariances différentes en raison des ordres de grandeur des valeurs impliquées. C'est la raison pour laquelle on définit la covariance des variables centrées réduites :

- le ***coefficient de corrélation linéaire***

$$r = \frac{c_{XY}}{s_X s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \bar{y}^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2}}$$

Il est égal à la covariance des variables centrées réduites, il est indépendant des unités, il est symétrique et $-1<r<1$. Les valeurs 1 et -1 correspondent à une liaison linéaire parfaite c'est-à-dire $ax_i+by_i+c=0\ \forall i$.

Cependant r ne mesure que le caractère linéaire et son usage doit être réservé à des nuages de points répartis suivant une tendance linéaire. De plus r est très sensible aux individus extrêmes et n'est donc pas robuste.

La corrélation n'est pas transitive : on peut observer que x est corrélée avec y et y est très corrélée avec z, sans pour autant que x soit corrélée avec z.

La figure ci-dessous illustre la nécessité de la visualisation des données avant l'usage du coefficient de corrélation linéaire.

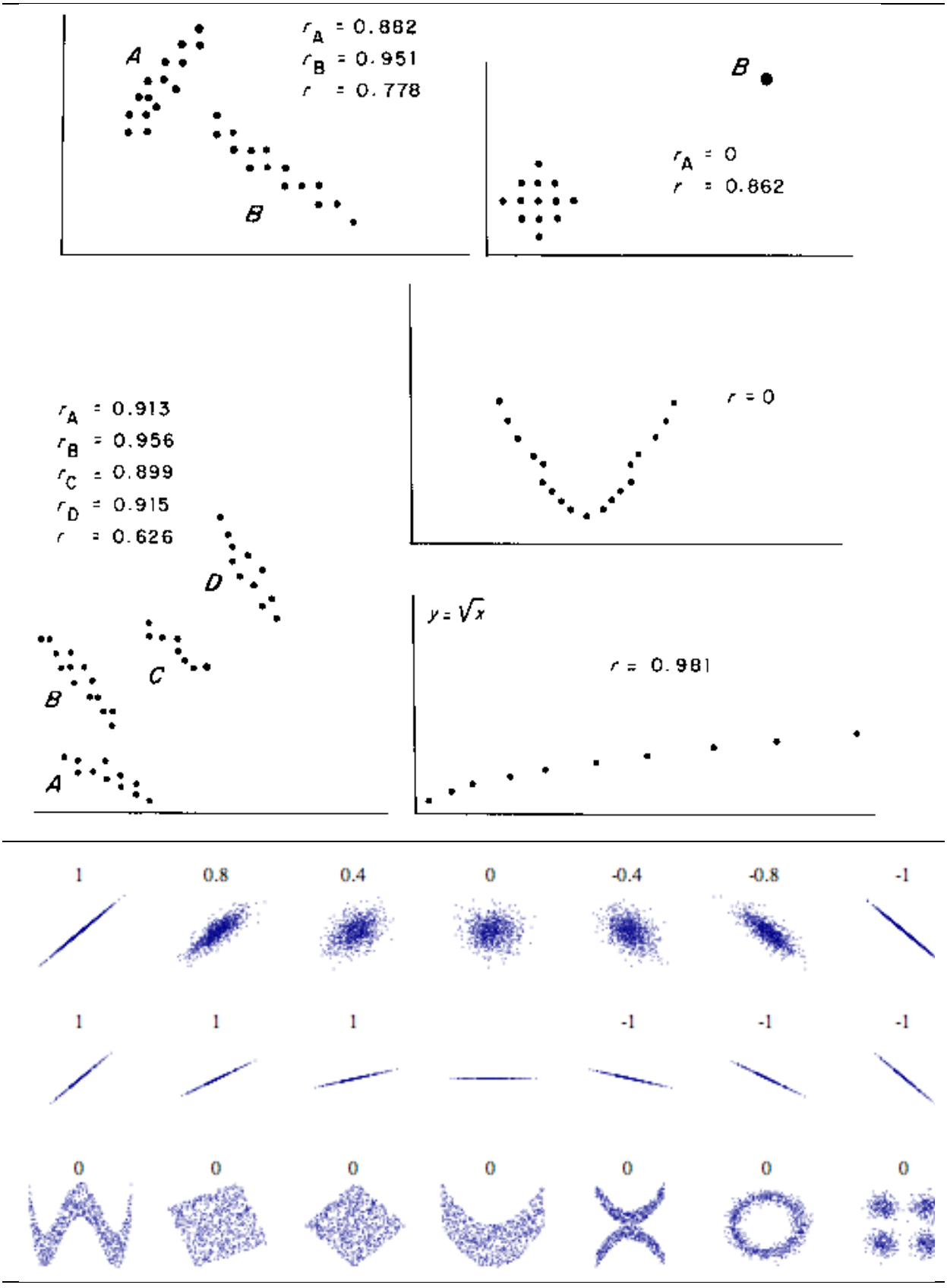

Figure 4 : usage du coefficient de corrélation linéaire.